"DO YOU KNOW WHAT YOU DON'T KNOW?" EXPLORING MONITORING ACCURACY ACROSS DOMAINS OF GENERAL KNOWLEDGE, FINANCIAL CALCULATION, AND PROBABILITY CALCULATION*

STELLA DENTAKOS

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN CLINICAL DEVELOPMENTAL PSYCHOLOGY

YORK UNIVERSITY
TORONTO, ONTARIO

DECEMBER, 2019

ABSTRACT

Confidence and its accuracy have been most commonly examined in domains such as general knowledge and learning, with less study of other domains, such as applied knowledge and problem-solving. Monitoring accuracy in real-world competencies may depend on characteristics of the domain. The current study examined whether monitoring accuracy, both calibration (resistance to overconfidence) and resolution (discrimination) indices, are stable within individuals and across tasks that represent highly diverse areas. The well-established domain of general knowledge and two understudied applied domains of financial calculation and probability calculation were examined. In addition, correlations between monitoring accuracy and cognitive abilities (intellectual ability and working memory) and several aggregated judgments regarding each task as a whole (ratings of predicted and postdictive performance, task difficulty, and effort required), as well self-perceptions relating to test anxiety and academic self-concept were explored. Calibration was significantly positively correlated across tasks, reflecting a person-centered trait, but not resolution. Cognitive abilities were predictive of both calibration and resolution across tasks, while other task-specific judgments and self-perception variables demonstrated varied and tasks specific associations. Monitoring accuracy was not predictive of real-world outcomes including academic average and learning challenges. Overall study findings support that when considering a wide range of domains, calibration displays domain-generality, while resolution displays domain-specificity.

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation I have received a great deal of support and assistance.

First and foremost, I would like to thank my supervisor Dr. Maggie Toplak for her invaluable advice, dedication, and mentorship. Maggie, I cannot imagine not only creating this project, but overcoming the many challenges of these doctoral years without your support. Know that I am beyond thankful and grateful.

I would also like to extend my thanks to Dr. Rakefet Ackerman for her support and insight into this project, particularly the calculation of monitoring indices. Her expertise was extremely helpful and valued.

I would like to thank my committee members, Drs. Gary Turner and Geoff Sorge for generously offering their time, support, and continuous engagement in the preparation and review of this dissertation. I would like to also thank Dr. Emma Climie for taking on the role of external examiner, Dr. Jim Bebko for chairing the defense, and Dr. Alison Macpherson for participating as my internal/external member on my defense. Your insights and valuable advice were greatly appreciated.

I would also like to thank my fellow lab member, Wafa Saoud, for taking on this project with me. The combination of our strengths and ideas made for a great team. Special thanks to Amanda Edwards for all the administrative support and helping this study get under way. I would also like to acknowledge Cameron Amini for the important data entry work.

Last, I would not have been able to complete this without some amazing people for whose help I am immeasurably grateful: my parents, my Nonna, my family, and my incredible friends. Thank you for supporting this dream over all these years. Through changing cities, financial struggles, illness - your belief in me has been unwavering. This journey has not always been easy, but it has always been worth it.

I am extremely grateful to my husband, Brandon. Thank you for your constant love, encouragement, understanding, and patience. And for holding down the fort during all those late nights and early mornings spent writing at the coffee shop. I would not be where I am today without you.

Last, and definitely not least, my most heartfelt thanks to my son, Logan.

Logan, I dedicate this to you.

That I was able to write this document during your first year of life is a testament to how special you are and how much you motivate me. You have made me stronger, better, and more fulfilled than I could have ever imagined. You are my greatest accomplishment.

Stella Dentakos

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

"Do you know what you don't know?" Exploring monitoring accuracy across domains of general knowledge, financial calculation, and probability calculation

The discrepancy between knowledge and metacognitive monitoring of one's knowledge has been examined extensively in both the metacognition (Kleitman & Stankov, 2007; Koriat, 2008, 2012a; Stankov, Kleitman, & Jackson, 2014) and in the judgment and decision-making (Bruine de Bruin, Parker, & Fischhoff, 2007; Parker & Fischhoff, 2005; Stanovich, West, & Toplak, 2016; Yates, Lee, & Bush, 1997; West & Stanovich, 1997) fields. Metacognitive monitoring represents a subjective assessment of how well a task has been, is, or will be performed (Ackerman & Thompson, 2017; Bjork, Dunlosky, & Kornell, 2013; Nelson & Narens, 1980). Several indicators of monitoring accuracy have been used, including calibration and resolution, as detailed below. In addition to different methods for assessing monitoring accuracy, the nature of the domain examined may also impact accuracy (Erickson & Heit, 2015; West & Stanovich, 1997). In this study, calibration and resolution were examined within the classic domain of general knowledge and compared to additional domains that have been scarcely studied in this context despite being common in real-life scenarios: financial calculation and probability calculation. In addition, the question of domain generality versus specificity was also explored by examining whether calibration and resolution were related across these diverse domains. Cognitive abilities, represented by intellectual abilities and working memory, task-specific judgments including predictive and postdictive confidence, task difficulty and effort ratings, as well as self-perceptions of academic self-concept and cognitive test anxiety, were explored as potential monitoring accuracy correlates in each of these domains. The relationship between monitoring accuracy and real-world outcomes of academic success and learning-related challenges were also examined.

Indices of Monitoring Accuracy

Metacognition is traditionally defined as "thinking about thinking" (Flavell, 1979) or "knowing about knowing" (Metcalfe & Shimamura, 1995) and includes aspects of metacognitive monitoring and control (Schraw & Dennison, 1994). Metacognitive monitoring reflects the ability to evaluate current states of knowledge, such as judging that performance on a problem is poor. Metacognitive control refers to the ability to engage in regulation strategies, such as reworking the problem or double-checking calculations (Grainger, Williams, & Lind, 2016; Zabrucky, 2010). Metacognitive monitoring is not only thought to precede control, but to also dictate subsequent decision-making and effort regulation (Zabrucky, 2010). There are two ways that accuracy of metacognitive monitoring can be measured: calibration and resolution.

Calibration. Calibration represents the degree of fit between subjective feelings of certainty or correctness about one's answer, known as confidence judgments, and the objective accuracy of such judgments (Keren, 1991). The more closely overall confidence judgments match success rates, the better calibrated the individual is considered to be. Calibration is viewed as a measure of *precision*, assessing whether a specific confidence judgment matches performance exactly (Schraw, 2009). Calibration is therefore considered to be a measure of absolute accuracy.

Poor calibration can happen in both directions, including overconfidence and underconfidence. Overconfidence occurs when confidence judgments are greater than actual performance and this bias reflects a failure to detect errors (e.g., being confident in an incorrect response; Pallier, Wilkinson, Danthir, Kleitman, Knezevic, Stankov, & Roberts, 2002; Rinne & Mazzocco, 2014). In contrast, underconfidence occurs when confidence judgments are lower than actual performance and reflects the false detection of errors (e.g., lacking confidence in

correct responses; Pallier et al., 2002; Rinne & Mazzocco, 2014). Calibration has a direct impact on reasoning and decision-making by regulating and directing subsequent behaviors. For example, if performance on a math problem is judged as poor, an individual will likely rework the problem or double-check calculations, in order to increase performance. However, if performance on the same problem is judged as satisfactory, further regulation strategies will not be initialized. Overconfidence can therefore lead to a false sense of mastery resulting in allocating less cognitive resources than required to solve a problem. In contrast, underconfidence can lead to unnecessary and continued allocation of resources to a problem. Well-developed calibration skills are therefore critical for effective resource allocation. Individuals however tend to be poor judges of their own knowledge state, such that both children and adults are likely to display biased confidence judgments, with a tendency towards over- rather than underconfidence (Bjork et al., 2013; Soderstrom, Yue, & Bjork, 2015; Stanovich, West, & Toplak, 2016). Overconfidence has also been found to exist cross-culturally, with some minor variations in the degree of overconfidence. For example, overconfidence in general knowledge has been found to be typically stronger among Asian participants, compared to Western participants (Yates, Lee, & Bush, 1997), although both groups display a general tendency towards overestimating performance.

In the current study, participants indicated which response out of a possible four alternatives was correct and were then asked to assess their confidence in that response, on a scale from 25% (not confident at all, just guessing) to 100% (very confident). Resistance to overconfidence was equal to one minus the absolute difference between mean confidence and percentage correct across task items, such that higher scores reflected better performance (Bruine

de Bruin et al., 2007; see Method section for detailed description of calibration indices and corresponding formulas).

Resolution. Resolution, also known as discrimination or relative accuracy, is another measure of metacognitive monitoring accuracy. Resolution indicators can be used to assess whether a person is able to discriminate between correct versus incorrect performance (Koriat, 2012a) and are viewed as a measure of *consistency* (Schraw, 2009). Resolution is therefore considered to be a measure of relative, rather than absolute, accuracy.

Resolution is important for guiding people to effectively choose which materials to invest additional effort in to make the best use of their time (e.g., Destan & Roebers, 2015; Thiede, Anderson, & Therriault, 2003). Using the same example as above, calibration can inform whether performance on a math problem is poor, whereas resolution will direct to which parts of the problem are in need of additional effort investment. As evidenced by this example, both calibration and resolution have different functions regarding metacognitive monitoring and their measures are dissociable. That is, in the same experimental paradigm, calibration might be high, while resolution might be low, or vice versa (Maki, Shields, Wheeler, & Zacchilli, 2005; see Thiede, Mueller, & Dunlosky, 2015, for a review). There are various ways of measuring resolution including the Pearson correlation coefficient, the Discrimination Index (Schraw, 2009), the Goodman–Kruskal Gamma correlation (Nelson, 1984), and the confidence-judgment accuracy quotient (CAQ; Jackson & Kleitman, 2014). The choice of index depends on the literature (i.e., educational psychology literature versus decision-making literature), whether confidence judgments are elicited pre, during, or post task, and the design of the confidence judgment scale. Similar to other studies within the decision-making field, the current study measured resolution with the Goodman–Kruskal Gamma correlation. Gamma correlations

inform whether responses that receive greater confidence judgments are also the responses that result in greater performance and whether responses that receive lower confidence judgments are also the responses that result in poorer performance (Maki et al., 2005). Similar to the Pearson correlational coefficients a gamma value of "0" indicates a lack of relationship (i.e., lack of resolution) and greater values indicate stronger relationships (i.e., strong ability to discriminate between correct versus incorrect responses; see Method section for detailed description of resolution index).

Thus, these two aspects of monitoring accuracy, calibration and resolution, have different functions and inform on different aspects of metacognitive monitoring. That is, in the same experimental paradigm, calibration might be high, while resolution might be low, or vice versa (Koriat, Sheffer, & Ma'ayan, 2002). Yet, monitoring accuracy may not only be impacted by distinct measurement methods, but also by distinct knowledge domains (Erickson & Heit, 2015; West & Stanovich, 1997). The relationship between indices of monitoring accuracy and domain knowledge was therefore considered in the present study.

Domain-Generality versus Domain-Specificity

Overconfidence has been displayed across various domains, including predictions of sports outcomes (Ronis & Yates, 1987), reading comprehension (Glenberg & Epstein, 1987; Lin & Zabrucky, 1998), problem solving (Ackerman & Zalmanov, 2012; García et al., 2016), financial decision-making (Malmendier & Tate, 2008; Zacharakis & Shephard, 2001; Schrand & Zechman, 2012) and general knowledge (Koriat, Lichtenstein, & Fischhoff, 1980; Yates, Lee, & Bush, 1997). Regarding resolution, some domains are characterized by low resolution (e.g., reading comprehension; Thiede, Griffin, Wiley, & Anderson, 2010), while others typically show high resolution (e.g., problem solving; Ackerman & Zalmanov, 2012). Despite the breadth of

studied domains, these domains have been largely explored in isolation and have rarely been examined in parallel in the same study (Erickson & Heit, 2015). As such, an important consideration is whether calibration and resolution are general and independent of the domain under study, or whether they are specific and closely related to content knowledge.

According to the domain-general hypothesis, being able to endorse confidence judgments that accurately match one's performance reflects a skill, or trait, that one can apply across different areas of functioning (e.g., Erickson & Heit, 2015; Kleitman; 2008; Kleitman & Stankov, 2001; Pallier et al., 2002; Veenman & Verheij, 2003). In particular, when considering several tasks, a picture of the calibration bias being a characteristic of individuals has consistently emerged. For example, Schraw, Dunkle, Bendixen, and Roedel (1995) found that students' accuracy scores were correlated across five different subject domains, concluding that calibration is governed by general metacognitive processes, independent of content knowledge. Similarly, Jackson and Kleitman (2014) found that confidence ratings on a medical decision task and cognitive ability indicators measured by solving Raven Matrices and a vocabulary questionnaire showed a robust bias across tasks. People who tend to be overconfident on one type of task therefore tend to be overconfident on other types of tasks.

In contrast, the domain-specific hypothesis suggests that calibration reflects the ability to assess specific content knowledge and that these abilities can vary depending on individuals' comfort with the content area in question. Studies in support of the domain-specific view have demonstrated that calibration and resolution are in fact inconsistent across subject areas and can be sensitive to task domain (e.g., Glaser, 1991).

Findings from research that has explored the question of domain-generality versus domain-specificity (Erickson & Heit, 2015; Kelemen, Frost, & Weaver, 2000; Klayman,

González-Vallejo, & Barlas, 1999; Pallier et al., 2002; Perfect, 2004; Scott & Berman, 2013; Veenman, Van Hout-Wolters, & Afflerbach, 2006; Veenman & Verheij, 2003; West & Stanovich, 1997) have yielded inconclusive findings. In an effort to consolidate these two views, some researchers have suggested a developmental model in which abilities first emerge as domain-specific in early childhood and then eventually become general and applied across domains (Erickson & Heit, 2015; Veenman & Spans, 2005). In general, metacognitive abilities are thought to formally appear between the ages of eight and ten, with expected age-related improvements taking place during middle and late childhood, and more developed abilities being fully formed by adolescence and early adulthood (Lockl & Schneider, 2006; Veenman et al., 2006). According to the mixed-model view, emerging adults (18-29 years of age) would be expected to display domain-general abilities, given that they are at the later end of this developmental trajectory. Domain-specificity however may also be apparent, depending on the domain in question and the reliance on content knowledge.

One complicating factor for comparing across different domains is item difficulty (Koriat, 2012a). For example, one cannot fairly compare calibration in general knowledge and math calculation if the test items in each respective domain are not matched for difficulty. Item difficulty has been shown to be related to calibration, termed the hard-easy effect (Juslin, Winman, & Olsson, 2000; Lichtenstein & Fischhoff, 1977), where easy test items tend to result in high accuracy and underconfidence but difficult test items result in low accuracy and overconfidence. Similarly, those with low knowledge are known to be more overconfident than those with high knowledge (see Miller & Geraci, 2011, for a review). These findings highlight the importance of using tasks with similar difficulty levels for when comparing monitoring accuracy across domains, in order to control for these differences (e.g., Erickson & Heit, 2015).

Thus, the experimental tasks developed for this study were piloted in order to try to match task difficulty across domains.

Three Domains. To contribute to the understanding of domain similarities and differences, calibration and resolution were examined in the well-studied domain of general knowledge (Lichtenstein, Fischhoff, & Phillips, 1977; Yates, Lee, & Bush, 1997) and compared to two understudied domains in this context: financial calculation and probability calculation. For each of these domains, the research has been focused either on calibration or on resolution, but no study to date has had both measures and none have compared these measures within individuals across these domains.

General knowledge. General knowledge refers to accumulated and informal knowledge across several topics, as opposed to in-depth knowledge about one particular topic. General knowledge questions are the most well-known and widely-used means of measuring calibration (Lichtenstein et al., 1977; Yates et al., 1997). Overall, individuals tend to display overconfidence in their general knowledge (Bruine de Bruin et al., 2007; Stanovich, West, & Toplak, 2016; West & Stanovich, 1997; Yates et al., 1997). Overconfidence in the domain of general knowledge therefore reflects a well-established bias in the confidence judgement literature.

Numeracy. There has never been a time in human history that we have been faced with so much numerical information to process in our everyday lives. Making choices in finances and decisions involving probability information have become very prevalent in our modern society, and the complexity of these decisions have increased significantly with the availability of more financial aids than ever and more data available in domains such as probability outcomes of medical treatments. Given the omnipresence of numerical thinking and the importance of well-developed calculation skills, literacy in the twenty-first century has been proposed as not only

including the ability to read and write, but also the ability to apply numerical thinking (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, and Woloshin, 2007). For these reasons, both domains of financial calculation and probability calculation were explored in this study.

Financial calculation. Financial literacy reflects the ability to use numeric information and to make decisions regarding financial planning, wealth accumulation, debt, and pensions (Lusardi & Mitchell, 2014). Financial calculation represents the computational process through which financial knowledge is applied. In real-world financial problems, both literacy and calculation skills are simultaneously required. For example, calculating an interest rate requires interest rate knowledge in conjunction with basic computational abilities. In general, it has been reported that individuals tend to overestimate their ability to apply numeracy skills to financial contexts (Lusardi & Mitchell, 2014). This overestimation has been found to be widest for emerging and older adults, two time points in which financial numeracy may be most important (Chen & Volpe, 1998; Lusardi & Mitchell, 2014). That is, older adults are at the end of their financial life cycle, when decisions regarding debt, savings, and pension are most likely to take place. Similarly, young adults are in the beginning of their financial life cycle, when decisions regarding credit cards, loans, investments, and mortgages are more likely to be made. As such, the two time points in which individuals are making the most critical financial decisions coincide with the two time points in which individuals are most likely to display overconfidence, and therefore make biased financial-related decisions.

Research that has explored financial literacy in emerging adult samples has concluded that young adults lack the appropriate knowledge and skills of basic financial concepts, including the ability to understand and calculate inflation, risk diversification, and interest rates (Lusardi, Mitchell, & Curdo, 2019). Poor financial literacy can have negative consequences on financial

decision-making, such as increased challenges with debt (Lusardi & Tufano 2015), increased difficulty accumulating and managing wealth (Hilgert, Hogarth, & Beverly, 2003), and a decreased likeliness of retirement planning (Lusardi & Mitchell, 2007). For these reasons, we chose to study financial calculation in the current sample of undergraduate students. Emerging adults are in the developmental stage where they are leaving home, expanding financial responsibilities (e.g., rent, tuition, groceries, travel) and relying on credit and loans as they generally have not had sufficient opportunity to accumulate savings. Despite this increase in the need for financial competency, most emerging adults display low level finance skills and do not possess adequate financial knowledge (Chen and Volpe, 1998; Mandell, 2008). To assess this domain, financial calculation items that reflect ecologically and developmentally valid financial problems most relevant to emerging adults' everyday financial decisions were developed, including conversion currency rates, costs and savings calculations, credit card interest rate calculations, and calculating bank interest rates.

Probability calculation. A second numeracy domain considered in this study involved calculating probabilities in real-world contexts. From interpreting news headlines to understanding information about medical tests and procedures, probabilistic thinking has become a skill required in our everyday lives (Gigerenzer et al., 2007). Probability reasoning (e.g. "What are the chances that…") plays a central role in decision-making, particularly in situations involving risk estimation. Research has demonstrated however that when information is presented via probabilities, individuals display significant difficulties in accurately interpreting risk level (Gigerenzer et al., 2007). In general, less numerate individuals have been found to have less wealth and to display poorer health (Peters & Bjalkebring, 2015). Individuals who display poor numeracy skills have also been found to be less accurate in comprehending medical

information and to make poorer health-related decisions (Gigerenzer et al., 2007). For example, Miron-Shatz, Hanoch, Doniger, Omer, and Ozanne, (2014) examined the impact of numeracy upon willingness to pay for BRCA1/2 testing in high-risk women reporting a family history of breast and ovarian cancer, by exploring how objective and subjective probability skills relate to willingness to pay for genetic testing. Study results found that subjective, but not objective, numeracy skills were correlated with willingness to pay (Miron-Shatz, Hanoch, Doniger, Omer, & Ozanne, 2014) As such, it was participants' confidence in their ability to understand probabilistic information (i.e., subjective numeracy), and not their actual mathematical ability (i.e., objective numeracy), that had a direct impact on their health-related decision-making (Miron-Shatz et al., 2014).

By exploring the question of domain generality versus specificity within these three domains of general knowledge, financial calculation, and probability calculation, the current study aimed to better understand whether accuracy is content- or person-dependent. However other factors such as cognitive abilities, aggregate judgments of task performance, difficulty, and effort, as well as personal dispositions may also impact individuals' ability to effectively monitor their performance (Ackerman, 2019; Ackerman & Thompson, 2017; Thompson, 2009; West & Stanovich, 1997). Thus, a better understanding of how individual level differences impact metacognitive monitoring was also warranted.

Correlates of Monitoring Accuracy

A variety of cues are thought to underlie the formation of metacognitive judgments, such as task specific information and ones' own subjective experience. Although the most commonly studied individual difference index has been cognitive abilities, relatively little is known about the role of other person-related factors on confidence judgments in reasoning and problem-

solving contexts, including the role of metareasoning processes (see Ackerman & Thompson, 2017, for a review) and individual and dispositional differences (Thompson, 2009; West & Stanovich, 1997). In a review by Ackerman (2019) three levels of cues for metacognitive judgments were identified: (1) self-perceptions (i.e., domain knowledge, personality traits such as anxiety, "I am good/bad at this type of task"); (2) task characteristics (i.e., difficulty, format); and (3) momentary experiences (i.e., perceived ease of processing, familiarity). The current study, therefore, not only explored variability of confidence judgments across domains, but also across differences in cognitive abilities, task-specific judgments, and self-perceptions.

Cognitive abilities. Research exploring person-related differences in confidence judgments has largely focused on two indicators of cognitive abilities: intelligence and executive functions (Stanovich & West, 1998). Intelligence is usually measured through the assessment of both crystallized and fluid abilities, whereas executive functions have been reflected by measuring working memory skills (Komori, 2016). Performance on both general knowledge and calculation tasks have been found to be positively associated with intelligence (Stanovich & West, 1998; Peters & Bjalkebring, 2015) and with working memory (Komori, 2016).

Performance on knowledge calibration paradigms, such as the overconfidence index, has also been associated with cognitive abilities, such as self-reported SAT scores (Stanovich, West, & Toplak, 2016) and verbal and nonverbal intellectual abilities (Bruine de Bruin, Parker, & Fischhoff, 2007; Stanovich & West, 1998; Stanovich et al., 2016). Better calibration is therefore associated with better cognitive abilities. The explanation for this association has been attributed to the simulating and hypothetical reasoning that is required for successful performance on several decision-making and rational thinking tasks (Stanovich, 2011). That is, the mechanisms of cognitive decoupling allow individuals to simulate alternative worlds and to consider

hypothetical scenarios that are not in the immediate environment, requiring engagement of analytic processes to construct these scenarios. If poor calibration results from individuals' exposure to environmentally unrepresentative knowledge, then at least some cognitive decoupling would be required to achieve better calibration (West & Stanovich, 1997). Cognitive decoupling is often indexed by individual differences in general cognitive abilities, such as intellectual abilities and executive functions (Stanovich, 2009). Thus, individual differences in cognitive decoupling mechanisms should be associated with calibration, which was expected to be replicated using the present study's three cognitive tasks (Bruine de Bruin et al., 2007; Stanovich & West, 1998).

Similarly, resolution has also been associated with cognitive abilities, particularly with depth of processing on reading comprehension tasks (Thiede et al., 2003). Studies that have examined methods to encourage greater depth of processing in learners have shown that greater processing improves resolution (Anderson & Thiede, 2008; Fukaya, 2013; Thiede, Dunlosky, Griffin, & Wiley, 2005; Thiede, Wiley, & Griffin, 2011). Recently, calibration of reading comprehension and problem solving was also found to be improved by task characteristics which call for depth of processing, although this improvement was more effective in computerized environments than in paper-based environments (Lauterman & Ackerman, 2014; Sidi, Spigelman, Zalmanov, & Ackerman, 2017). Thus, resolution was also expected to be associated with cognitive abilities in the current study.

Task-specific judgments. Task-specific judgments relate to the information and beliefs individuals entertain about factors impacting overall performance in a task (Ackerman, 2019). Such task characteristics can include task-related abilities, task difficulty, and amount of processing effort required by the task (Ackerman, 2019). In order to gain a better understanding

of associations between task-specific judgments and monitoring accuracy, a series of aggregated judgments were collected, including pre-and postdictive confidence judgments, as well as ratings of task difficulty, required effort, and feeling of effort.

Predictive and postdictive confidence judgments. Predictive confidence judgments reflect individuals' subjective confidence in their knowledge or abilities, prior to a task, whereas postdictive judgments reflect individuals' subjective confidence in overall performance, after completing a task. Predictive judgments are hypothesized to guide allocation and regulation of mental resources for a given task (see Ackerman & Thompson, 2017). If an individual estimates poor confidence in their ability to complete a task, they will likely invest more effort into the task, in order to be successful. As such, predictive confidence judgments require individuals to estimate a priori performance based on stored memory structures relevant to a given domain. Predictive judgments are considered to be theory-based, guided by individuals own theories about their abilities in a certain domain, as well as about the characteristics of the domain (Ackerman, 2019; Koriat, 1997).

In contrast, postdictive confidence judgments require individuals to use their perceived performance on each task as a reference point for their rating. Postdictive judgments are considered to be experience-based, guided from the real-time experience of having recently completed the task and from elicited cues, such as ease of processing (Ackerman, 2019; Koriat, 1997). Both predictive and postdictive confidence judgments of performance have been shown to be positively associated with success rates across different domains (Erickson & Heit, 2015) and were considered in this study.

Task difficulty, effort, and feeling of effort. In addition to judgments of performance, participants were asked to rate task difficulty, effort required, and their affective reaction to

engaging in mental effort, as aggregated judgments following each experimental task (Finn, 2010; Hsu, Eastwood, & Toplak, 2017; Hsu et al., 2018).

Task difficulty has been shown to impact confidence ratings (Juslin et al., 2000; Klayman & Soll 1999). The "hard-easy effect" is a well-known occurrence in calibration research and refers to the covariance between confidence judgments and task difficulty (Juslin et al., 2000), where individuals tend to overestimate the probability of success when a task is perceived as hard, and to underestimate the probability of success when a task is perceived as easy. That is, harder items are more likely to promote overconfidence, whereas easier items are more likely to promote underconfidence (Juslin et al., 2000). A similar negative correlation between confidence, indices of monitoring accuracy, and task difficulty was expected in the current study.

In contrast, effort, and specifically the subjective feeling of effort as it relates to monitoring accuracy, has received less empirical attention. Studies that have explored the role of mental effort in confidence judgments have revealed somewhat mixed findings. Some findings support that individuals tend to attribute lower confidence ratings to tasks perceived as effortful, as the increased investment of effort is acknowledged and taken into account when estimating performance (for review, see Ackerman 2019). Other research (Chen, 2002) suggests that the experience of increased effort exertion leads to higher confidence ratings, due to an overestimation of performance (e.g., "If I worked this hard, I must be right"). Regarding the feeling of effort, individuals high in mental effort tolerance have been found to rate tasks as less difficult (Dornic, Ekehammar, & Laaksonen, 1991). That is, individuals with a greater ability to tolerate the aversiveness of increased effort expenditure are more likely to judge a task as less difficult. Hsu and colleagues (2018) also experimentally distinguished between the phenomenological experience of mental effort expenditure and that of discomfort, supporting

that both difficulty and discomfort were correlated, yet distinct from one another. Considering the relationship between overconfidence and difficulty described in the hard-easy effect, it can be hypothesized that poorer monitoring accuracy would not only be related to increased task difficulty, but also to increased effort and discomfort.

As such, the present study asked participants to provide a separate rating of task difficulty, as well as required and experienced effort, for each task. It was important that these three ratings were associated with their respective task so that participants had a specific reference point for their subjective ratings in each domain. For each task, correlations between monitoring accuracy indices (i.e., calibration and resolution) and these additional aggregated judgments were examined.

Self-perceptions. Individual self-perceptions were the final set of monitoring accuracy correlates explored in this study. Self-perceptions represent individuals' beliefs about their traits, abilities, and knowledge and reflect one's confidence in the ability to succeed on a given task or domain (Ackerman, 2019). However, the role of self-perceptions on indices of monitoring accuracy, such as calibration and resolution, have been generally less explored in the metacognitive research. Self-concept (defined as one's perception of the self) and anxiety (defined as one's physio-emotional reaction when thinking about or performing a task; Stankov, Lee, Luo, & Hogan, 2012) have both been identified as non-cognitive predictors of student achievement (Lee, 2009). Given the positive relationship between academic achievement and metacognitive monitoring (Hacker et al., 2008), as well as the cognitive nature of the present study tasks, self-perceptions of both academic self-concept and cognitive test anxiety were explored as potential monitoring accuracy correlates.

Academic self-concept. Self-concept represents a collection of descriptive and evaluative beliefs and perceptions about the self (e.g., "I am bad at mathematics"; Bong & Skaalvik, 2003; Kröner & Biermann, 2007). Kröner and Biermann (2007) suggested that when individuals are asked to form a confidence judgment on a task they are unsure of, they are likely to base confidence judgments on how they perceive themselves to perform on similar tasks. This is what Ackerman (2019) refers to as a heuristic cue of self-perception, from which confidence estimates are formed. Given that an individual's self-concept can change depending on the dimension being considered (e.g., academic self-concept versus social self-concept), measures of self-concept should be closely related to performance domain. Academic self-concept, defined as knowledge and perceptions about the self in academic, achievement, and testing situations, was considered in this study, using a short form of the Academic Self Concept Scale (ASCS; Reynolds, Ramirez, Magriña, & Allen, 1980). The ASCS captures the academic facet of general self-concept and includes items surveying self-beliefs about academic achievement (e.g., "*All in all, I am proud of my grades in college*.") and abilities (e.g., "*I often expect to do poorly on exams*").

Cognitive test anxiety. Similar to self-concept, anxiety can also be context-dependant, such that anxiety measures should be closely related to performance domain. Studies examining the role of anxiety on performance have noted that it is the cognitive component of anxiety, such as negative thoughts, worries, and rumination, that is most detrimental and that accounts for the largest proportion of variability in performance decline (Cassady & Johnson, 2002). That is, anxiety can negatively impact performance through the disruption and overloading of cognitive processes, such as working memory, as well as through the increased reliance on avoidance mechanisms. Given the current study's focus on knowledge and processing domains, cognitive

anxiety, defined as cognitive reactions and internal dialogues regarding evaluative and performance situations (Cassady & Johnson, 2002), was explored.

Research on how anxiety may impact confidence judgments is relatively sparse. Studies that have explored this association have provided some support for the relationship between increased anxiety and decreased monitoring accuracy (Everson, Smodlaka, and Tobias, 1999; Legg & Locker, 2009). For example, a study by Legg and Locker (2009) found that metacognitive skill moderated the relationship between math anxiety and performance, such that performance decreased as a function of increased anxiety. Similarly, Everson, Smodlaka, and Tobias (2009) demonstrated that increased test-taking anxiety was related to decreased metacognitive motoring on a reading comprehension task, independent of actual reading ability. In contrast, some research demonstrates that certain levels of state anxiety can be helpful to high-risk situations, given that feelings of uncertainty can lead to the allocation of greater effort and attentional resources (Macher, Papousek, Ruggeri, & Paechter, 2015). Considering these differing views, Morsanyil, Cheallaigh, and Rakefet (2019) have proposed that a new area of investigation connecting metacognition literature with anxiety research be explored. Such investigations can potentially help better our understanding of the metacognitive differences between anxious and non-anxious learners (e.g., task-related approach versus avoidance behaviours) and whether anxious learners are more or less prone to overconfidence, compared to non-anxious leaners (Morsanyil, Cheallaigh, & Rakefet, 2019).

Based on this literature, the current study predicted that individuals with poorer academic self-concept and greater cognitive anxiety would display poorer calibration and resolution scores. Given that self-concept and anxiety have been found to be dissociable yet related (Lee, 2009), a positive correlation between both scales was also expected.

Real World Outcomes of Monitoring Accuracy

Reviewed so far is how metacognitive accuracy may relate to domain, cognitive abilities, task-specific judgments, and self-perceptions. A final consideration would be to better understand how such skills relate to real-world consequences (Bruine de Bruin et al., 2007). Overconfidence has been associated with a number of real-world outcomes, such as several risk behaviors, externalizing behaviors, and substance use (Parker & Fischhoff, 2005) and several negative decision outcomes that vary in severity, from throwing out food to having a mortgage or loan foreclosed (Bruine de Bruin, Parker & Fischhoff, 2007). Efficient use of metacognitive monitoring skills, however, is not only relevant to decision-making research, but also to broader learning contexts. Particularly, in the current era of high-stakes testing, test performance has become increasingly important with direct effects on educational placement and admission, choice of major or specialization, and graduation (Hacker, Bol, & Keener, 2008). Research has demonstrated that the development and use of metacognitive skills are critical for student success across a variety of domains, including mathematics (Pugalee, 2001), reading (Presley, 2002), and writing (Pugalee, 2001). Monitoring accuracy has also been linked to academic achievement (Hacker et al., 2008). Following this line of research, a final, exploratory goal of the current study was to consider potential real-world metacognitive monitoring outcomes. Specifically, relationships between indices of monitoring accuracy, academic average, and self-reported learning challenges were investigated. Monitoring accuracy, cognitive abilities, and self-perceptions were also explored as predictors of achievement and learning outcomes.

Present Study

Goals of the current study were to build upon and extend existing research on confidence judgments (for Study Overview, See Figure 1).

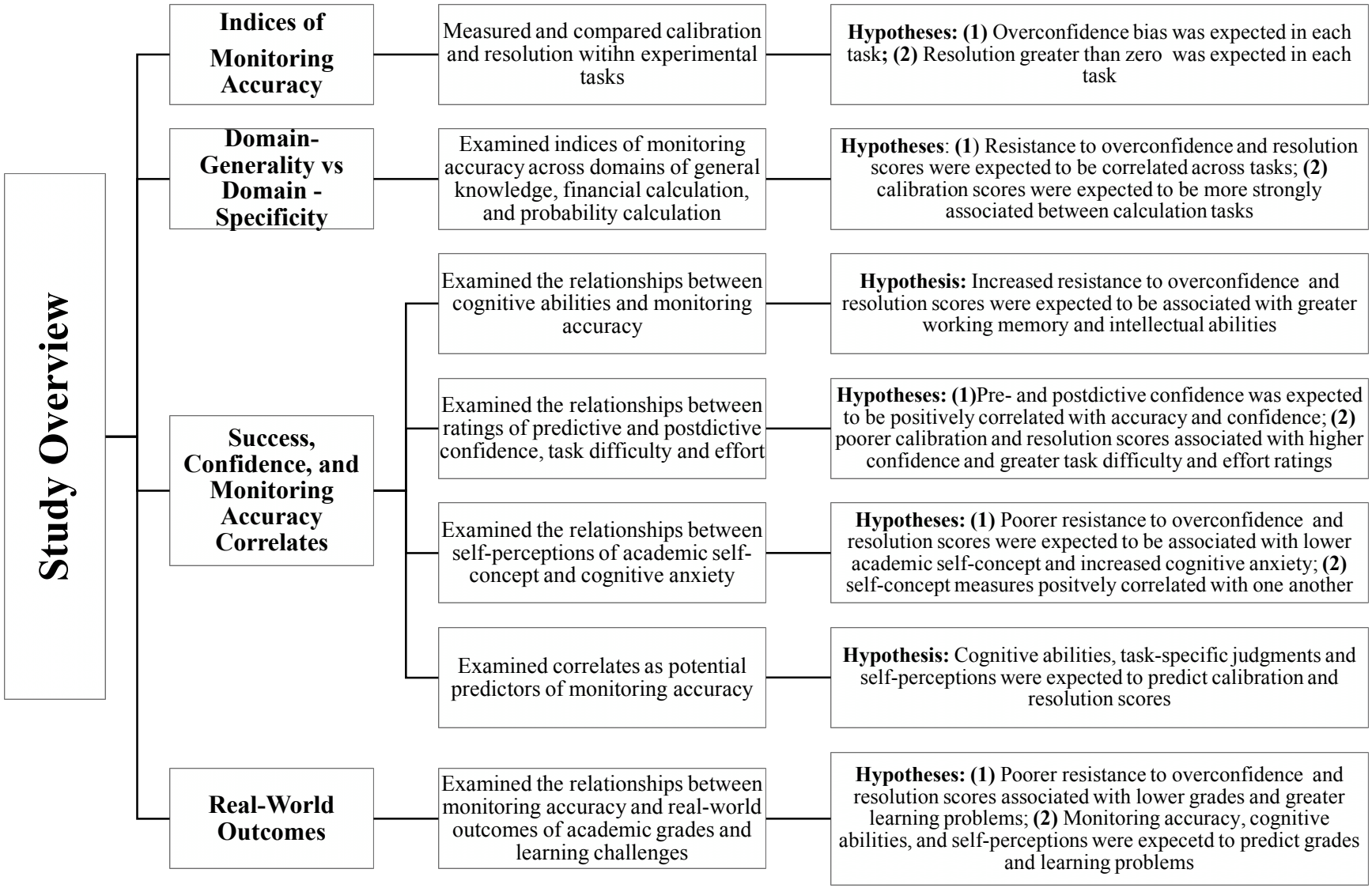


Figure 1. Present study overview. This figure outlines the study goals, research questions, and hypotheses.

First, both calibration and resolution, two related yet distinct measures of monitoring accuracy, were compared within and across domains. Second, the question of domain-generality versus domain-specificity was explored through the development and use of novel tasks that measured real-life financial decision-making and probability calculations. Third, relationships between cognitive abilities, including intellectual abilities and working memory, and monitoring accuracy were assessed. Fourth, task-specific judgments, including predictive and postdictive confidence ratings, post-task ratings of task difficulty, effort required, and feeling of effort, as well as self-perceptions of cognitive anxiety and academic self-concept were explored as potential monitoring accuracy correlates. Last, the relationship between monitoring accuracy and real-world outcomes, including academic grades and self-reported learning problems, were explored. Research hypotheses were as follows:

Hypotheses relating to indices of monitoring accuracy.

Hypothesis 1. As has been reported in the literature, an overconfidence bias was expected across all experimental tasks (Bjork et al., 2013; Soderstrom et al., 2015; Stanovich et al., 2016).

Hypothesis 2. Given that a gamma value of "0" indicates a lack of relationship, and therefore lack of resolution, resolution was expected to be greater than zero across all tasks.

Hypotheses relating to domain-generality versus domain-specificity.

Hypothesis 3a. Resistance to overconfidence scores were expected to be correlated across experimental tasks. That is, better calibration on one task was expected to be positively correlated with better calibration on the other tasks (Jackson & Kleitman, 2014; Schraw et al., 1995).

Hypothesis 3b. Even though domain-generality of calibration was hypothesized, in order to take into account the literature supporting domain-specificity (Glaser, 1991), it was hypothesized that resistance to overconfidence scores between similar domains (i.e., financial and probability calculations) would reflect stronger correlations compared to resistance to overconfidence scores between less similar domains (i.e., general knowledge versus calculations).

Hypothesis 4. Although resolution has been found to be less consistent (Jackson et al., 2016) and has been less explored across domains, it was hypothesized that resolution scores would also be correlated across experimental tasks, such that better resolution on one task would be positively correlated with better resolution on the other tasks.

Hypotheses relating to success, confidence, and monitoring accuracy correlates.

Hypothesis 5. Based on previous research (Bruine de Bruin et al., 2007; Stanovich & West, 1998; Stanovich et al., 2016), working memory and intellectual abilities were expected to be positively associated with increased calibration and resolution scores, across general knowledge, financial calculation, and probability calculation tasks.

Hypothesis 6a. Predictive confidence judgments were expected to be positively correlated with accuracy and confidence, as has been previously reported in the literature (Erickson & Heit, 2015).

Hypothesis 6b. Similarly, postdictive confidence judgments were also expected to be positively correlated with accuracy and confidence (Erickson & Heit, 2015).

Hypothesis 7. Poorer resistance to overconfidence scores were expected to be associated with higher confidence and increased ratings of task difficulty, required effort, and feeling of effort (Ackerman, 2019; Chen, 2002; Hsu et al., 2018).

Hypothesis 8. Poorer resolution scores were expected to be associated with higher confidence and increased ratings of task difficulty, required effort, and feeling of effort (Ackerman, 2019; Chen, 2002; Hsu et al., 2018).

Hypothesis 9. In line with previous studies (Everson et al, 1999; Legg & Locker, 2009), poorer calibration scores were expected to be associated with poorer academic self-concept and increased cognitive anxiety.

Hypothesis 10. Similar to calibration, poorer resolution scores were also expected to be associated with poorer academic self-concept and increased cognitive anxiety.

Hypothesis 11. Self-perceptions of academic self-concept and cognitive anxiety were expected to be positively correlated with one another (Lee, 2009).

Hypothesis 12. Correlates of monitoring accuracy including cognitive abilities, task-specific judgments, and self-perceptions were expected to predict calibration scores.

Hypothesis 13. Correlates of monitoring accuracy including cognitive abilities, task-specific judgments, and self-perceptions were expected to predict resolution scores.

Hypotheses relating to real-world outcomes of monitoring accuracy.

Hypothesis 14. Poorer resistance to overconfidence scores were expected to be associated with lower grades and greater self-reported learning problems (Hacker et al., 2008).

Hypothesis 15. Poorer resolution scores were predicted to be associated with lower grades and greater self-reported learning problems.

Hypothesis 16a. Indices of monitoring accuracy, cognitive abilities, and self-perceptions were expected to predict real-world outcomes of academic achievement (Hacker et al., 2008).

Hypothesis 16b. Indices of monitoring accuracy, cognitive abilities, and self-perceptions were expected to predict real-world outcomes of learning challenges (Hacker et al., 2008).

Method

Participants

The current study was a subset of a larger project investigating confidence judgments in emerging adults. Participants consisted of 153 undergraduate students attending York University, in Toronto, Canada. The data were collected during a single term. Participants were compensated two undergraduate research participant pool credits for their time.

Of the 153 participants, 17 participants were excluded from data analyses (i.e., four participants reported English as a second language and identified language-related difficulties in understanding and/or completing study tasks; 11 participants did not complete full task battery due to time constraints; one participant did not complete tasks correctly; and one participant did not have adequate variability to calculate a resolution score). The remaining sample therefore consisted of 136 participants (102 females; 34 males) whose ages ranged from 18 to 30, with the mean age being 20.43 (*SD* = 2.77). Most participants were in their first year of the undergraduate program (56.6%), reported their ethnicity as White/European (31.6%), identified English as first language (61.76%; participants who did not indicate English as a first language reported between six to 28 years of English speaking), and reported a current academic average between 70-79% (45.6%). Table 1 displays a summary of all participant demographic characteristics.

Table 1
Descriptive Statistics for Socio-Demographic Variables

Socio-Demographic Variable	Frequency	Percent	*N*
Age (years)			129
18-21	99	76.70	
22-25	22	17.0	
26-30	8	6.20	

Gender			136
Male	34	25.0	
Female	102	75.0	
Year in university			136
1st year undergrad	77	56.61	
2nd year undergrad	28	20.58	
3rd year undergrad	18	13.24	
4th year undergrad	8	5.88	
5th year undergrad	2	1.47	
Post-BA Continuing	1	0.74	
Other	2	1.47	
Ethnicity			136
White/European			
Black	43	31.61	
Asian	12	8.82	
South-Asian	23	16.91	
Arab	33	24.26	
Latino-Hispanic	5	3.68	
Other	6	4.41	
	14	10.29	
English as first language			134
Yes	84	61.76	
No	50	36.76	
Mother highest level of education			133
Less than 7th Grade	4	3.00	
Junior high/Middle school	5	3.75	
Partial high school	8	6.01	
High school graduate	22	16.54	
Partial college/university	26	19.54	
College/university education	48	36.09	
Graduate/Professional degree	20	15.03	

Father highest level of education			133
Less than 7th Grade	6	4.51	
Junior high/Middle school	2	1.50	
Partial high school	5	3.75	
High school graduate	23	17.29	
Partial college/university	12	9.02	
College/university education	49	37.84	
Graduate/Professional degree	36	27.06	
			132
Family income			
Well below average	2	1.51	
Below average	16	12.12	
Average	75	56.81	
Above average	34	25.76	
Well above average	5	3.78	
Current academic average			130
50-59%	5	3.84	
60-69%	36	27.69	
70-79%	62	47.69	
80-100%	27	20.76	
Self-reported socio-emotional difficulties			132
No	74	55.06	
Yes – minor difficulties	43	32.57	
Yes – definite difficulties	14	10.61	
Yes – severe difficulties	1	0.75	
Self-reported learning difficulties			132
No	80	60.60	
Yes – minor difficulties	45	34.09	
Yes – definite difficulties	6	4.54	
Yes – severe difficulties	1	0.75	

Measures

Demographics. All participants were asked to report their age, gender, academic year, ethnicity, current academic average, parental level of education, family income, and perceived

level of socio-emotional and academic difficulties (i.e., no difficulties, minor difficulties, definite difficulties, severe difficulties; see Table 1 for descriptive statistics). Appendix A contains the demographic form used in the current study.

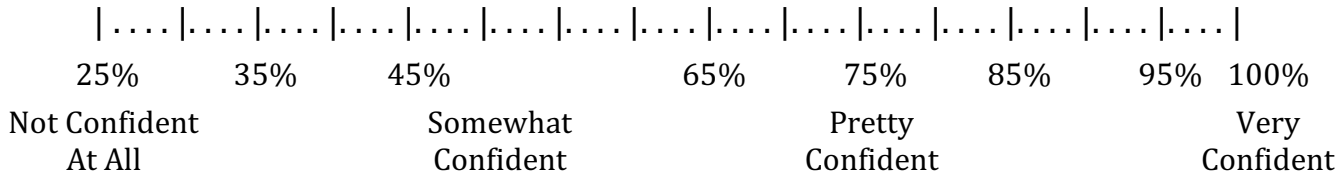
Experimental tasks. Three monitoring accuracy tasks in different domains, including general knowledge, financial calculation, and probability calculation were developed for the purpose of this study. All three tasks had the same multiple-choice format with four alternatives and the same confidence rating scale. The tasks were developed to be matched on total number of items, success rates, and confidence ratings. These measures were first piloted using a paper-and-pencil version of the tasks with a sample of 18 undergraduate and graduate students to ensure that the instructions were clear and that items were not too easy or too difficult. Overall success rate was greatest for the Financial Calculation Task (82.18%), followed by the Probability Calculation Task (71.06%), and the General Knowledge Task (60.38%). Given that the pilot sample was comprised of several psychology graduate students, it was expected that success rate would be lower in the study sample, particularly on calculation tasks, given that graduate students had all completed PhD level statistical instruction and thus accumulated greater numeracy knowledge and skills. Items were therefore judged to be at a moderate level of difficulty, which permitted adequate variability in performance to calculate the different monitoring accuracy indices. Appendix B contains the pilot data for each task including item-by-item success rate.

Confidence rating scale. Confidence judgments were measured on an item-by-item basis with each item on the general knowledge, financial calculation, and probability calculation tasks being followed with a confidence rating scale. All tasks contained four multiple choice item responses, such that confidence judgments were marked on a scale ranging from 25% (i.e.,

chance of responding correctly by guessing) to 100% (i.e., absolute confidence in answer provided). This confidence scale was based on previous confidence rating scales used in the decision-making literature (Bruine de Bruin et al., 2007). Details on how to use the confidence rating scale was presented to participants both orally and visually, with examiners providing instructions via projected PowerPoint slides. Practice trials were also presented prior to each task, to provide feedback and to ensure that participants understood how to complete the confidence rating scales.

The general knowledge task was presented first with elaborated instructions for explaining the responses required and examples for completing the confidence ratings. The instructions were as follows:

> In this survey, you will be asked to answer general knowledge questions. There will be four options for each question (A, B, C, and D). Please choose the option that you think is correct for each question.
>
> Following each question, you will be asked to indicate how sure you are about your answer on a scale from 25% (*Just Guessing, Not Confident at All)* to 100% (*Very Confident*). Since there are 4 options for each question, 25% would mean you are just guessing and you are not very confident. Then, 100% would mean you are very confident and certain. Respond to each confidence rating immediately after answering the question. Spend no more than 5 seconds on each confidence rating.
>
> Here is an example:
>
> **What is the name of a young sheep?**
>
> A. Lamb
> B. Calf
> C. Baby
> D. Steer



Please write the number that you just circled: _____

Participants were instructed to first select one of the options (A, B, C, or D) and then to indicate how confident they were in their answer by circling the number on the scale and also writing down the actual number that they circled. Following these instructions participants were provided with examples of three potential responses on the confidence scale. In this example question, option A (i.e., Lamb) was the correct response. Using the response of "Lamb" (i.e., Choice A) for these examples, participants were instructed that a response of 25% would indicate they are guessing, a response of 100% would indicate that they are 100% sure, and a response of 50% would indicate that they are somewhat confident. These examples were displayed visually on the response form.

General Knowledge Test. The General Knowledge Test assessed knowledge about a broad range of facts and subjects. Each multiple-choice item on the General Knowledge Test corresponded to four possible alternatives. The items were selected from published general knowledge norms reported in Tauber, Dunlosky, Rawson, Rhodes, and Sitzman (2013; which were updated from norms first published by Nelson & Narens, 1980). These norms provided recall accuracy and confidence ratings; items selected reflected a wide range of difficulty and confidence ratings, while avoiding floor and ceiling effects. Each question was presented separately, followed by the confidence rating scale. An example item is (* = correct response):

What is the name of the largest ocean on earth?

A. North Sea
B. Atlantic
C. Pacific*
D. Indian

Initially, the General Knowledge Test included 24 items. Yet, to calculate whether participants can discriminate between correct versus incorrect responses, adequate variability is needed, which is problematic with items that reflect either ceiling or chance performance

(Ackerman & Goldsmith, 2011; Leming & Flau, 2014). Thus, as is standard practice in conventional resolution research, floor items with an overall success rate of less than 10% or ceiling items with an overall success rate greater than 90% were removed. Although this is not common practice in calibration research, this was done for all calculations, in order to allow comparison between monitoring accuracy indices. As displayed in Table 2, eight items (three below 10% success rate and five above 90% success rate) were removed on the General Knowledge Test, leaving 16 items on this task for analysis[1]. Cronbach's alpha for the General Knowledge Test was 0.21. Appendix C contains test questions, item-by-item success rate, and removed items.

Table 2
Success Rate Variability for General Knowledge Test

Success Rate (%)	Frequency	Percent
0-10	3	12.50
10-20	2	8.33
20-30	1	4.17
30-40	2	8.33
40-50	4	16.67
50-60	4	16.67
60-70	1	4.17
70-80	1	4.17
80-90	2	8.33
90-100	5	20.83

Note. Items with overall success of less than 10% or greater than 90% were removed.

[1] For the general knowledge task, due to a technical printing error, 38 participants missed completing two items on the task. Overall accuracy between participants who missed the two items and the remaining 98 participants was compared. There were no differences in overall accuracy. Thus, scores of these 38 participants were pro-rated for the statistical analyses.

Financial Calculation Test. The Financial Calculation Test assessed individuals' knowledge and calculation abilities related to financial decision-making. Each multiple-choice item on the Financial Calculation Test corresponded to four possible alternatives and was presented separately, followed by the confidence rating scale. Test items were chosen based on real-world scenarios and calculations relevant to emerging adults. Test items include conversion currency rates (6 items), costs and savings calculations (6 items), credit card interest rates on unpaid balances (6 items) and bank interest rates (6 items). An example item is (* = correct response):

> Janice has a job where she earns $2000 per month. She spends $900 for rent and $150 for groceries each month. She also spends $250 per month on transportation. If she budgets $100 each month for clothing, $200 for restaurants and $250 for everything else, how long will it take her to save $750?
>
> A. 3 months
> B. 4 months
> C. 5 months*
> D. 6 months

Initially, the Financial Calculation Test included 24 items. Similar to the General Knowledge Test, items with success rate corresponding to 10% or less, as well as items with success rate corresponding to 90% or more were subsequently removed (for success rate variability, see Table 3). Six items were therefore removed from analyses such that the Financial Calculation Test corresponded to 18 items (Cronbach's alpha = 0.77). Appendix C contains the test questions, item-by-item success rate, and removed items.

Probability Calculation Test. The Probability Calculation Test assessed individuals' ability to make judgments based on probabilities. Test items included calculating the likelihood of an event (such as a coin toss, dice toss or pulling random balls out of a bag; 9 items),

calculating decimals from fractions to identify the correct response (11 items), and comparing probabilities in decimal format (4 items). Each multiple-choice item on the Probability Calculation Test corresponded to four possible alternatives. Each question presented separately, followed by the confidence rating scale. An example item is (* = correct response):

> A chance of miscarriage is approximately 15% during the first 20 weeks of pregnancy. Drug use can triple the incidence of miscarriage during this time. If a pregnant woman uses drugs during her first 20 weeks, what chance does she have of having a miscarriage?
>
> A. 15%
> B. 30%
> C. 45%*
> D. 60%

Table 3
Success Rate Variability for Financial Calculation Test

Success Rate (%)	Frequency	Percent
0-10	0	0
10-20	0	0
20-30	0	0
30-40	1	4.17
40-50	1	4.17
50-60	5	20.83
60-70	5	20.83
70-80	3	12.5
80-90	3	12.5
90-100	6	25

Note. Items with overall success of less than 10% or greater than 90% were removed.

Similar to the General Knowledge and Financial Calculation tests, the Probability Calculation Test initially included 24 items. Items with success rate corresponding to 10% or less, as well as items with success rate corresponding to 90% or more were subsequently removed (for a success rate variability, see Table 4). Three items were therefore removed from

analyses such that the Probability Calculation Test corresponded to 21 items (Cronbach's alpha = 0.69). Appendix C contains the test questions, item-by-item success rate, and removed items.

Table 4
Success Rate Variability for Probability Calculation Test

Success Rate (%)	Frequency	Percent
0-10	0	0
10-20	0	0
20-30	0	0
30-40	2	8.33
40-50	6	25
50-60	2	8.33
60-70	6	25
70-80	4	16.67
80-90	1	4.17
90-100	3	12.5

Note. Items with overall success of less than 10% or greater than 90% were removed.

Indices of monitoring accuracy. The degree of fit between confidence judgments and performance can be examined in various ways. The present study focused on two monitoring accuracy indices: (1) calibration (i.e., bias, overconfidence, and resistance to overconfidence) and (2) resolution. Figure 2 summarizes indices used in this study.

Calibration. Calibration represents how well subjective confidence judgments match actual performance. There are different ways to index calibration, such that various measures of calibration have been used within the metacognition, education, and decision-making literature. The present study focused on three calibration indices: (1) the bias index, (2) the overconfidence index, and (3) the resistance to overconfidence index.

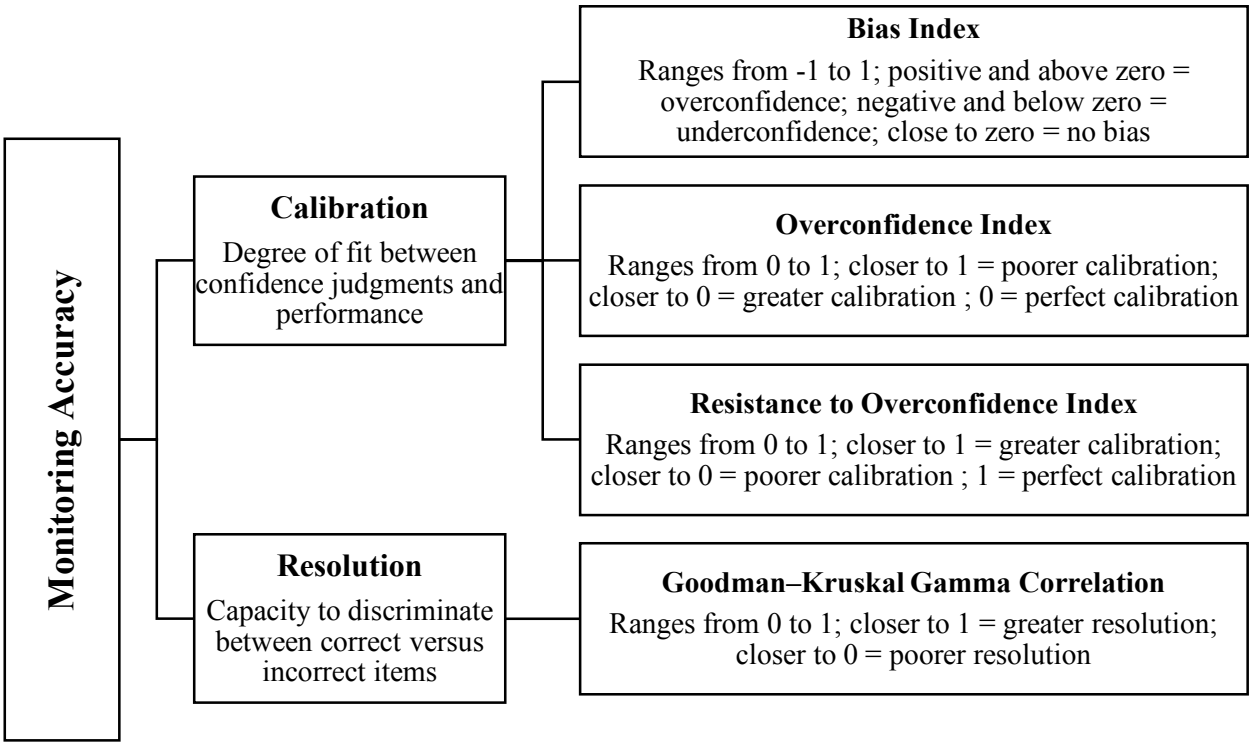


Figure 2. Indices of monitoring accuracy. This figure summarizes monitoring accuracy indices used in the current study.

Bias index (Ronis & Yates, 1987). Bias measures the degree of over or underconfidence, as well as the direction of judgment errors. It is the difference between the mean subjective probability of a correct answer (i.e., mean confidence rating) and the proportion of correct answers (i.e., percent of items answered correctly). The bias index is represented by the following formula:

Bias index = mean confidence - percent correct

The bias index ranges from -1 to +1 with -1 indicating underconfidence, zero indicating perfect calibration and +1 indicating overconfidence. Scores close to zero indicate a lack of bias, such that confidence is matched with performance (see Figure 3).

Overconfidence index (Bruine de Bruin et al., 2007; Parker & Fischhoff, 2007). For investigations including individual difference variables, where scores are unidirectional and range from zero to one, the bias index cannot be examined in correlational analyses with these very variables, as both ends of the bias index measure a type of miscalibration (see Figure 3).

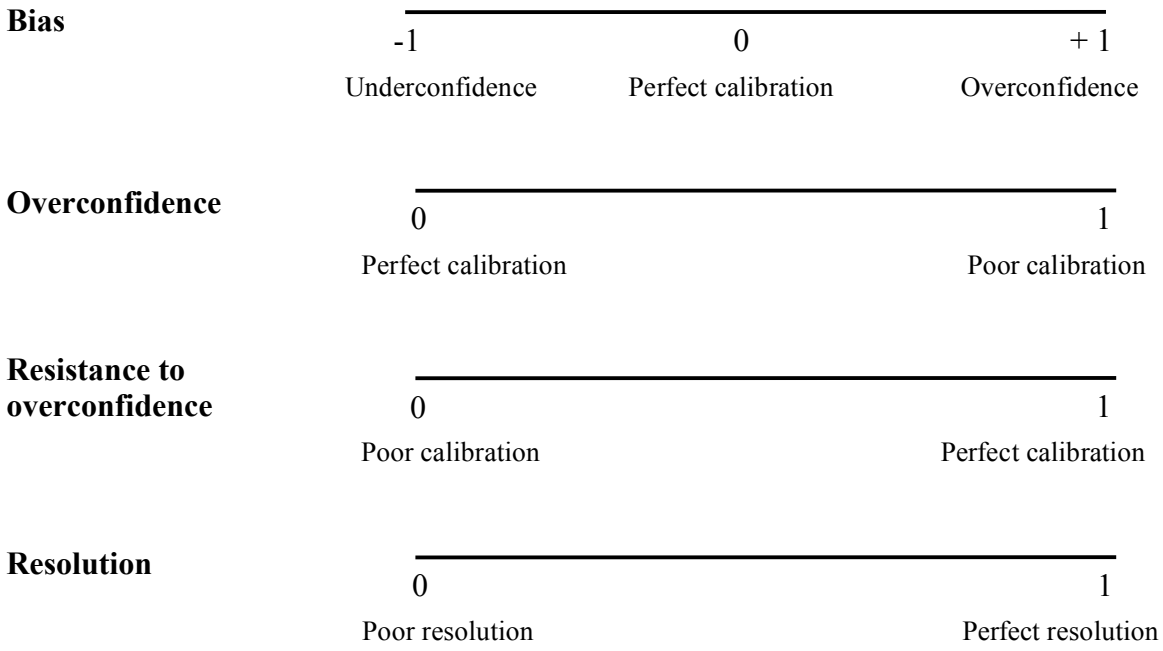


Figure 3. Depiction of monitoring accuracy indices. This figure depicts the range and direction of each monitoring accuracy index.

Thus, a measure to index the magnitude of the bias score to measure the raw degree of miscalibration was needed. As the tendency for participants is to display over- rather than underconfidence, and to be consistent with the decision-making literature (Stanovich, West & Toplak, 2016), we called this variable the overconfidence index. As such, the overconfidence index reflects the difference between confidence judgments and performance. It is the absolute difference between mean confidence and percentage correct. The overconfidence index is represented by the following equation:

$$\text{Overconfidence index} = | \text{mean confidence} - \text{percent correct} |$$

Overconfidence values range from 0 to 1, with higher scores reflecting a greater difference between confidence and performance (i.e., poorer calibration) and lower scores reflecting a smaller difference between confidence and calibration (i.e., better calibration). A score of zero means that confidence is perfectly matched with performance (see Figure 3).

Resistance to overconfidence index (Bruine de Bruin et al., 2007; Parker & Fischhoff, 2007). The overconfidence index assesses the magnitude of the miscalibration rather than the direction of the miscalibration. Consequently, overconfidence was subtracted from one to measure resistance to overconfidence, so that higher scores indicate better calibration; this is also consistent with analyses of the resolution index where a higher score indicates better resolution. The resistance to overconfidence index was therefore formulated for the current study and is the inverse of the overconfidence index described above. It is represented by the following equation:

Resistance to overconfidence index = 1- | mean confidence - percent correct |

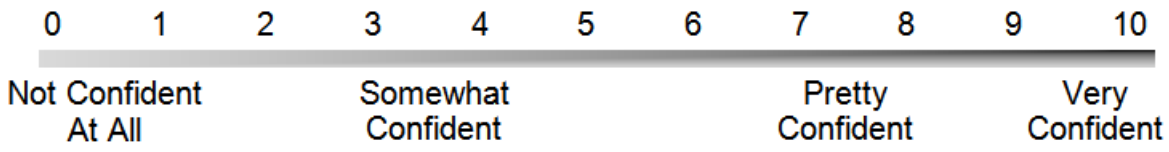
As described above, overconfidence values range from 0-1, with higher scores reflecting poor calibration. The resistance to overconfidence scores also range from 0 – 1, but higher scores reflect better calibration. A score of 1 means that confidence is perfectly matched with performance (see Figure 3).

Resolution (Koriat, 2012a). Resolution, also known as discrimination or relative accuracy, is another metacognitive index used to examine monitoring accuracy. Resolution reflects an individual's capacity to discriminate between correct versus incorrect items and to assign an appropriate level of confidence to each. High resolution occurs if participants can accurately assign higher probability of being correct to correct answers (e.g., 100% confidence in a correct answer) and a lower probability of being correct to incorrect answers (e.g., 25% confidence in an incorrect answer). The resolution index was measured with the use of the

Goodman–Kruskal Gamma correlation (Nelson, 1984; Masson & Rotello, 2009; Schraw, 2009) and involved a within-subjects measure of the relationship between confidence judgments and the correctness on each individual item (Koriat et al., 2009). Resolution values range from 0 to 1, with higher values representing higher resolution and lower values representing lower resolution (see Figure 3). Mean resolution scores were also compared to 0 by one-sample t-test, to support that participants were not using the confidence scale arbitrarily and were meaningfully discriminating between correct and incorrect responses.

Task-specific judgments.

Predictive judgments. Prior to completing each task, participants were asked for an aggregated predictive judgment regarding their ability to answer general knowledge questions (i.e., How confident are you in your ability to correctly answer general knowledge questions, such as "What is the capital of Greece?"), their ability to solve financial calculation problems (i.e., How confident are you in your ability to correctly solve real-world calculation problems, such as "How many Canadian dollars in 30 US dollars?") and their ability to solve probability calculation problems (i.e., How confident are you in your ability to correctly solve probability questions, such as "Two coins are tossed, what is the probability that two heads are obtained?"). Participants were asked to rate their confidence judgments on the following scale:



Postdictive judgments. Following each of experimental task, participants were asked to re-rate their performance on the same scale described above (e.g., "How confident are you in your ability to correctly solve general knowledge questions?"; 0 = Not Confident at All; 10 =

Very Confident).

Task difficulty, effort, and effort required. After each experimental task, participants were also asked to rate aggregated judgments of experienced difficulty (i.e., "How difficult did you find this task?"; 0 = Not Difficult at All; 10 = Very Difficult), required effort (i.e., "How much effort was required of you to complete this task?"; 0 = No Effort at All; 10 = Extreme Effort) and the feeling associated with expending effort (i.e., "How did using that level of effort make you feel?"; 0 = Extremely Pleasant; 10 = Extremely Unpleasant). Tables 5, 6, and 7 contains means, standard deviations, ranges, and normality estimates for the aggregated ratings of each experimental task. The ratings were approximately normally distributed.

Cognitive ability tasks. Cognitive ability encompassed two domains of cognitive functioning: (1) intellectual ability and (2) working memory.

Intellectual ability. The Shipley-2 (Shipley, Gruber, Martin, & Klein, 2009***)*** provided an estimate of general intelligence and included two subtests: Vocabulary and Block Patterns. The Shipley-2 can be administered individually or in groups, with an administration time of 20 minutes. On the Vocabulary subtest, participants were asked to choose amongst four alternatives which definition most closely matches the target word. On the Block Patterns subtest, participants were asked to choose amongst four alternatives which block best completes the design. Scores on the Vocabulary subtest are thought to reflect crystalized abilities, whereas scores on the Block Patterns subtest represent fluid reasoning abilities. In addition to looking at subtests separately, raw scores (not age corrected) were also standardized and summed to create a composite score of general intelligence, called the Intelligence Raw Score Composite. The Shipley Vocabulary test has been reported to range from .85 to .92 across age groups and the

Block Patterns has been reported to range from .88 to .94 across age groups for internal consistency (Shipley et al., 2009). Higher scores indicated better intellectual abilities.

Working memory. Based on the methods of Turner and Engle (1989), the reading span task provided a measure of working memory. This task was group administered with sentences presented on PowerPoint slides via a projector screen (see Appendix D). This task included 12 blocks, each consisting of two, three, four, or five sentences. There was a total of 42 items (i.e., two sets of two, three, four and five sentences). Participants were provided with a response form to follow along with while direct instructions were given by the examiner. Participants were provided with the following instructions:

> You will see a sentence on the screen. Your job is to read the sentence out loud, along with me. As soon as you have finished reading the sentence, decide if the sentence is True or False by checking off either True or False on the sheet of paper in front of you. There will be 12 sets of sentences, each set containing 2-5 sentences. After each set, you will be prompted to write down the last word of each sentence from that set (e.g., "What was the last word in each sentence that you read in Set #1?) Don't worry about spelling! **Please put your pencils down as soon as you have finished writing the words.**

Participants were presented with a practice block with two sentences before the actual test blocks were started. For each trial, participants were asked to circle on the response form whether sentences are true or false (e.g., "Cucumbers are green"), while also having to commit the last word in each sentence to memory (e.g., "green"). At the end of each block, participants were asked to recall the to-be-remembered words from the entire block. Recall accuracy was the dependent measure on this task. Higher scores reflected better working memory abilities. Present Cronbach's alpha = 0.66.

Self-perception scales. Self-perceptions measured in the present study included academic self-concept and cognitive anxiety.

Academic Self-Concept Scale (Reynolds, Ramírez, Magriña, & Allen, 1980). The Academic Self-Concept Scale (ASCS) assessed the academic facet of general self-concept. A short-form 22-item version was used in this study. Participants were asked to rate their level of agreement with various statements on a six-point Likert-type scale (1= "disagree strongly", 2 = "disagree moderately", 3= "disagree slightly", 4= "agree slightly", 5= "agree moderately", 6 = "agree strongly"). Higher total scores on the ASCS indicated greater academic self-concept. Two examples items are "*I often get discouraged about school*" and "*If I try hard enough, I will be able to get good grades.*" Present Cronbach's alpha = 0.92.

Cognitive Test Anxiety Scale (Cassady & Johnson, 2002). The Cognitive Test Anxiety Scale (CTAS) is a 27-item measure that assessed the cognitive dimension of performance-related anxiety. Participants were asked to rate their level of agreement with various statements on a six-point Likert-type scale (1= "disagree strongly", 2 = "disagree moderately", 3= "disagree slightly", 4= "agree slightly", 5= "agree moderately", 6 = "agree strongly"). Higher total scores on the CTAS usual indicate less cognitive test anxiety. To facilitate comparisons with other variables in the present study which were scaled in the positive direction, scores on the CTAS were reversed, such that higher scores indicated increased cognitive anxiety. Two examples items are "*I tend to freeze up on things like intelligence tests and final exams*" and "*The prospect of taking a test in one of my courses would not cause me to worry.*" Present Cronbach's alpha = 0.93.

Self-perceptions raw composite score. For regression analyses, reported later, a composite self-perception score that combined scores on measures of academic self-concept and cognitive test anxiety was used, as measures where highly correlated ($r(136) = -0.72$, $p > 0.01$)

and allowed to reduce the number of predictors in the overall regression models. To form this index, scores on the ASCS and on the CTAS were standardized and summed.

Real-world outcome variables. Real-world outcome variables considered included academic average and learning challenges. Table 9 contains means, standard deviations, range and normality estimates for real-world outcome variables. All variables were approximately normally distributed.

Academic average. Participants were asked to select which of the following categories best represented their current academic average: (1) below 49%, (2) 50-59%, (3) 60-69%, (4) 70-79%, or (5) 80-100%. Higher scores indicated higher self-reported academic average.

Learning challenges. Participants were asked to rate their perceived level of learning difficulties by rating the question "Overall, do you think that you have difficulties in learning or academics?" on a four-point Likert-type scale (1 = "no"; 2 = "minor difficulties"; 3 = "definite difficulties"; 4 = "severe difficulties"). Higher scores indicated greater perceived learning challenges.

Procedure

Testing sessions were conducted in group format and were 120 minutes in length. A maximum of 10 participants were tested at one time and each testing session included two examiners. Participants were first instructed to complete and sign informed consent and demographic forms. Once initial forms were collected, participants were then engaged in the group-administered reading span working memory task presented on PowerPoint slides via a projector screen, followed by instructions to complete both the Vocabulary and Block Patterns subtests of the Shipley-2. Participants were then each provided with a calculator and asked to independently complete the general knowledge, financial calculation, and probability calculation

tasks. Tasks from the current study were counterbalanced with tasks for another study, so that participants received the tasks in one of two possible order formats. There were no statistically significant differences in success rate between the tasks based on presentation order ($ps > 0.05$). Once all study tasks where completed, participants were asked to answer a questionnaire that included measures of academic self-concept and cognitive anxiety.

Results

Data Management

All variables of interest were first examined for missing values and accurate data entry. For the general knowledge task, due to a technical printing error, 38 participants missed completing two items on the task. Overall success rate between participants who missed the two items and the remaining 98 participants was compared. There were no statistically significant differences in overall success rate ($p > 0.05$) between both sets of participants. Scores of these 38 participants were therefore pro-rated for the statistical analyses. Other missing values were dealt with by mean substitution. Out-of-range values, plausible means, and standard deviations were also examined for all variables of interest. Variables of interest showed normal distributions with skewness values below 3 and kurtosis values below 10 (Kline, 2005).

There were no significant sex differences in resistance to overconfidence for general knowledge, $t(134) = 0.72$, $p = 0.47$, financial calculation, $t(135)= -0.36$, $p = 0.72$, or probability calculation, $t(135)= 0.52$, $p = 0.60$. Similarly, there were no significant sex differences in resolution for general knowledge, $t(134) = 0.72$, $p = 0.11$, financial calculation, $t(124)= -0.36$, $p = 0.23$, or probability calculation, $t(133)= 0.52$, $p = 0.26$. As such, the present study did not find any sex differences in monitoring accuracy.

Success, Confidence, and Indices of Monitoring Accuracy

The first goal of the present study was to measure the relationships between accuracy, confidence, and indices of monitoring accuracy, including calibration (i.e., bias and resistance to overconfidence) and resolution. Definitions for the various variables of interest are summed in Table 5.

Table 5
Definitions of Study Variables

Variable	**Definition**
Metacognition	
Accuracy/ Success Rate	Accuracy represents a measure of task performance and refers to whether items are successfully completed.
Confidence	Confidence represented the subjective feeling about the correctness of one's answer.
Monitoring accuracy	Metacognitive monitoring represents a subjective assessment of task performance. Monitoring accuracy refers to the fit between assessments of confidence and actual performance. Calibration and resolution represent measures of monitoring accuracy.
Calibration	
Calibration	The degree of fit between confidence judgments and the objective accuracy of those judgments. Various indexes represent calibration, including bias and resistance to overconfidence.
Bias	Refers to the direction of miscalibration: underconfidence occurs when confidence judgments are lower than actual performance; overconfidence occurs when judgments are greater than actual performance.
Resistance to overconfidence	Resistance to overconfidence represents the discrepancy between confidence and accuracy and assesses the magnitude of miscalibration. Higher scores indicate better calibration.
Resolution	
Resolution	The degree to which participants can discriminate between correct and incorrect items, measured by the Goodman-Kruskal Gamma correlation (*G*). *G* represents the mean within-participant gamma correlation between confidence and accuracy. Higher scores indicate better resolution.

Success rate and confidence. Means and standard deviations for success rate and mean confidence are presented in Tables 6, 7, and 8 for the general knowledge, financial calculation, and probability calculation tasks respectively.

The scores were approximately normally distributed for all variables. Three participants had extremely low mean confidence scores. These scores were retained as there was no justification for their removal. Analyses were conducted with and without the extreme scores, and the scores did not impact the pattern of results.

As designed to be, mean success rates (global mean = 60.0%) and confidence ratings (global mean = 74.3%) on all tasks were in an intermediate level of difficulty, which allowed confidence variability above and below actual success rates. The small differences in success and confidence among the tasks allowed for comparison across the tasks (general knowledge: $M_{success}$ = 0.53 (*SD* = 0.13), $M_{confidence}$ = 0.72 (*SD*=0.12); financial calculation: $M_{success}$ = 0.63 (*SD* = 0.20), $M_{confidence}$ = 0.74 (*SD*=0.18); probability calculation: $M_{success}$ = 0.59 (*SD* = 0.17), $M_{confidence}$ = 0.75 (*SD* = 0.16)).

Pearson's correlations were run to assess the relationships between success and confidence (see Table 9). For all tasks, there was a moderate to large statistically significant relationship between mean success and mean confidence, $r(136) = 0.41$ to 0.71, $ps < 0.01$), suggesting that increased success was related to increased confidence across task domains. Figure 4 depicts mean success and mean confidence for all experimental tasks.

Calibration. Means and standard deviations for monitoring accuracy indices of calibration (i.e., bias and resistance to overconfidence) are presented in Tables 6, 7, and 8 for the general knowledge (M_{bias} = 0.19, *SD* = 0.14; $M_{overconfidence}$ = 0.80, *SD* = 0.12), financial

Table 6
Descriptive Statistics for General Knowledge Test Dependent Variables (N = 136)

Dependent Measure	Mean (SD)	Potential Range	Observed Range	Skewness	Kurtosis
Pre-Task Ratings					
Predictive confidence (10 = very confident)	5.80 (2.09)	0 – 10	1 – 10	- 0.01	- 0.73
Task Performance and Monitoring					
Success rate	0.53 (0.13)	0 – 1	0.21 – 0.94	0.03	0.20
Confidence	0.72 (0.12)	0 – 1	0.28 – 0.96	- 0.76	0.91
Bias	0.19 (0.14)	-1 – 1	-0.15 – 0.48	- 0.20	- 0.12
Resistance to overconfidence[a]	0.80 (0.12)	0 – 1	0.52 – 1.00	- 0.38	- 0.48
Resolution	0.35 (0.28) *	0 – 1	0 – 1	0.46	- 0.79
Post-Task Ratings					
Postdictive confidence (10 = very confident)	5.87 (1.88)	0 – 10	0 – 10	- 0.33	0.31
Task difficulty (10 = very difficult)	3.90 (2.10)	0 – 10	0 – 10	0.12	- 0.27
Effort required (10 = extreme effort)	4.62 (2.07)	0 – 10	0 – 9	- 0.23	- 0.38
Feeling of effort (10 = extremely unpleasant)	4.60 (1.74)	0 – 10	0 – 9	- 0.48	0.89

Note. Asterisks indicate the significance of one sample t-tests for differences of calibration and resolution from zero.
[a]Higher scores indicate larger discrepancy between confidence and accuracy.

Table 7
Descriptive Statistics for Financial Calculation Test Dependent Variables (N = 136)

Dependent Measure	Mean (SD)	Potential Range	Observed Range	Skewness	Kurtosis
Pre-Task Ratings					
Predictive confidence (10 = very confident)	5.35 (2.24)	0 – 10	0 – 10	0.01	-0.44
Task Performance and Monitoring					
Success rate	0.63 (0.20)	0 – 1	0.17 – 0.94	-0.18	-0.91
Confidence	0.74 (0.18)	0 – 1	0.29 – 1.00	-0.27	-0.74
Bias	0.11 (0.15)	-1 – 1	-0.21- 0.52	0.41	0.42
Resistance to overconfidence[a]	0.86 (0.11)	0 – 1	0.29 – 0.52	0.24	0.14
Resolution	0.43 (0.32) *	0 – 1	0.00 – 1.00	0.03	-0.12
Post-Task Ratings					
Postdictive confidence (10 = very confident)	5.54 (2.95)	0 – 10	0-10	-0.13	0.08
Task difficulty (10 = very difficult)	5.62 (2.46)	0 – 10	0-10	-0.38	-0.23
Effort required (10 = extreme effort)	6.47 (2.24)	0 – 10	0-10	-0.48	0.10
Feeling of effort (10 = extremely unpleasant)	5.93 (2.22)	0 – 10	0-10	-0.20	0.16

Note. Asterisks indicate the significance of one sample t-tests for differences of calibration and resolution from zero.
[a]Higher scores indicate larger discrepancy between confidence and accuracy.

Table 8

Descriptive Statistics for Probability Calculation Test Dependent Variables (N = 136)

Dependent Measure	Mean (SD)	Potential Range	Observed Range	Skewness	Kurtosis
Pre-Task Ratings					
Predictive confidence (10 = very confident)	6.24 (2.06)	0 – 10	2 -10	0.20	-0.76
Task Performance and Monitoring					
Success rate	0.59 (0.17)	0 – 1	0.14 – 0.95	-0.15	-0.78
Confidence	0.75 (0.16)	0 – 1	0.25 – 1.00	-0.56	-0.23
Bias	0.16 (0.15)	-1 – 1	-0.27 - 0.51	0.04	0.27
Resistance to overconfidence[a]	0.82 (0.12) *	0 – 1	0.00 – 0.51	0.81	0.03
Resolution	0.37 (0.28) *	0 – 1	0.00 – 1.00	0.25	-0.97
Post-Task Ratings					
Postdicitve confidence (10 = very confident)	5.31 (2.65)	0 – 10	0-10	-0.20	-0.85
Task difficulty (10 = very difficult)	5.47 (2.17)	0 – 10	0-10	-0.28	0.02
Effort required (10 = extreme effort)	6.14 (2.12)	0 – 10	0-10	-0.27	-0.04
Feeling of effort (10 = extremely unpleasant)	5.69 (1.96)	0 – 10	0-10	0.36	0.66

Note. Asterisks indicate the significance of one sample t-tests for differences of calibration and resolution from zero.

[a]Higher scores indicate larger discrepancy between confidence and accuracy

Table 9
Correlations Between Mean Success Rate, Mean Confidence, Resistance to Overconfidence and Resolution

	1	2	3	4
General Knowledge Task				
1. Mean success rate	--	0.41**	0.62**	0.11
2. Mean confidence		--	-0.38**	0.24**
3. Resistance to overconfidence[a]			--	-0.10
4. Resolution				--
Financial Calculation Task				
1. Mean success rate	--	0.71**	0.52**	-0.15
2. Mean confidence		--	-0.07	-0.21**
3. Resistance to overconfidence[a]			--	0.00
4. Resolution				--
Probability Calculation Task				
1. Mean success rate	--	0.60**	0.50**	0.24**
2. Mean confidence		--	-0.25**	0.10
3. Resistance to overconfidence[a]			--	0.20*
4. Resolution				--

[a] A higher score indicated better calibration.
*$p < .05$. **$p < .01$.

calculation ($M_{bias} = 0.86$, $SD = 0.15$; $M_{overconfidence} = 0.14$, $SD = 0.11$), and probability calculation tasks ($M_{bias} = 0.16$, $SD = 0.15$; $M_{overconfidence} = 0.82$, $SD = 0.12$).

Bias index. The significance of one-sample t-tests for differences of bias from zero are marked by asterisks near the means in Tables 6, 7, and 8. The bias index was significantly greater than zero for the general knowledge task, $t(135) = 16.22$, $p < 0.001$, the financial calculation task, $t(135) = 8.61$, $p < 0.001$, and the probability calculation task, $t(135) = 12.55$, $p < 0.001$. These findings indicated that across experimental tasks, confidence was not well matched with performance.

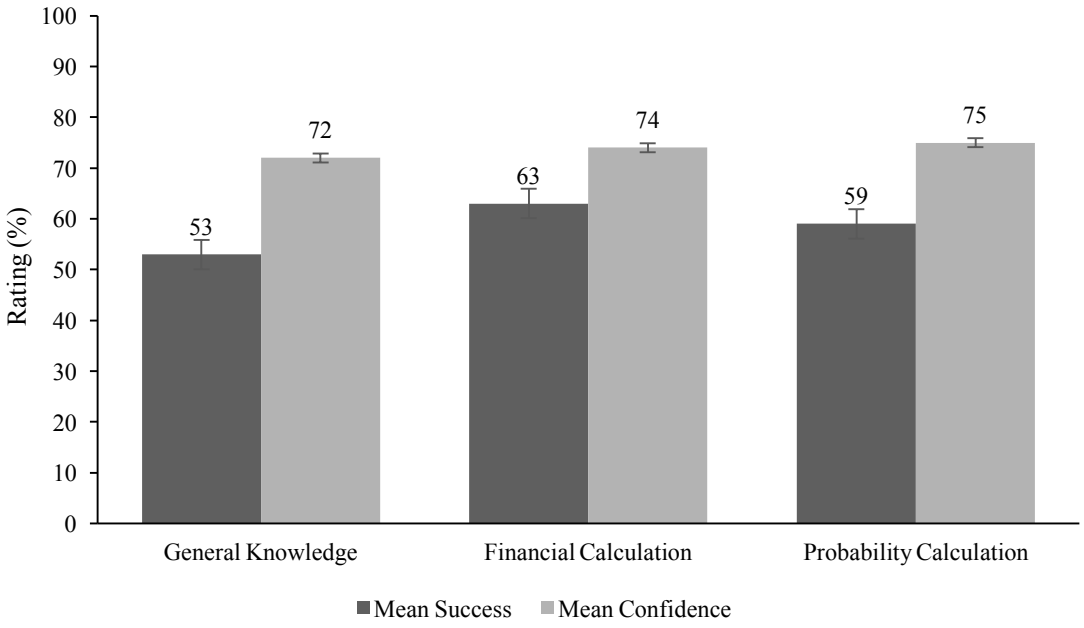


Figure 4. Mean success and mean confidence across experimental tasks. Mean percentage of participant *(N* = 136) success rate and confidence on the general knowledge, financial calculation and probability calculation tasks.

Figure 5 depicts the number of participants who were overconfident (i.e., bias index scores above zero), underconfident (i.e., bias index scores below zero), and perfectly calibrated (i.e., bias index scores equal to zero). On the general knowledge task, out of 136 participants, 14 participants were underconfident, one participant was calibrated, and 121 participants were overconfident. On the financial calculation task, 21 participants were underconfident, two participants were calibrated, and 113 participants were overconfident. On the probability calculation task, 13 participants were underconfident, two participants were calibrated, and 121 participants were overconfident.

As previously described, the bias index cannot be examined in correlational analyses for investigations including variables where scores are unidirectional and range from zero to one. The resistance to overconfidence index was therefore used as for all subsequent calibration analyses.

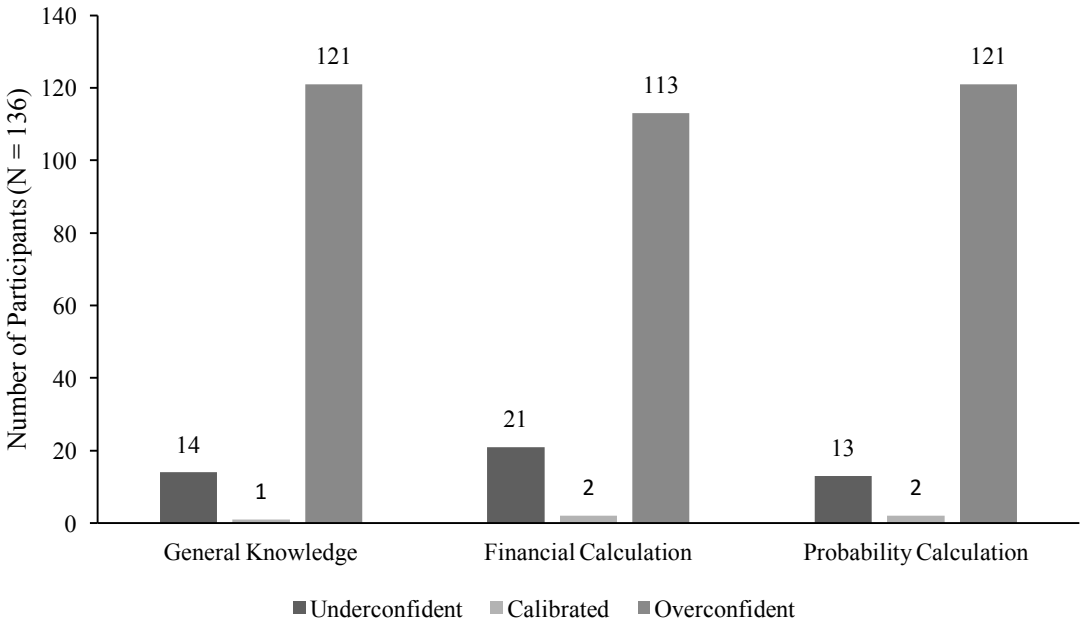


Figure 5. Confidence levels across experimental tasks. Number of participants *(N* = 136) who were underconfident, overconfident, and perfectly calibrated on the general knowledge, financial calculation and probability calculation tasks.

Resistance to overconfidence. Pearson correlations were run to assess the relationships between the resistance to overconfidence index, mean success rate, and mean confidence (see Table 9). There was a large statistically significant relationship between resistance to overconfidence and mean success rate across all three tasks, $r(136) = 0.50$ to 0.62, $ps < 0.01$. Higher accuracy was therefore positively correlated with better calibration.

The relationship between resistance to overconfidence and mean confidence seemed to vary across tasks. There was a moderate negative correlation between mean confidence and resistance to overconfidence within the general knowledge task, $r(136) = -0.38$, $p < 0.01$, suggesting that increased confidence in general knowledge was related to less awareness of accuracy. Similarly, there was a small negative correlation within the probability calculation task, $r(136) = -0.25$, $p < 0.01$, also suggesting that increased confidence in probability

calculations was related to poorer accuracy awareness. There was however no statically significant relationship between confidence and resistance to overconfidence on the financial calculation task ($p > 0.05$).

The significance of one-sample t-tests for differences of calibration from zero are marked by asterisks near the means in Tables 6, 7, and 8. The resistance to overconfidence index was significantly greater than zero for the general knowledge task, $t(135) = 79.93$, $p < 0.001$, the financial calculation task, $t(135) = 86.97$, $p < 0.001$, and the probability calculation task, $t(135) = 77.23$, $p < 0.001$. The resistance to overconfidence index in the study however reflects the magnitude of miscalibration, but does not take into account underconfidence in responses. In a separate analysis, accuracy was subtracted from confidence for each task. Obtained mean differences were then subjected to a one sample t-test. Even when underconfident responses were included, a significant overconfidence effect was still obtained across all tasks, $t(135)=8.61$ to 16.22, $p<.0001$. These findings indicate that across experimental tasks, overall confidence was not well matched with performance, with a tendency towards over- rather than underconfidence.

Resolution. Means and standard deviations for resolution, as measured by Gamma correlations[2], are presented in Tables 6, 7, and 8 for the general knowledge ($M= 0.35$, $SD = 0.28$), financial calculation ($M= 0.43$, $SD = 0.32$), and probability calculation tasks ($M = 0.37$, $SD = 0.28$) respectively. The gamma coefficient (G) represents the mean within-participant gamma correlation between the confidence and accuracy. The significance of one-sample t-tests for differences of resolution from zero are marked by asterisks near the means in Tables 6, 7, and 8. The resolution index was significantly greater than zero for the general knowledge task, $t(135) =$

[2] Another resolution index called the confidence-judgment accuracy quotient (CAQ; Jackson & Kleitman, 2014; Schraw, 2009) which provides a difference score between the average confidence assigned to correct and incorrect items, was also examined. Across all of the analyses, parallel findings were found using the Gamma and CAQ indices.

14.87, $p < 0.001$, the financial calculation task, $t(135) = 15.02$, $p < 0.001$, and the probability calculation task, $t(134) = 15.12$, $p < 0.001$. Consistent with prior research, significant resolution effects were observed across tasks, indicating that participants discriminated successfully between correct and incorrect responses.

Pearson's correlations were run to assess the relationships between resolution, mean success, and mean confidence (see Table 9). These relationships seemed to vary across tasks. On the general knowledge task, there was a small positive correlation between mean confidence and resolution, $r(136) = 0.24$, $p < 0.01$, suggesting that greater confidence in a response was related to a better ability to discriminate between correct and incorrect answers. On the financial calculation task however, an inverse relationship was obtained, such that there was a small negative correlation between mean confidence and resolution, $r(136) = -0.21$, $p < .01$. This result suggests that on financial calculation items, greater confidence in a response is associated with poorer ability to discriminate between correct and incorrect. On the probability calculation task, there was no statistically significant correlation between mean confidence and resolution ($p > 0.05$), yet there was a small positive correlation between mean success and resolution, $r(136) = 0.24$, $p < 0.01$, and between resistance to overconfidence and resolution, $r(136) = 0.20$, $p < 0.05$. As such, on probability calculation items, greater resolution was associated with greater accuracy and calibration, but not with confidence.

Calibration versus resolution. The relationship between calibration, measured by the resistance to overconfidence index, and resolution, measured by *G*, was explored within each experimental task. As indicated in Table 9, resistance to overconfidence and resolution scores were not significantly correlated for both the general knowledge task and the financial calculation task (ps > 0.05). These findings suggest that calibration was not associated with

successful discrimination between incorrect and correct responses for either the general knowledge or the financial calculation domain. However, there was a small positive association between resistance to overconfidence and resolution scores within the probability calculation task, $r(136) = 0.20$, $p < 0.05$. As such, within the domain of probability calculations, greater calibration was associated with greater discrimination between incorrect and correct responses.

Domain Generality Versus Domain Specificity

A second goal of the present study was to explore the question whether monitoring accuracy indices of resistance to overconfidence and resolution remain stable across domains or rather fluctuate as a function of domain knowledge.

Pearson's correlations were first run to assess the relationships between the resistance to overconfidence index and each experimental task (see Table 10).

Table 10
Correlations Between Resistance to Overconfidence and Experimental Tasks

	1	2	3
1. General knowledge resistance to overconfidence [a]	--	0.34**	0.22*
2. Financial calculation resistance to overconfidence [a]		--	0.49**
3. Probability calculation resistance to overconfidence [a]			--

[a] A higher score indicated better calibration.
*$p < .05$. **$p < .01$.

There was a small statistically significant relationship between resistance to overconfidence on the general knowledge task and resistance to overconfidence on the probability calculation task, $r(136) = 0.22$, $p < 0.05$. There was a moderate statistically significant relationship between resistance to overconfidence on the general knowledge task and resistance to overconfidence on the financial calculation task, $r(136) = 0.34$, $p < 0.01$. There was a large statistically significant relationship between resistance to overconfidence on the financial

calculation task and resistance to overconfidence on the probability calculation task, $r(136) = 0.49$ $p < 0.01$. These correlations indicated that better calibration on one task was associated with better calibration on the other tasks. However, as expected, stronger associations were observed between tasks that shared domain similarities, such as the calculation requirements on the financial calculation and probability calculation tasks (i.e., moderate versus small effect size). Using a Fisher r-to-z transformation which allows to statistically compare the differences between Pearson correlations, it was found that correlations between general knowledge and financial calculation and between general knowledge and probability calculations were statistically significantly different than the correlation between financial calculation and probability calculation ($p < 0.05$), further supporting this difference in correlation strength between domains.

Pearson's correlations were also run to assess the relationship between resolution and experimental tasks (see Table 11). There were no statistically significant relationships between resolution and experimental tasks (ps > 0.05). The ability to discriminate between correct and incorrect responses on one task was therefore not related to the ability to discriminate between correct and incorrect responses on another other task.

Table 11
Correlations Between Resolution and Experimental Tasks

	1	2	3
1. General knowledge resolution	--	-0.17	0.16
2. Financial calculation resolution		--	0.08
3. Probability calculation resolution			--

*$p < .05$. **$p < .01$.

Success, Confidence, and Monitoring Accuracy Correlates

A third goal of the present study was to further explore potential correlates of accuracy, confidence, and monitoring accuracy, including: (1) cognitive abilities (i.e., working memory and intellectual abilities), (2) task-specific judgments (i.e., pre- and postdictive confidence, aggregated ratings of task difficulty, effort required, and feeling of effort) and (3) self-perceptions (i.e., academic self-concept and cognitive test anxiety). Relationships between monitoring accuracy and real-world outcomes measures of academic grades and learning challenges were also explored.

Cognitive abilities. Descriptive statistics including means and standard deviations for working memory scores, Shipley Vocabulary scores (i.e., verbal abilities), Shipley Block Pattern scores (i.e., nonverbal abilities), and the Intelligence Raw Score Composite (i.e., summed and standardized Shipley Vocabulary and Block Pattern scores) are displayed in Table 12.

Table 12
Descriptive Statistics for Cognitive Ability Measures

	Mean (SD)	Potential Range	Observed Range	Skewness	Kurtosis
Working Memory Total Raw Score	32.10 (4.09)	0 – 42	20 – 42	-0.20	0.07
Shipley- 2 Vocabulary Raw Score	25.92 (4.35)	0 – 40	16 – 37	-0.1	-0.10
Shipley-2 Block Patterns Raw Score	16.18 (4.92)	0 – 26	5 – 26	0.01	-0.79
Intelligence Raw Score Composite	42.10 (6.94)	0 – 66	21 – 59	-0.01	-0.32

Note. $N = 136$

Cognitive abilities and success rates. Pearson's correlations were run to assess the relationships between cognitive abilities, success, and confidence (see Table 13). Across all

Table 13
Correlations Between Cognitive Abilities, Mean Success Rate, Mean Confidence, Resistance to Overconfidence, and Resolution

	Working Memory Raw Score	Shipley 2- Vocabulary Raw Score	Shipley 2- Block Patterns Raw Score	Intelligence Raw Score Composite
General Knowledge Task				
Mean success rate	0.38**	0.41**	0.29**	0.46**
Mean confidence	0.11	0.37**	0.16	0.35**
Resistance to overconfidence [a]	0.22**	0.10	0.12	0.14
Resolution	0.13	0.17	0.22**	0.26**
Financial Calculation Task				
Mean success rate	0.17*	0.20*	0.49**	0.47**
Mean confidence	0.05	0.12	0.42**	0.37**
Resistance to overconfidence [a]	0.10	0.14	0.17	0.21*
Resolution	-0.20*	-0.01	-0.01	-0.02
Probability Calculation Task				
Mean success rate	0.12	0.27**	0.51**	0.53**
Mean confidence	0.01	0.05	0.37**	0.29**
Resistance to overconfidence [a]	0.09	0.22*	0.23**	0.30**
Resolution	-0.02	0.14	0.18*	0.22*

[a] A higher score indicated better calibration.
*$p < .05$. **$p < .01$.

tasks, there was a small to large positive statistically significant relationship between cognitive abilities and accuracy $r(136) = 0.17$ to 0.53, $ps < 0.05$). These findings suggest that greater working memory, nonverbal, and verbal abilities are all related to greater performance. The one exception was the lack of association between working memory and mean success rate on the probability calculation task ($p > 0.05$).

Cognitive abilities and confidence. Associations between cognitive abilities and confidence varied across tasks. For the general knowledge task, there was a moderate positive statistically significant relationship between verbal abilities and mean confidence $r(136) = 0.37$, $p < 0.01$, and between mean confidence and intelligence composite scores, $r(136) = 0.35$, $p < 0.01$. Greater verbal abilities and intelligence composite scores were therefore related to greater confidence in general knowledge items. There were no statistically significant relationships between mean accuracy and working memory or nonverbal abilities ($ps > 0.05$). For both the financial calculation and probability calculation tasks, there was a moderate positive statistically significant relationship between mean confidence and nonverbal abilities, $r(136) = 0.37$ to 0.42, $ps < 0.01$, and between mean confidence and intelligence composite scores, $r(136) = 0.29$ to 0.37, $ps < 0.01$. As such, greater nonverbal abilities and intelligence composite scores were related to greater confidence in both financial and probability calculation items. There were no statistically significant relationships between mean confidence, working memory, and verbal abilities ($ps > 0.05$) on the financial calculation and probability calculation tasks.

Cognitive abilities and indexes of monitoring accuracy. Pearson's correlations were run to assess the relationships between cognitive abilities, resistance to overconfidence, and resolution (see Table 13). Associations between cognitive abilities and monitoring accuracy indexes varied across tasks. For the general knowledge task, there was a small positive

statistically significant relationship between resistance to overconfidence and working memory, $r(136) = 0.22$, $p < 0.01$, suggesting that greater working memory abilities were associated with greater calibration on general knowledge items. Moderate positive statistically significant relationships were also found between resolution and nonverbal abilities, $r(136) = 0.22$, $p < 0.01$, and between resolution and intelligence composite scores, $r(136) = 0.26$, $p < 0.01$. These findings suggest that greater nonverbal abilities and intelligence composite scores were related to greater confidence in general knowledge items. All other associations were non-significant (ps > 0.05).

For the financial calculation task, there was a small positive statistically significant relationship between resistance to overconfidence and intelligence composite scores, $r(136) = 0.21$, $p < 0.05$, such that greater intelligence composite scores were associated with greater calibration on financial calculation items. A small negative statistically significant relationship was also found between resolution and working memory, $r(136) = -0.20$, $p < 0.05$. This finding suggests that greater working memory abilities are related to greater difficulty in discriminating between correct and incorrect items. All other associations between cognitive abilities and monitoring accuracy indices on the financial calculation task were non-significant (ps > 0.05).

For the probability calculation task, there were small positive statistically significant relationships between resistance to overconfidence and verbal abilities, $r(136) = 0.22$, $p < 0.01$, between resistance to overconfidence and nonverbal abilities, $r(136) = 0.23$, $p < 0.01$, and between resistance to overconfidence and intelligence composite scores, $r(136) = 0.30$, $p < 0.01$. Greater verbal, nonverbal, and intelligence composite scores were therefore associated with greater calibration on probability calculation items. There was no statistically significant relationship between resistance to overconfidence on the probability calculation task and

working memory ($p > 0.05$). Regarding resolution, small positive statistically significant relationships were found between resolution and nonverbal abilities, $r(136) = 0.18$, $p < 0.05$, and between resolution and intelligence composite scores, $r(136) = 0.22$, $p < 0.05$. Greater nonverbal abilities and composites scores were therefore related to greater discrimination between correct and incorrect probability calculation items. There were no statistically significant relationships between resolution, working memory, and verbal abilities, on the probability calculation task ($ps > 0.05$).

Task-specific judgments. A second series of potential correlates explored in the present study included task-specific judgments represented by aggregated pre- and postdictive confidence judgments and post-task ratings of task difficulty, effort required, and feeling about effort. Intercorrelations between task specific judgments across experimental tasks are displayed in Table 14.

Across all tasks, there was a medium to large positive statistically significant relationship between predictive and postdictive confidence ratings, $r(136) = 0.39$ to 0.58, $ps < 0.01$, suggesting that participants performance estimation prior to completing a task was associated with their estimation after having experienced the task. In each task, task difficulty was negatively associated with postdictive confidence ratings, $r(136) = -0.37$ to 0.54, $ps < 0.01$, and positively associated with both effort required and feeling of effort, $r(136) = -0.37$ to 0.69, $ps < 0.01$. Thus, as perceived task difficulty increased, post task confidence in performance decreased. Furthermore, the more difficult the task was perceived, the more effortful and aversive the task was experienced. As expected, effort required and feeling of effort were also positively related across experimental tasks, $r(136) = 0.27$ to 0.59, $ps < 0.01$.

Relationships between each task-specific judgment, success, confidence, and monitoring accuracy were then explored.

Table 14
Intercorrelations Between Task-Specific Judgments across Experimental Tasks

	1	2	3	4	5
General Knowledge Task					
1. Predictive confidence	--	0.58**	-0.26*	-0.07	-0.09
2. Postdictive confidence		--	-0.47**	-0.16	-0.27**
3. Task difficulty			--	0.47**	0.27**
4. Effort required				--	0.27**
5. Feeling of effort					--
Financial Calculation Task					
1. Predictive confidence	--	0.51**	-0.20*	-0.16	-0.22*
2. Postdictive confidence		--	-0.54**	-0.44**	-0.44
3. Task difficulty			--	0.69**	0.55**
4. Effort required				--	0.59**
5. Feeling of effort					--
Probability Calculation Task					
1. Predictive confidence	--	0.39**	-0.11	-0.13	-0.06
2. Postdictive confidence		--	-0.37**	-0.25**	-0.29**
3. Task difficulty			--	0.69**	0.39**
4. Effort required				--	0.42**
5. Feeling of effort					--

*$p < .05$. **$p < .01$.

***Predictive and postdictive confidence judgments*.** Means and standard deviations for pre- and postdictive confidence judgments are presented in Tables 6, 7, and 8 for the general knowledge (M_{pre} = 5.80, SD = 2.09; M_{post} = 5.87, SD = 1.88), financial calculation (M_{pre} = 5.35, SD = 2.24; M_{post} = 5.54, SD = 2.95), and probability calculation (M_{pre} = 6.24, SD = 2.06; M_{post} = 5.31, SD = 2.65) tasks. As previously described, pre- and postdictive confidence judgments

represented aggregated judgments regarding the ability to answer task-related questions (i.e., "How confident are you in your ability to correctly solve general knowledge questions?"), rather than item-by-item confidence.

Pre- and postdictive judgments for each experimental task were then compared. Participants predicted their probability calculation skills to be higher prior to completing the task, compared to after experiencing the task, $t(135) = 4.12$, $p < 0.0001$. No differences were obtained on the general knowledge and financial calculation ratings, $t < 1$; and $t < 1$, respectively.

Pearson's correlations were run to assess the relationships between pre- and postdictive confidence ratings and mean success rate, mean confidence, resistance to overconfidence, and resolution (see Table 15). Across all tasks, there was a small to large positive statistically significant relationship between pre- and postdictive ratings, mean success, and mean confidence $r(136) = 0.25$ to 0.69, $ps < 0.05$. These findings suggest that greater pre- and postdictive confidence judgments are related to greater item-by-item accuracy and confidence. The only exception was the lack of association between pre-task confidence and mean success on the general knowledge task ($p > 0.05$).

The relationship between pre- and postdictive confidence ratings and monitoring accuracy indexes of resistance to overconfidence and resolution however were varied across tasks. On the general knowledge task, there was a small negative statistically significant relationship between predictive confidence and resistance to overconfidence, $r(136) = -0.24$, $p < 0.01$, suggesting that greater predictive confidence was related to poorer calibration. There was no statistically significant relationship between predictive confidence and resolution ($p > 0.05$).

Table 15

Correlations Between Pre- and Postdictive Confidence, Task-Specific Judgments, Mean Success, Mean Confidence, Resistance to Overconfidence, and Resolution

	Predictive Confidence	Postdictive Confidence	Task Difficulty	Effort Required	Feeling of Effort
General Knowledge Task					
Mean success rate	0.07	0.25**	-0.15	-0.12	-0.04
Mean confidence	0.31**	0.57**	-0.48**	-0.25**	-0.17*
Resistance to overconfidence[a]	-0.24**	0.21*	0.21*	.010	0.13
Resolution	0.10	0.19*	-0.21*	-0.16	-0.07
Financial Calculation Task					
Mean success rate	-0.28**	0.59**	-0.47**	-0.45**	-0.36**
Mean confidence	0.31**	0.69**	-0.46**	-0.45**	-0.43**
Resistance to overconfidence[a]	0.08	0.10	-0.12	-0.14	-0.01
Resolution	-0.04	-0.16	0.12	0.19*	0.04
Probability Calculation Task					
Mean success rate	0.22*	0.42**	-0.10	-0.25**	-0.06
Mean confidence	0.30 **	0.60**	-0.27**	-0.26**	-0.25**
Resistance to overconfidence[a]	-0.01	-0.10	0.10	-0.06	0.20*
Resolution	0.08	0.08	-0.08	-0.16	-0.06

[a]A higher score indicated better calibration.

*$p < .05$. **$p < .01$.

Regarding postdictive confidence, there was a small positive statistically significant relationship between general knowledge and postdictive confidence and resistance to overconfidence, $r(136) = 0.21$, $p < 0.05$, as well as between both postdictive confidence and resolution, $r(136) = 0.19$, $p < 0.05$. Greater post-task confidence in general knowledge was therefore relater to greater on-task calibration and resolution. Similar relationships were not observed between predictive and postdictive confidence judgments and monitoring accuracy indexes, on neither the financial calculation nor the probability calculation tasks.

Task difficulty, effort, and feeling of effort. Means and standard deviations for task difficulty ($M_{knowledge} = 3.90$, $SD = 2.10$; $M_{financial} = 5.62$, $SD = 2.95$; $M_{probability} = 5.47$ ($SD = 2.17$)), required effort ($M_{knoweldge} = 4.62$, $SD = 2.07$; $M_{financial} = 6.47$, $SD = 2.24$; $M_{probability} = 6.14$, $SD = 2.12$) and feeling of effort ($M_{knowledge} = 4.60$, $SD = 1.74$; $M_{financial} = 5.93$, $SD = 2.22$; $M_{probability} = 5.69$, $SD = 1.96$) are presented in Tables 5, 6, and 7. The means of the aggregated confidence ratings ranged in the 4-6 range on a possible range of 0-10, indicating that none of the tasks were rated as extremely difficult, requiring extreme effort, or as extremely unpleasant.

Pearson's correlations were run to assess the relationships between aggregated judgments of perceived difficulty, required effort, and feelings of effort with mean accuracy, mean confidence, resistance to overconfidence, and resolution (see Table 15). The relationships between post-task ratings and mean accuracy varied across tasks.

For the general knowledge task, there was no statistically significant relationship between post-task ratings and mean success ($ps > 0.05$). There was a small to moderate negative statistically significant relationship between general knowledge post-task ratings and mean confidence, $r(136) = -0.17 – -0.48$, $ps < 0.05$, suggesting that greater perceived difficulty, required effort, and feelings of unpleasantness were related to lower confidence in general

knowledge items. Regarding monitoring accuracy indices, there was a small positive statistically significant relationship between perceived general knowledge task difficulty and resistance to overconfidence, such that greater perceived difficulty was related to greater calibration, $r(136) = 0.21$, $p < 0.05$. There was also a small negative statistically significant relationship between perceived general knowledge task difficulty and resolution. Greater perceived difficulty was therefore related to a poorer ability to discriminate between correct versus incorrect items, $r(136) = -0.21$, $p < 0.05$.

On the financial calculation task, there was a moderate negative statistically significant relationship between post-task ratings and both mean success and mean confidence $r(136) = -0.36 – -0.47$, $ps < 0.01$). These findings suggest that greater perceived difficulty, required effort, and feelings of unpleasantness were related to lower success and confidence in financial calculation items. Regarding monitoring accuracy indices, there was a small positive statistically significant relationship between perceived effort required on financial calculation items and resolution, $r(136) = 0.19$, $p < 0.01$, suggesting that greater required effort was related to greater discrimination between correct and incorrect items. There were no other statistically significant associations between post-task ratings and indices of monitoring accuracy, for the financial calculation task.

For the probability calculation task, there was a moderate negative statistically significant relationship between effort required and mean success rate, $r(136) = -0.25$, $p < 0.01$, such that greater perceived effort was associated with poorer performance. There were no statistically significant relationships between accuracy, task difficulty, and feeling of effort ($ps > 0.05$). There was a small negative statistically significant relationship between probability calculations post-task ratings and mean confidence $r(136) = -0.25 – -0.27$, $ps < 0.01$). Greater perceived

difficulty, required effort, and feelings of effort were therefore related to lower confidence in probability calculation items. Regarding monitoring accuracy indices, there was a small positive statistically significant relationship between feeling of effort and resistance to overconfidence, suggesting that greater unpleasantness was related to greater calibration, $r(136) = 0.20$, $p < 0.05$. There were no statistically significant relationships between resistance to overconfidence, task difficulty, and required effort (ps > 0.05). There were no statistically significant relationships between resolution, task difficulty, required effort, and feeling of effort (ps > 0.05).

Self-perceptions and real-world outcomes. A final series of potential correlates explored in the present study were self-perception measures of academic self-concept and cognitive test anxiety. Real-world outcomes of academic average and learning challenges were also considered as further, exploratory analyses. Descriptive statistics including means and standard deviations are displayed in Table 16. Pearson correlations among the measures are presented in Table 17.

Table 16
Descriptive Statistics for Self-Perceptions and Real-World Outcomes

	Mean (SD)	Potential Range	Observed Range	Skewness	Kurtosis
Academic self-concept	87.78 (16.97)	22 – 132	43 – 127	-0.11	-0.18
Cognitive test anxiety	86.13 (22.00)	27 – 162	35 – 155	0.51	0.55
Academic average	3.85 (0.79)	1 – 5	2 – 5	-0.22	-0.45
Learning challenges	1.46 (.62)	1 – 4	1 – 4	1.24	1.39

Note. $N = 136$

As expected, academic self-concept and anxiety were negatively related, such that poorer academic self-concept was related to greater cognitive test anxiety, $r(136) = 0.72$, $p > 0.01$. Cognitive test anxiety was also negatively related to learning challenges, such that increased self-

reported learning difficulties were related to less cognitive test anxiety ($p > 0.05$). Learning difficulties were also negatively related to academic average, indicating that poorer grades were associated with increased learning challenges ($p > 0.05$).

Table 17
Correlations Between Self-Perceptions and Real-World Outcome Variables

	1	2	3	4
1. Academic self-concept	--	-0.72**	0.02	0.01
2. Cognitive test anxiety		--	-0.05	-0.20*
3. Academic average			--	-0.23*
4. Learning challenges				--

*$p < .05$. **$p < .01$.

Pearson's correlations assessed the relationships between self-perceptions, success rates and confidence, monitoring accuracy indices, and real-world outcomes (see Table 18). For the general knowledge task, there was a small negative correlation between academic self-concept and mean success rate, $r(136) = -0.21$, $p < 0.05$, such that greater accuracy on general knowledge items was related to poorer academic self-concept. Regarding confidence, there was a small positive correlation between self-reported academic average and confidence, $r(136) = 0.29$, $p < 0.01$, and a moderate negative correlation between self-reported learning challenges and confidence, $r(136) = -0.31$, $p < 0.01$. These findings support that greater confidence in general knowledge items was associated with greater academic averages and with less self-reported learning challenges.

For the financial calculation task, there was a small positive correlation between mean accuracy and cognitive test anxiety, $r(136) = 0.20$, $p < 0.05$, suggesting that increased test anxiety was related to greater accuracy. There was a small negative correlation between mean accuracy

Table 18
Correlations Between Self-Perceptions, Real-World Outcomes, Mean Success Rate, Mean Confidence, Resistance to Overconfidence, and Resolution

	Academic Self-Concept	Cognitive Test Anxiety	Academic Average	Learning Challenges
General Knowledge Task				
Mean success rate	-0.21*	0.08	0.17	-0.16
Mean confidence	-0.18	0.13	0.29**	-0.32**
Resistance to overconfidence[a]	-0.04	-0.04	-0.03	0.08
Resolution	-0.09	-0.09	0.09	-0.17
Financial Calculation Task				
Mean success rate	0.02	0.20*	0.06	-0.20*
Mean confidence	0.01	0.02	0.02	-0.17
Resistance to overconfidence[a]	0.06	0.04	0.11	-0.11
Resolution	-0.04	0.02	-0.01	-0.16
Probability Calculation Task				
Mean success rate	0.04	0.02	0.16	-0.13
Mean confidence	0.02	-0.05	0.00	-0.10
Resistance to overconfidence[a]	-0.04	0.09	0.18*	-0.15
Resolution	0.00	0.06	0.02	-0.07

[a] A higher score indicated better calibration.
*$p < .05$. **$p < .01$.

and self-reported learning challenges, $r(136) = -0.20$, $p < 0.05$, supporting that greater reported learning challenges were related to poorer performance. There were no statistically significant relationships between confidence and self-perceptions on the financial calculation task (ps > 0.05). There were also no statistically significant relationships between success rate, confidence, and self-perception measures on the probability calculation task (ps > 0.05).

Contrary to the present's study initial hypothesis, academic self-concept and cognitive test anxiety were not related to indices of monitoring accuracy, across all experimental tasks (ps > 0.05). Similarly, self-reported academic average and learning challenges were largely not associated with monitoring accuracy indices (ps > 0.05), with the exception of a small statistically significant relationship found between self-reported academic average and resistance to overconfidence on the probability calculation task, $r(136) = 0.18$, $p < 0.05$.

Predictors of Monitoring Accuracy and Real-World Outcomes

Two series of regressions analyses were then conducted in order to better understand the relationships between monitoring accuracy, its correlates, and real-world monitoring accuracy outcomes. The first set of regressions included predicting calibration and resolution from cognitive abilities, task-specific judgments, and self-perception variables. The second set of regressions included predicting real-world outcomes of academic grades and self-reported learning challenges from indices of monitoring accuracy, cognitive abilities, and self-perceptions.

Assumptions. Prior to each regression, diagnostic tests were conducted to determine whether regressions were a viable procedure. As with all multiple regressions, assumptions of independence of observations, linearity, homoscedasticity, absence of multicollinearity, and lack of unusual points (i.e., outliers, high leverage point, highly influential points) were verified.

Given that each value of the outcome variables came from a separate case, independence of observations was assumed. There was also independence of residuals, as assessed by a Durbin-Watson statistic that ranged from 1.92 – 2.11 across regression analyses. Assumptions of linearity (i.e., predictor variables are collectively and independently linearly related to all outcome variables), homoscedasticity (i.e., equal residuals for all values of the outcome variables), and normally distributed errors (i.e., normal distribution of the residuals) were explored graphically, with no violations found. Variance inflation factor (VIF) values of less than 10 indicated a lack of multicollinearity. Finally, the absence of unusual points was verified (i.e., scores were less than three standard deviations from the mean, leverage values were below 0.20, and Cook's Distance values were above 1).

Predictors of calibration and resolution. Due to their centrality in the present study, predictors of calibration and resolution were further explored by using simultaneous regression analyses to identify unique predictors of monitoring accuracy. Cognitive abilities (i.e., working memory and intelligence composite scores), the various aggregated judgments, and a Self-Perceptions Raw Composite Score (i.e., summed and standardized cognitive test anxiety and academic self-concept scores) were entered as predictors of calibration and resolution on each of the monitoring accuracy tasks. The results of these regression analyses appear in Tables 18 and 19.

Regression analyses on calibration. Calibration was represented by the resistance to overconfidence index, where greater scores indicated better calibration. Regressions analyses to predict calibration are summed in Table 19.

Predicting general knowledge calibration. A multiple regression was run to predict resistance to overconfidence on the general knowledge task from cognitive abilities, aggregated

pre- and post-task judgment scores, and the self-perceptions raw composite score. The multiple regression model statistically significantly predicted resistance to overconfidence on the general knowledge task, $F(8,126) = 3.16$, $p = 0.003$, $R^2 = 0.17$, adjusted $R^2 = 0.11$. Both working memory, $\beta = 0.26$, $p = 0.005$, and ratings of post-task difficulty, $\beta = 0.26$, $p = 0.04$, uniquely contributed to predicting resistance to overconfidence, accounting for 6% and 5% of the variability in general knowledge calibration, respectively.

Table 19
Simultaneous Regression Results for Resistance to Overconfidence

	B	SE_B	β	t	Variance explained
Criterion Variable = Resistance to Overconfidence[a] on General Knowledge Test					
Working memory raw score	0.01	0.00	0.25	2.86 **	6%
Intelligence raw composite z-score	0.00	0.01	0.13	1.58	2%
Predictive confidence judgement	-0.01	0.01	-0.13	-1.22	1%
Postdictive confidence judgement	-0.00	0.01	-0.07	-0.58	<1%
Task difficulty	0.01	0.01	0.22	2.05*	5%
Effort required	-0.00	0.01	-0.03	-0.34	<1%
Feeling of effort	0.01	0.01	0.07	0.67	<1%
Self-perceptions raw composite z-score	-0.00	0.01	0.02	0.20	<1%
Overall Regression: F (8, 126) = 3.16**					
Multiple R = 0.41					
Multiple R^2 = 0.17					
Adjusted R^2 = 0.11					
Criterion Variable = Resistance to Overconfidence[a] on Financial Calculation Test					
Working memory raw score	.00	0.10	0.10	1.13	<1%
Intelligence raw composite z-score	.02	0.00	0.20	2.13*	4%
Predictive confidence judgement	.01	0.01	0.11	1.03	1%
Postdictive confidence judgement	-0.00	0.01	-0.08	-0.63	<1%
Task difficulty	-0.00	0.01	-0.06	-0.46	<1%

Effort required	-0.01	0.01	-0.17	-1.37	3%
Feeling of effort	0.01	0.01	0.16	1.39	3%
Self-perceptions raw composite z-score	0.01	0.01	0.12	1.31	1%
Overall Regression: F (8, 126) = 1.67					
Multiple R = 0.31					
Multiple R^2 = 0.10					
Adjusted R^2 = 0.04					
Criterion Variable = Resistance to Overconfidence[a] on Probability Calculation Test					
Working memory raw score	.00	0.00	0.02	0.19	<1%
Intelligence raw composite z-score	.03	0.01	0.30	3.42***	8%
Predictive confidence judgement	-0.00	0.01	-0.03	-0.24	<1%
Postdictive confidence judgement	-0.00	0.00	-0.09	-0.90	<1%
Task difficulty	0.01	0.01	0.12	1.03	<1%
Effort required	-0.16	0.01	-0.23	-1.97*	3%
Feeling of effort	0.17	0.01	0.23	2.42**	5%
Self-perceptions raw composite z-score	-0.01	0.01	0.09	-0.61	<1%
Overall Regression: F (8, 126) = 3.35**					
Multiple R = 0.41					
Multiple R^2 = 0.17					
Adjusted R^2 = 0.12					

Note. B = unstandardized regression coefficient; SE_B = standard error of the coefficient; β = standardized coefficient.
[a] A higher score indicated better calibration.
*$p < .05$. **$p < .01$. ***$p < .001$.

Predicting financial calculation calibration. A similar multiple regression was run to predict resistance to overconfidence on the financial calculation task from cognitive abilities, aggregated pre- and post-task judgment scores, and the self-perceptions raw composite score. The complete regression model did not statistically significantly predict resistance to overconfidence on the financial calculation task, $F(8, 126) = 1.67$ $p = 0.11$. The intelligence composite score however did significantly predict resistance to overconfidence, $\beta = 0.20$, $p = 0.04$, accounting for 4% of the variability in financial calculation calibration.

Predicting probability calculation calibration. A multiple regression was run to predict resistance to overconfidence on the probability calculation task from cognitive abilities, aggregated pre- and post-task judgment scores, and the self-perceptions raw composite score. The multiple regression model statistically significantly predicted resistance to overconfidence on the probability calculation task, $F(8, 126) = 3.25$, $p = 0.002$, $R^2 = 0.17$, adjusted $R^2 = 0.12$. The intelligence raw score composite, $\beta = 0.29$, $p < 0.001$, as well as post-task ratings of effort, $\beta = -0.23$, $p = 0.04$, and feeling of effort, $\beta = 0.23$, $p = 0.01$, significantly contributed to predicting resistance to overconfidence, accounting for 8%, 3%, and 5% of the variability in probability calculation calibration, respectively.

Regression analyses on resolution. Resolution was measured with the Goodman–Kruskal Gamma, where greater scores indicated a greater ability to discriminate between correct versus incorrect responses. Regression analyses to predict resolution are summed in Table 20.

Predicting general knowledge resolution. A multiple regression was run to predict general knowledge resolution from cognitive abilities, aggregated pre- and post-task judgment scores, and the self-perceptions raw composite score. The multiple regression model statistically significantly predicted general knowledge resolution, $F(8,126) = 2.19$, $p < 0.05$, $R^2 = 0.12$, adjusted $R^2 = 0.07$. The intelligence raw score composite, $\beta = 0.23$, $p < 0.01$, was the sole predictor uniquely contributing to predicting resolution, accounting for 4% of the variability in general knowledge resolution.

Predicting financial calculation resolution. A multiple regression was run to predict probability calculation resolution from cognitive abilities, aggregated pre- and post-task judgment scores, and the self-perceptions raw composite score. The regression model did not statistically significantly predict financial calculation resolution, $F(8,126) = 1.42$, $p = 0.20$.

Working memory raw scores however did significantly predict financial calculation resolution, β = -0.19, $p < 0.05$, accounting for 4% of the variability.

Table 20
Simultaneous Regression Results for Resolution

	B	SE_B	*β*	t	Variance explained
Criterion Variable = Resolution on General Knowledge Test					
Working memory raw score	0.00	0.00	0.05	0.57	<1%
Intelligence raw composite z-score	0.04	0.02	0.23	2.65**	4%
Predictive confidence judgement	0.00	0.01	0.02	0.14	<1%
Postdictive confidence judgement	0.02	0.02	0.13	1.09	<1%
Task difficulty	-0.02	0.02	-0.06	-0.54	<1%
Effort required	-0.02	0.01	-0.10	-1.01	<1%
Feeling of effort	0.00	0.01	0.02	0.23	<1%
Self-perceptions raw composite z-score	0.05	0.02	-0.06	-0.67	<1%
Overall Regression: F (8,126) = 2.19*					
Multiple R = 0.35					
Multiple R^2 = 0.12					
Adjusted R^2 = 0.07					
Criterion Variable = Resolution on Financial Calculation Test					
Working memory raw score	-0.20	0.01	-0.19	-2.07*	4%
Intelligence raw composite z-score	0.10	0.02	0.05	0.49	<1%
Predictive confidence judgement	0.01	0.02	0.04	0.36	<1%
Postdictive confidence judgement	-0.02	0.01	-0.16	-1.24	3%
Task difficulty	-0.02	0.02	-0.02	-0.13	<1%
Effort required	0.03	0.02	0.18	1.48	3%
Feeling of effort	-0.02	0.02	-0.10	-0.90	1%
Self-perceptions raw composite z-score	-0.00	0.03	-0.01	-0.06	<1%
Overall Regression: F (8,117) = 1.43					
Multiple R = 0.30					
Multiple R^2 = 0.09					

Adjusted $R^2 = 0.03$					
Criterion Variable = Resolution on Probability Calculation Test					
Working memory raw score	-0.04	0.01	-0.19	-2.05*	4%
Intelligence raw composite z-score	0.04	0.02	0.05	0.49	<1%
Predictive confidence judgement	0.00	0.01	0.04	0.37	<1%
Postdictive confidence judgement	0.00	0.01	-0.16	-1.25	3%
Task difficulty	0.00	0.02	-0.02	-0.13	<1%
Effort required	-0.02	0.02	0.18	1.45	3%
Feeling of effort	0.00	0.01	-0.01	-0.89	<1%
Self-perceptions raw composite z-score	-0.00	0.03	-0.02	-0.16	<1%
Overall Regression: F (8,126) = 1.42					
Multiple R = 0.29					
Multiple $R^2 = 0.09$					
Adjusted $R^2 = 0.03$					

Note. B = unstandardized regression coefficient; SE_B = standard error of the coefficient; β = standardized coefficient.
*$p < .05$. **$p < .01$.

Predicting probability calculation resolution. A multiple regression was run to predict probability calculation resolution from cognitive abilities, aggregated pre- and post-task judgment scores, and the self-perceptions raw composite score. The regression model did not statistically significantly predict probability calculation resolution, $F(8, 125) = 1.45$, $p = 0.22$. Working memory however did significantly predict probability calculation resolution, $\beta = -0.19$, $p < 0.05$, accounting for 4% of the variability.

Taken together, the most consistent predictors of both calibration and resolution were cognitive abilities, such that either greater working memory total scores or greater intelligence raw composite scores predicted greater resistance to overconfidence and resolution scores, across experimental tasks. All other predictors rarely had unique explanatory power.

Predictors of real-world outcomes. A second series of multiple regression analyses were conducted in order to further explore and identify unique predictors of real-world outcomes. Cognitive abilities (i.e., working memory and intelligence composite scores), indices of monitoring accuracy on each of the monitoring accuracy tasks (i.e., resistance to overconfidence and resolution), and the self-perceptions raw composite score were entered as predictors of self-reported academic grades and learning challenges. The results of these regression analyses appear in Tables 21 and 22.

Regression analyses on academic grades. Regressions analyses to predict academic grades are summarized in Table 21.

Table 21
Simultaneous Regression Results for Academic Grades

	B	SE_B	β	t	Variance explained
Regression #1					
Working memory raw score	0.01	0.02	0.06	0.67	<1%
Intelligence raw composite z-score	0.10	0.05	0.19	2.02*	4%
Resistance to Overconfidence[a] on General Knowledge Test	-0.41	0.62	-0.07	-0.79	<1%
Resolution on the General Knowledge Test	0.07	0.27	0.04	0.48	<1%
Self-perceptions raw composite z-score	0.04	0.07	-0.04	-0.42	<1%
Overall Regression: F (5, 123) = 1.08					
Multiple R = 0.21					
Multiple R^2 = 0.04					
Adjusted R^2 = 0.03					
Regression #2					
Working memory raw score	0.00	0.02	0.01	0.14	<1%
Intelligence raw composite z-score	0.02	0.01	0.19	1.98*	4%
Resistance to Overconfidence[a] on the Financial Calculation Test	0.50	0.65	0.07	0.78	<1%
Resolution on the Financial Calculation Test	-0.00	0.24	-0.00	-0.02	<1%
Self-perceptions raw composite z-score	-0.02	0.07	-0.02	-0.25	<1%
Overall Regression: F (5,115) = 1.32					

Multiple R = 0.23					
Multiple R^2 = 0.05					
Adjusted R^2 = 0.01					
Regression #3					
Working memory raw score	0.01	0.02	0.04	0.41	<1%
Intelligence raw composite z-score	0.09	0.05	0.16	1.99*	3%
Resistance to Overconfidence[a] on the Probability Calculation Test	0.93	0.60	0.14	1.54	2%
Resolution on the Probability Calculation Test	-0.14	0.25	-0.05	-0.56	<1%
Self-perceptions raw composite z-score	-0.03	0.07	-0.04	-0.49	<1%
Overall Regression: F (5, 123) = 1.55					
Multiple R = 0.24					
Multiple R^2 = 0.06					
Adjusted R^2 = 0.02					

Note. B = unstandardized regression coefficient; SE_B = standard error of the coefficient; β = standardized coefficient.
[a] A higher score indicated better calibration.
*$p < .05$. **$p < .01$.

Predicting academic grades from general knowledge task. A multiple regression was run to predict academic grades from cognitive abilities, indices of monitoring accuracy from the general knowledge task, and the self-perception raw composite score. The regression model was not statistically significant, $F(5, 123) = 1.07$, $p = 0.38$. However, the intelligence raw score composite, $\beta = .19$, $p < 0.05$ significantly contributed to academic grades, accounting for 3% of the variability.

Predicting academic grades from financial calculation task. A multiple regression was run to predict academic average from cognitive abilities, indices of financial calculation monitoring accuracy, and self-perception raw composite scores. The regression model was not statistically significant, $F(5, 123) = 1.10$, $p = 0.36$. Intelligence raw composite scores however did significantly contribute to predicting academic grades, $\beta = 0.19$, $p < 0.05$, accounting for 4% of the variability.

Predicting academic grades from probability calculation task. A multiple regression was run to predict academic average from cognitive abilities, indices of probability calculation monitoring accuracy, and self-perception raw composite scores. The regression model was not statistically significant, $F(5, 123) = 1.55$, $p = 0.18$. Intelligence raw composite scores however did significantly contribute to predicting academic grades, $\beta = 0.16$, $p < 0.05$, accounting for 3% of the variability.

Across regression analyses, only the intelligence raw composite scores emerged as a significant predictor of academic grades.

Regression analyses on learning challenges. Regression analyses to predict self-reported learning challenges are summarized in Table 22.

Table 22
Simultaneous Regression Results for Learning Challenges

	B	SE_B	β	t	Variance explained
Regression #1					
Working memory raw score	-0.00	0.01	-0.01	-0.10	<1%
Intelligence raw composite z-score	-0.02	0.00	-0.16	-1.85	2%
Resistance to Overconfidence[a] on the General Knowledge Test	0.37	0.47	0.07	0.78	<1%
Resolution on the General Knowledge Test	-0.28	0.20	-0.12	-1.39	2%
Self-perceptions raw composite z-score	-0.16	0.05	-0.26	-3.09**	7%
Overall Regression: F (5,126) = 1.96					
Multiple R = 0.27					
Multiple R^2 = 0.07					
Adjusted R^2 = 0.04					
Regression #2					
Working memory raw score	0.00	0.01	0.05	0.51	<1%
Intelligence raw composite z-score	-0.02	0.01	-0.18	-1.97	3%
Resistance to Overconfidence[a] on the Financial Calculation Test	-0.35	0.49	-0.06	-0.71	<1%
Resolution on the Financial Calculation Test	0.34	0.18	0.17	1.92	3%
Self-perceptions raw composite z-score	-0.16	0.06	-0.26	-2.92**	7%

Overall Regression: F (5, 115) = 3.60**					
Multiple R = 0.37					
Multiple R^2 = 0.14					
Adjusted R^2 = 0.10					
Regression #3					
Working memory raw score	0.00	0.01	0.03	0.27	<1%
Intelligence raw composite z-score	-0.01	0.04	0.00	-1.67	2%
Resistance to Overconfidence[a] on the Probability Calculation Test	-0.47	0.48	-0.09	-1.01	<1%
Resolution on the Probability Calculation Test	-0.05	0.20	-0.02	-0.24	<1%
Self-perceptions raw composite z-score	-0.16	0.05	-0.25	-2.90**	6%
Overall Regression: F (5, 124) = 2.99*					
Multiple R = 0.33					
Multiple R^2 = 0.11					
Adjusted R^2 = 0.07					

Note. B = unstandardized regression coefficient; SE_B = standard error of the coefficient; β = standardized coefficient.
[a]A higher score indicated better calibration.
*$p < .05$.

Predicting learning challenges from general knowledge task. A multiple regression was run to predict learning challenges from cognitive abilities, indices of general knowledge monitoring accuracy, and self-perception raw composite scores. The regression model was statistically significant, $F(5, 125) = 3.52$, $p < 0.05$. The self-perceptions raw score composite, $\beta = -0.26$, $p < 0.01$, was the sole predictor of learning challenges, accounting for 7% of the variability.

Predicting learning challenges from financial calculation task. A multiple regression was run to predict learning challenges from cognitive abilities, indices of financial calculation monitoring accuracy, and self-perception raw composite scores. The regression model was statistically significant, $F(5, 115) = 3.60$, $p < 0.01$. The self-perceptions raw score composite, $\beta = -0.26$, $p < 0.01$, was the sole predictor of learning challenges, accounting for 7% of the variability.

Predicting learning challenges from probability calculation task. A multiple regression was run to predict learning challenges from cognitive abilities, indices of probability calculation monitoring accuracy, and self-perception raw composite scores. The regression model was statistically significant, $F(5, 124) = 2.99$, $p < .01$. The self-perceptions raw score composite, $\beta = -0.25$, $p < 0.01$, was the sole predictor of learning challenges, accounting for 7% of the variability.

Across analyses, findings support that cognitive abilities, specifically the intelligence raw composite score, were the most consistent predictor of academic grades, whereas individual differences in self-perceptions was the most significant predictor of learning challenges. All other predictors did not emerge as statistically significant.

Discussion

The present study examined individuals' monitoring accuracy across three tasks from different domains. In addition to general knowledge, two of the tasks used have been rarely examined from a metacognitive perspective: financial calculation and probability calculation. The association between them, and the well-studied general knowledge domain extends the existing literature on understanding the stability of monitoring accuracy, both within individuals and across domains.

Four aims were considered: (1) exploring indices of monitoring accuracy across domains; (2) investigating whether monitoring accuracy scores reflect domain-general or domain-specific abilities; (3) exploring potential correlates as well as predictors of success, confidence and monitoring accuracy, including cognitive abilities, task-specific judgments, and self-perceptions; and (4) examining the relationship between monitoring accuracy and real-world outcomes of academic achievement and learning challenges. Overall, an overconfidence bias was found in all experimental tasks. Calibration scores were also correlated across tasks, suggesting domain-

generality. In contrast, although reliable resolution was obtained in all tasks, resolution was not correlated across tasks, suggesting domain-specificity of discrimination abilities. Associations between task-specific judgments, self-perceptions, and monitoring accuracy indices varied across tasks. Cognitive abilities were the sole consistent predictor of both calibration and resolution. In this study, monitoring accuracy did not predict real-world achievement and learning outcomes. However, cognitive abilities emerged as a predictor of academic grades, whereas self-perceptions predicted learning difficulties. A summary of research aims, hypotheses, and findings are presented in Table 23 and discussed below.

Performance, Confidence, and Monitoring Accuracy

The first goal of the present study was to compare calibration and resolution within experimental tasks. In all three tasks, confidence was not well matched with performance, with a tendency towards over- rather than underconfidence. Participants' confidence in their task performance was therefore greater than their actual performance. This finding is consistent with the overconfidence effect reported in many domains (Dunning, Heath, & Suls, 2004), including the judgment and decision-making literature (Bruine de Bruine et al., 2007; Lichtenstein & Fischhoff, 1977; Stanovich et al., 2016; West & Stanovich, 1997; Yates, Lee & Bush, 1997). Resistance to overconfidence and accuracy were also found to be positively related, such that in each task, better performance was associated with better calibration. Increased success on task items was therefore associated with an increased ability to estimate actual performance. In contrast, resistance to overconfidence and confidence demonstrated a negative relationship, with poorer confidence being associated with better calibration (except for the financial calculation task, which was nonsignificant). This negative association between confidence and calibration can be best understood in the context of the overconfidence effect. If individuals are in fact prone

Table 23
Summary of Study Aims, Hypotheses, and Results

Aims	Hypotheses	Results		
		General Knowledge	Financial Calculation	Probability Calculation
Indices of Monitoring Accuracy	*Hypothesis 1*: Overconfidence bias across all tasks	X	X	X
	Hypothesis 2. Resolution greater than zero across all tasks	X	X	X
Domain-Generality vs Domain-Specificity	*Hypothesis 3a.* Calibration scores correlated across tasks	X	X	X
	Hypothesis 3b. Calibration scores between similar domains reflect stronger correlations		X	X
	Hypothesis 4. Resolution scores correlated across tasks.			
Success, Confidence, and Monitoring Accuracy Correlates	*Hypothesis 5.* Cognitive abilities associated with increased calibration and resolution	X	X	X
	Hypothesis 6a. Predictive confidence positively correlated with: • accuracy • confidence	 X	 X X	 X X
	Hypothesis 6b. Postdictive confidence positively correlated with accuracy and confidence	X	X	X
	Hypothesis 7. Poorer calibration associated with: • higher confidence • increased difficulty • required effort • feeling of effort	 X X		 X X

	Hypothesis 8. Poorer resolution associated with: • higher confidence • increased difficulty • required effort • feeling of effort	X X	X X	
	Hypothesis 9. Poorer calibration associated with poorer academic self-concept and increased cognitive anxiety			
	Hypothesis 10. Poorer resolution scores associated with poorer academic self-concept and increased cognitive anxiety			
	Hypothesis 11. Academic self-concept and cognitive anxiety positively correlated	X	X	X
	Hypothesis 12. Cognitive abilities, task-specific judgments, and self-perceptions predict calibration scores	Cognitive abilities and difficulty ratings only	Cognitive abilities only	Cognitive abilities, effort, and feeling of effort only
	Hypothesis 13. Cognitive abilities, task-specific judgments, and self-perceptions predict resolution scores	Cognitive abilities only	Cognitive abilities only	Cognitive abilities only
Real-World Outcomes	*Hypothesis 14.* Poorer calibration associated with lower grades and greater learning problems			Academic average only
	Hypothesis 15. Poorer resolution associated with lower grades and greater learning problems.			
	Hypothesis 16a. Indices of monitoring accuracy, cognitive abilities, and self-perceptions predict academic achievement	Cognitive abilities only	Cognitive abilities only	Cognitive abilities only
	Hypothesis 16b. Indices of monitoring accuracy, cognitive abilities, and self-perceptions predict learning challenges	Self-perceptions only	Self-perceptions only	Self-perceptions only

Note. X indicates a significant finding.

to overestimating their knowledge, then it makes sense that lower confidence ratings are actually better matched with objective performance. In support of this explanation, a study by Koriat (1980) in which participants confidence judgments were experimentally manipulated, found that a decrease in mean confidence was better matched with mean performance, thus leading to improved calibration scores.

In addition, reliable resolution was found across all three tasks, as indicated by mean resolution scores being significantly different from zero. This supports that participants were meaningfully using the confidence scale and were able to successfully discriminate between correct and incorrect responses. However, associations between resolution, accuracy, and confidence varied across tasks. Thus, in contrast to calibration, resolution did not demonstrate a consistent pattern from one task to another. The finding that each task in isolation showed reliable resolution strengthens the inference of domain-specificity in resolution that emerged across tasks (Jackson et al., 2016).

There were no correlations between calibration and resolution in either the general knowledge or financial calculation tasks. Calibration and resolution were significantly positively correlated in the probability calculation task, although this effect was small. The ability to judge actual performance was therefore not consistently related to the ability to discriminate between correct and incorrect responses across tasks. Consequently, individuals may be able to assess when, for example, knowledge prior to an exam is poor and that they must study in order to increase performance (i.e., good calibration), yet not be able to discriminate between which parts of the material they have internalized, versus which parts would require more review (i.e., no resolution). This distinction between calibration and resolution is of importance when wanting to

target metacognitive abilities, suggesting that different indices of monitoring accuracy may warrant different levels of intervention and support.

Domain-General Calibration, Domain-Specific Resolution

A second study goal was to explore the question of whether metacognitive monitoring represents a general ability, similar across domains, or rather a specific ability that varies as a function of domain considered. Overall, calibration was significantly and positively correlated across experimental tasks, which suggests domain-generality. This finding further supports previous studies showing that calibration is a domain-general construct, reflecting a stable individual difference trait (Jackson & Kleitman, 2014; 2016; Stanovich & West, 1997).

Nonetheless, correlations between calculation tasks were stronger, compared to correlations between general knowledge and either financial or probability calculations. There was also a significant difference between correlation coefficients with the general knowledge task, then with the numeracy-based tasks. Thus, although calibration demonstrated domain-generality, there may be some sensitivity to task domain. That is, tasks in which the underlying content knowledge is more closely related may display more analogous calibration abilities, compared to tasks that rely on different knowledge bases (e.g., general knowledge versus calculations). The idea that the *strength* of relatedness, rather than its existence, can vary across knowledge domains may help consolidate inconsistent findings in the literature which support both domain-generality (e.g., Erickson & Heit, 2015; Scott & Berman, 2013; Veenman and Verheij, 2001) and domain-specificity (e.g., Fitzgerald, Arvaneh, & Dockree, 2017; Glaser, 1991). A series of studies exploring monitoring accuracy across various domains conducted by Schraw and colleagues (1995) lends some support to this view. In the first experiment of this study, Schraw and colleagues (1995) used experimental tasks of more diverse domains, whereas

in the second experiment task knowledge and content were more closely related. Although authors concluded a general tendency for domain-generality of confidence, they noted that correlations between similar tasks were stronger, compared to those reflecting a more diverse set of content knowledge (Schraw et al., 1995). An interaction between person-related factors and content knowledge was therefore proposed, in which domain-general processes may support domain-specific performance (Schraw et al., 1995). That is, although calibration abilities display a similar pattern across diverse domains, individual expertise in domain considered may impact this general metacognitive trait.

In contrast, there was a lack of correlations in resolution, across experimental tasks. The ability to discriminate between correct and incorrect responses on one task was therefore not related to the ability to discriminate between correct and incorrect responses on another task. Given that previous studies have compared resolution across sets of similar tasks (Ackerman & Beller, 2007; Finn & Metcalfe, 2008), findings from the present study suggest that when tasks are diverse, resolution might be domain-specific, rather than general. That is, the same individual may have strong resolution abilities on a general knowledge task, but less on a calculation task. This further supports that, in contrast to calibration, the ability to discriminate right from wrong answers may vary across tasks (Jackson, Kleitman, Howie, & Stankov, 2016). This distinction reinforces previous claims that calibration and resolution are conceptually different and empirically separable indicators of monitoring accuracy (Ais, Zylberberg, Barttfeld, & Sigman, 2016; Koriat et al. 2002; Koriat, 2012b; Maki et al., 2005; Thiede et al., 2015) and can yield important insights into interventions targeting metacognitive skill and development. For example, if the goal is to increase general metacognitive accuracy skills, calibration abilities should be targeted and supported, as they can apply across areas. In contrast, if the goal is to

increase performance in a given domain, discrimination skills relating to specific content knowledge should be the goal. Calibration and resolution are therefore complementary measures, both critical for determining effective effort regulation (see Ackerman, Parush, Nassar, & Shtub, 2016, for a review).

Taken together, study findings support that calibration is a person-centered, domain general metacognitive ability that is relatively consistent across areas. In contrast, resolution represents a domain-specific ability, sensitive to content knowledge and task demands. The assessment of monitoring accuracy is therefore contingent on the metric used for assessment. Calibration is a measure of absolute accuracy, representing metacognitive precision, whereas resolution is a measure of relative accuracy, reflecting metacognitive consistency (Schraw, 2009). As such, an individual's performance may be precise yet inconsistent, and vice versa, with both metrics impacting accuracy and performance. A highly accurate individual would need to possess well-developed abilities across both indices of calibration and resolution. Considering these findings, both domain-specific and domain-general processes seem to be in operation when preforming in different domains (Schraw et al., 1995).

Cognitive Abilities Associated with and Predictive of Monitoring Accuracy

A third goal of the current study was to explore various potential correlates of success, confidence, and monitoring accuracy. In general, there was a positive relation between cognitive abilities (either working memory and/or intelligence composite score) and accuracy, calibration, and resolution. That is, participants with greater cognitive abilities were not only more successful on tasks items, but were also better able to estimate their performance and to discriminate between correct versus incorrect responses. These findings corroborate previous research linking both general knowledge and calculation skills with increased intelligence (Stanovich & West,

1998; Peters & Bjalkerbring, 2015) and can be attributed to mechanisms of cognitive decoupling (Stanovich, 2011). Cognitive decoupling requires individuals to simulate alternative worlds and to consider hypothetical scenarios that are not in the immediate environment, necessitating engagement of analytic processes to construct these scenarios. However, cognitive decoupling can be challenging to achieve as decoupling mechanisms must be continuously ongoing and demand that mental simulations be sustained while keeping hypothetical scenarios decoupled (Stanovich, 2011). For these reasons, better decoupling abilities can be indexed with measures of working memory and intellectual abilities (Stanovich, 2011; West & Stanovich, 1997). The current study supports that monitoring accuracy requires some cognitive decoupling, as monitoring accuracy also requires simulating and hypothetical reasoning in order to best evaluate current states of knowledge. Some cognitive decoupling may therefore be required to achieve better monitoring accuracy, which supports the association between increased monitoring accuracy and cognitive abilities.

Intelligence raw composite scores were also positively associated with greater confidence. However, the relationship between confidence and intelligence varied depending on task. That is, confidence on the general knowledge task was positively associated with verbal, rather than nonverbal abilities. In contrast, confidence on both financial and probability calculation tasks were positively associated with nonverbal, rather than verbal abilities. These findings further support that, as designed, tasks developed for this study tapped into diverse domains, such that general knowledge may have relied more heavily on crystalized abilities, whereas calculation tasks may have required greater fluid reasoning abilities (Shipley et al., 2009).

When exploring potential monitoring accuracy predictors, cognitive abilities emerged as a statistically significant variable. Across all experimental tasks, higher cognitive abilities predicted better calibration. These findings are consistent with other empirical studies that have examined overconfidence (Bruine de Bruin et al., 2007; Stanovich & West, 1998; Stanovich et al., 2016). Similarly, across all experimental tasks, higher cognitive abilities predicted better resolution, congruent with previous research supporting the relationship between intelligence and the ability to successfully discriminate between correct versus incorrect responses (Jackson & Kleitman, 2014). Overall, better cognitive abilities were associated with, and predictive of, better performance and monitoring accuracy.

Predictive and Postdictive Confidence, Difficulty, and Effort

A second set of potential correlates explored in the current study were task-specific judgments, including pre-and postdictive confidence judgments and post-task ratings of difficulty, effort required, and feeling of effort.

Across tasks, predictive and postdictive judgments were positively correlated to both confidence and accuracy. This corroborates previous findings positively linking predictive and postdictive judgments of performance to increased accuracy, across a variety of domains (Erikson & Heit, 2015). The relationships between confidence and post-task ratings were also similar across tasks, with greater confidence being negatively associated with task difficulty, effort required, and feeling of effort. That is, the more confident in performance, the less difficult and effortful the task was experienced. Ackerman and Zalmanov (2012) demonstrated that quickly produced answers, such as in the case of multiple choice responses, are generally accompanied by higher confidence ratings. Building on this finding, it can be suggested that rapidly generated responses are also perceived as less difficult and less effortful.

In contrast, there were no clear patterns that emerged in the relationships between accuracy and post-task ratings of difficulty, effort, and feeling of effort. The one exception was that across both financial and probability calculation tasks, success rate was negatively associated to perceptions of effort. That is, poorer accuracy on calculation items, but not on general knowledge items, was related to increased perceived effort. It can be hypothesized that increased effort exertion on calculation items was experienced as a result of the computational demands of these tasks. Responses on the general knowledge task relied on memory retrieval processes, whereas responses on the calculation task required additional computational skills and response generation.

Across domains, overall task-specific judgments did not emerge as predictors of monitoring accuracy. Two exceptions to this trend included ratings of task difficulty on the general knowledge task, as well as ratings of perceived effort and feeling of effort on the probability task, emerging as predictors of calibration. These findings suggest that perceived workload and effort required may importantly bear on monitoring accuracy in certain domains. The NASA Task Load Index (TLX; Hart & Staveland, 1988) has been used to index individual differences in perceived workload, which may be useful to include in future studies, to better our understanding of this potential association.

Academic Self-Concept and Cognitive Test Anxiety

A final set of potential monitoring accuracy correlates explored in this study included individual self-perceptions of academic self-concept and cognitive test anxiety. Contrary to initial hypotheses, no significant correlations were found between self-perceptions and monitoring accuracy. Previous studies however have suggested that cognitive biases negatively impact individual confidence levels (Koriat, Lichtenstein, & Fischhoff, 1980). In a study by

Erickson (2015), it was found that math anxiety becomes disruptive to metacognitive processes as a function of task difficulty, with easier tasks being less disruptive. Considering the overall success rate of present study tasks (general knowledge = 53.19%, financial calculation = 62.99%, probability calculation = 58.89%), it can be hypothesized that the impact of cognitive test anxiety may have emerged with more complex and difficult tasks. Furthermore, disposition measures explored in this study were narrowed, targeting self-perceptions about a particular task type (i.e., test-taking) and domain (i.e., academics). Perhaps an inclusion of broader considerations would have yielded different results. For example, in a study by Erickson (2015) focused on math anxiety, it was suggested that considering a less specific view of anxiety, such as generalized anxiety, would better capture the effect of anxiety on metacognitive abilities, particularly through its association with decreased working memory.

Real-World Outcomes: Academic Achievement and Learning Challenges

A final study goal was to explore monitoring accuracy, cognitive abilities, and self-perceptions as potential predictors of real-world outcomes of academic grades and learning challenges. Across experimental tasks, intelligence predicted academic grades, showing that higher intelligence was correlated with higher academic average. Neither calibration, resolution, nor self-perceptions however emerged as predictors of academic grades. Regarding learning challenges, only self-perceptions emerged as a significant predictor of self-reported difficulties. This finding is in line with previous literature supporting that poorer self-perceptions and self-efficacy beliefs negatively impact learning (Stankov et al., 2012).

However, contrary to initial hypotheses, indices of monitoring accuracy did not significantly predict any of the real-world outcomes considered in this study. Studies that have explored real-world consequences of monitoring accuracy have largely focused on decision-

making, such as composite scores that include a diverse set of real world outcomes, such as whether to throw out food or have a mortgage or loan foreclosed (Bruine de Bruin et al., 2007). These studies have found that poorer monitoring accuracy leads to poorer outcomes. Perhaps a consideration related to achievement or learning-related outcomes may be associated with variables pertaining to conscientious behaviours, such as whether to seek additional support, attend lectures, or investment of more study time. It can be hypothesized that if individuals are overconfident in their academic and learning abilities or unable to discriminate between areas in which they have mastered knowledge and those they have not, that in turn they would be less likely to engage in choices and behaviours that would favor the improvement of their learning and performance.

Taken together, present findings suggest that when wanting to identify individuals with poorer academic grades or learning difficulties, indices of monitoring accuracy may not be the most ideal variables to consider. Nonetheless, it is important to note that both academic average and learning challenges in this present study were self-reported and identified. Given that the study sample showed a tendency towards overconfidence, individuals may have not accurately reported their learning difficulties, thus masking results. Future studies should consider more objective measures of both grades and learning difficulties.

Limitations, Implications, and Future Directions

Limitations.

Calculating indices. Methodological and scoring considerations of monitoring accuracy indices should be taken into account. In the resolution literature, variability in item difficulty is critically important, but this screening is not used in the overconfidence literature. To be able to calculate the gamma index and whether a participant can discriminate between correct and

incorrect items, there needs to be variability in confidence judgments, such as assigning high confidence to correct responses and low confidence to incorrect responses (Fleming & Lau, 2014). If, for example, a participant assigns high confidence to all responses, not only is there lack of variability, but this response pattern can also suggest that the participant is not using the confidence scale meaningfully and is assigning arbitrary judgments. When calculating calibration however, this is less of a methodological concern, as the overconfidence formula does not require variability to be interpreted meaningfully. In the context of calibration, it is possible that a participant assigns high or low confidence ratings to all responses, due to a strong over- or underconfidence bias.

In order to use the same set of responses and participants, and directly compare findings across accuracy monitoring indices, the current study eliminated items that had very low or very high success rates (see Tables 2, 4, and 6). To our knowledge, this is the first study to directly and empirically compare calibration and resolution indices to address issues of domain-generality versus specificity. What was considered the most conservative approach was used, in order to directly compare these indices, and demonstrates the feasibility of this experimental strategy.

Another potential approach to comparing calibration and resolution could have been with the use of hybrid scores. In order to target calculation differences between absolute (i.e., calibration) and relative (i.e., resolution) accuracy, some researchers have proposed the use of hybrid indices (Schraw, 1990), such as the Brier score (Karen, 1991; Yates, 1990). Considering findings from this study that elucidate critical differences between calibration and resolution, the use of a synthesized score may have led to unrepresentative findings (Schraw, 1990). Evidenced by the current study, calibration and resolution reflect very different metacognitive monitoring

processes and should be considered separately. Future studies should aim for a more in-depth consideration of issues relating to scoring criteria, inclusion and exclusion of items, and choice of metacognitive indices, which may have important implications.

Sample size. A second limitation of the current study is the sample size considering the number of completed analyses. Given the novelty of our research question and methodology, current hypotheses needed to be approached in a systematic way. That is, the first step was to replicate findings (i.e., significant overconfidence and resolution) and establish the current study as being consistent with the existing literature. The second step was to then build on findings and empirically explore issues of domain generality and specificity. As a result of this dual approach, a greater number of hypotheses and analyses were conducted. In order to circumvent this potential issue of sample size, a Bonferroni correction was applied to our targeted aim of cross-domain index comparison (see Tables 10 and 11) and findings remained significant. For other exploratory analyses, effect sizes rather than p values can be used when interpreting results, as effect sizes are independent of sample size (Kline, 2005). Future research may wish to include an a priori power analysis in order to support sample size choice.

Another potential sample-related issue is the predominantly female sex distribution of the current sample. The current study did not find any sex differences in monitoring accuracy, although this finding may have been impacted by the skewed female sample. Future studies should aim for a more even distribution and further explore any potential sex differences.

Item difficulty. A third consideration critical for future studies would be that of item difficulty. Difficulty can impact confidence ratings, such that individuals tend to be underconfident in easy questions and overconfident in harder questions (Juslin et al., 2000). One of the main reasons why domain-generality has been less explored within the literature is the

challenge in designing experimental tasks from differing domains that have items of similar difficulty. Although the current study developed tasks that were close in difficulty, thus providing evidence that that inter-domain comparisons are possible, continued efforts to control for item difficulty will be of importance in future comparisons. Furthermore, given the above-described methodological challenges of calculating indices of calibration and resolution, consideration of not only similar item difficulty between domains, but also of similar variability will be of importance.

General knowledge task reliability. In the current study, the general knowledge task yielded low internal consistency. Future studies can improve on issues of task reliability by including a larger set of general knowledge task items and by increased piloting of items.

Small variance explained. Although significant predictors did emerge when predicting monitoring accuracy indices, resulting variance explained were small. Future research may wish to include other potential predictors of monitoring accuracy.

Task format. Monitoring accuracy can be sensitive to test format, such that open-ended responses may yield different results, compared to multiple choice. In a multiple-choice format, participants have greater opportunity to consider each response option, to recognize a correct response or to successfully guess the correct answer (Ackerman, 2019). In contrast, open-ended questions offer less potential response cues. Given that this study is one of the first to explore domain generality and specify not only within domains but also across, it would be of value to extend findings and to replicate across test formats, in order to further support that calibration is a domain-general, person-centered trait, whereas resolution is a domain-specific and knowledge-dependent ability.

Trial effects. Indices of monitoring accuracy were explored within a single trial experiment. Some research indicates that there may be a "trial effect," such that confidence judgments are sensitive to test-retest scenarios (Koriat et al, 2002). In particular, Koriat and colleagues (2002) found that whereas calibration decreases with practice, resolution, in contrast, increases. Future studies should build upon the findings of this study and explore domain generality and specificity not only across domains, but also across time points.

Implications.

Theoretical implications. Meta-reasoning involves the monitoring and control of reasoning and problem-solving, such as those captured by the cognitive demands of the current study tasks. In a review by Ackerman and Thompson (2017), a Meta-Reasoning framework synthesizing the existing literature was proposed (see Figure 6). In their model, Ackerman and Thompson (2017) distinguished between object-level processes, as captured by the left column of the model, and meta-level processes, as captured by the middle and right columns of the model. Object-level processes are responsible for basic cognitive activities, such as perceiving, remembering, planning, and classifying, whereas meta-level processes are responsible for governing object-level processes by (1) monitoring their function (i.e., metacognitive monitoring; middle column of model) and (2) allocating resources when required (i.e., metacognitive control; right column of model). Monitoring accuracy is considered to be a measure of metacognitive monitoring (Ackerman & Thompson, 2017; Bjork, Dunlosky, & Kornell, 2013; Nelson & Narens, 1980), thus captured by the middle column. Given this model, a subsequent consideration is how do findings from the current study fit into this proposed model, and what may be some points of convergence and divergence.

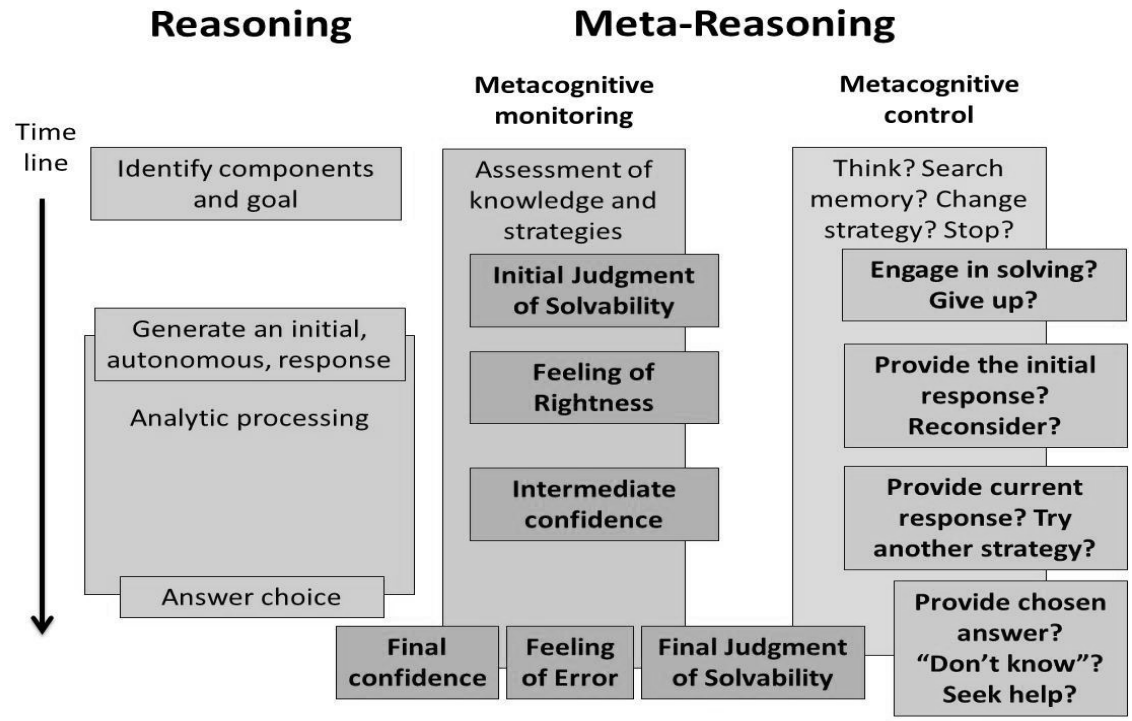


Figure 6. Proposed Meta-Reasoning Framework. Reprinted from Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607-617.

Predictive and postdictive confidence judgments in the current study can somewhat be likened to the initial and final judgments of solvability. Judgments of solvability are defined as the subjective probability that a problem is solvable, with the initial judgment being made from the first impression, and the final judgment being made after working on the problem (Ackerman & Thompson, 2017). A main difference however is that the predictive and postdictive judgments from the current study represent aggregate ratings, rather than item-specific ratings. Aggregate judgments are conceptually different from judgments made immediately before, during, or after an item is complete, as various cognitive factors are employed to pool together and reflect on aggregate performance (Thiede et al., 2005; Veenman et al., 2006). In addition, predictive and postdictive judgments from our study assessed participants' confidence in their overall performance ability within a particular domain (e.g. "How confident are you in your ability to

correctly solve general knowledge questions?") rather than their confidence in being able to solve a specific problem. This assessment of overall ability in a certain knowledge area taps into the cue-utilization view of metacognition, which states that judgments are the result of inferences based on a number of cues derived from beliefs about the self, prior knowledge about the topic area, and previous task-specific experience (Koriat, 1997). That is, when participants judge their ability to correctly solve, for example, general knowledge questions, beliefs about their personal abilities, knowledge, and previous experience with general knowledge problems are accessed and used to form an aggregate confidence judgment (Ackerman & Goldsmith, 2011; Koriat, 1997). Thus, findings from the current study suggest that the middle monitoring column of the Meta-Reasoning Framework also include aggregate judgments. It can be hypothesized that these aggregate impressions also impact subsequent control processes from the right column, such as whether or not to engage in problem-solving, or whether to seek outside support. For example, if overall knowledge in financial investments is judged as poor, an individual might seek out additional resources, such as advice from their financial institution. In contrast, if overall knowledge is judged as sufficient, this behavior may not be engaged in, regardless of whether or not it is objectively needed. Thus, aggregate judgments of knowledge and performance play a role not only in the monitoring of object-level cognitive processes, but also in the subsequent control mechanisms related to resource allocation and behavioral response.

Regarding monitoring accuracy indices, calibration and resolution are best captured by the final confidence in the proposed model, as both represent the subjective probability that the final response to a problem is correct. However, the current study supports that although calibration and resolution are both indicators of monitoring accuracy, these measures are dissociable and represent different aspects of metacognitive monitoring. Feeling of error from

the model does not quite capture relative accuracy, as it reflects the subjective feeling that an error was made (Ackerman & Thompson, 2017), as opposed to the ability to discriminate between correct and incorrect responses. One reason why differentiating monitoring accuracy indices within this model is important can be related to the diverse feedback effect that control strategies can have on each index. For example, in the context of practice and increased exposure to a problem, studies have shown that calibration becomes increasingly impaired, while resolution improves (Koriat, 1997, 2002). This difference in outcome further supports that there may be distinct cognitive processes that underlie calibration and resolution, such that these indices need to be separately considered and interpreted.

A final point regarding Ackerman And Thompson's (2017) Meta-Reasoning Framework would be the inclusion of individual difference variables, as the current study has shown how these may be associated with meta-reasoning. For example, cognitive abilities were not only positively correlated with monitoring accuracy, but also emerged as a predictor of monitoring accuracy, across both calibration and resolution. Perceived workload and effort required may also importantly bear in certain domains, as evidenced by present findings. These individual-level variables can have critical implications for control processes. Individuals can vary in their engagement with effortful and cognitively-demanding tasks and subsequently how much resources they allocate to such tasks (Hart & Staveland, 1988; Hsu et al., 2017). As such, monitoring accuracy is not only impacted by online, cognitive processes, but also by individual and trait-based variables. Historically, the process-based literature and individual differences literature have evolved separately; the current study proposes that in order to gain a better understanding of how cognitive processes and individual differences influence monitoring accuracy and subsequent control and regulation strategies, both lines of research be integrated.

Overall, findings from the current study support the proposed Meta-Reasoning Framework, with some minor differences. Specifically, the model is generally task-specific, capturing monitoring processes on a task and item-level. Theoretical modifications suggested by present study findings include a consideration of aggregate judgments, a clear differentiation between indices of calibration and resolution, and an inclusion of individual difference variables.

Translational implications. Beyond theory, findings from the current study also provide important implications for real-world contexts, such as metacognitive training and intervention, as well as implications directly concerned with education, particularly in the context of math.

Metacognitive training and intervention. Considering that the overconfidence effect was a robust finding, present across domains, it will be important to further support individuals in better monitoring their knowledge. Individuals' subjective confidence directly influences the decisions they make, the effort they invest, and whether to seek additional information. In high stakes domains like financial literacy, miscalibration can have large and long-lasting consequences, such as making poor investments or overestimating one's ability to manage increased debt. Studies have shown that metacognitive training can be beneficial in targeting accuracy, particularly for low-achieving groups (Cardell-Ellawar, 1995; Krugger& Denning, 1999), which are the most vulnerable for poor decision-making and potential long-lasting negative consequences. Metacognitive training typically involves improved strategy use, checking behavior, problem solving, and time and accuracy monitoring (Legg & Locker, 2009).

Furthermore, not only did an overconfidence effect emerge, but this tendency was reflected across different domains. Demonstrating generality in miscalibration across a diverse set of tasks is a novel and central contribution. Framing calibration as a person, rather than task-centred ability has direct implications for future interventions. Psychological and person-related

traits are usually thought of as developing in childhood and remaining somewhat fixed over time, becoming more resistant to intervention as a function of age. Given that metacognitive skills appear between the ages of eight and ten (Veenman et al., 2006), metacognitive training may be most beneficial during middle and late childhood, in order to support the development of emerging monitoring skills. An aim of future research should be to examine the effects of metacognitive training in reducing bias and whether this reduction in bias transfers to other domains. If calibration is domain-general, then it would be expected that metacognitive training can lead to calibration improvements across tasks.

Beyond targeting improved confidence, another training and intervention implication of the current study is that resolution should also be supported. The ability to discriminate between hits and misses is just as important to monitoring accuracy as is the ability to judge overall performance and should also be developed for efficient decision-making. Current findings showed that within the same task, successful calibration was not correlated with good resolution, suggesting that improvements in overconfidence would not necessarily lead to improvements in discrimination. This distinction is of importance, as it is not sufficient to accurately judge, for example, one's ability in financial literacy, but to also be able to distinguish between elements within that same domain that are more or less skilled. For example, an individual may be adept at interest rate calculations, yet less knowledgeable in mortgages. This individual's overall confidence in financial skills may therefore prevent them from seeking external support and advice, despite poor mortgage-related knowledge. Compared to calibration, there is much less research on ways to improve resolution through metacognitive training. One study by Koriat and colleagues (2002) did demonstrate that, in contrast to overconfidence, resolution abilities actually improved with repeated practice. This finding can be best understood within the context

of domain-specificity, such that as individuals accumulate greater content knowledge, performance in that domain also increases. A potential intervention pathway for discrimination would therefore be to increase exposure to domain knowledge and to provide ongoing feedback on performance, thus sensitizing individuals to aspects of content-related knowledge that have been more or less internalized. Thus, in contrast to calibration that can be supported as a generalized trait, findings from the present study suggest that resolution abilities are domain-specific and would require domain-specific interventions in order to help individuals better discriminate between correct and incorrect responses.

Education. The current study purposefully chose to examine mathematical decision-making when selecting which domains to compare to the more established area of general knowledge. Not only do mathematics represent an important academic domain in education, but mathematical skills have direct implications in real-world type problems, such as financial literacy (Chen & Volpe, 1998) and the use of probabilistic thinking in everyday decision-making (Gigerenzer et al, 2007).

Although understanding individual differences in math performance has been investigated in the literature, math has usually been explored in relation to anxiety (Erikson & Heit, 2015; Morsanyi et al., 2019). Exploring math from a metacognitive framework can provide a different narrative of why, beyond differences in actual computational knowledge and abilities, some individuals are more mathematically competent that others. As such, a novel contribution of this study is that it provides an empirical method for exploring how internal criteria, such as confidence judgments and the various meta-level processes described in Ackerman and Thompson's (2017) model, can help individuals monitor and adjust their mathematical performance.

Considering described differences between calibration and resolution, distinct intervention targets would be required to help support overall monitoring of math performance. Due to the domain-generality of overconfidence, increased training in math-related content may not necessarily lead to better performance calibration. Some studies have actually shown that interventions directly targeting financial literacy have led to only small improvements in financial competency, and that such improvements decayed over time (Fernandes, Lynch, & Netemeyer, 2014). Thus, another method for approaching mathematical competence can be based on how individuals engage in and adjust performance, based on their subjective confidence ratings. In the current study, confidence was related to performance, across all tasks. Similarly, Stankov, Morony, and Lee (2012) demonstrated that compared to other self-constructs of self-efficacy, self-concept, and anxiety, confidence emerged as the best non-cognitive predictor of academic achievement. A study by Jacobse and Harskamp (2012) also found that monitoring accuracy explained 16 to 36% of the variance in mathematics achievement. Thus, increased awareness of the accuracy of internal judgments of performance may be vital to supporting math engagement and performance.

In contrast, the domain-specificity of resolution abilities suggests that to better aid learners in discriminating between correct versus incorrect responses, increased computational and financial literacy exposure would be required. That is, increased mathematical knowledge would lead to better discrimination between concepts that are more or less known. Given these differences between calibration and resolution, it is going to be of importance to delineate what are individuals' specific areas of need, when targeting math performance. That is, is the learner biased and overconfident in their computational abilities? Or, are they able to appropriately gage their abilities, yet have difficulty discriminating between what is more or less understood? Or

rather, is it some combination of the two? A better understanding of which metric of metacognition accuracy is compromised can then lead to tailored intervention targets and improved metacognitive training, directly supporting areas of metacognitive monitoring that are in need.

Overall, the general educational implication of this current study is that to promote better correspondence between learner's perceptions of math performance and their actual performance, general metacognitive training targeted at increasing awareness of subjective monitoring states is of importance. Research has generally found that the ability to solve mathematical problems improves with the use of metacognitive training (e.g., Jacobse & Harskamp, 2012), further supporting the need to integrate confidence awareness and monitoring when targeting mathematical performance.

Future directions. Considering the novelty of this research and that the primary goal was to provide empirical evidence that questions of domain-generality and specificity can be experimentally explored, there are many ways in which current findings can be built upon and refined.

First, given the preliminary nature of the current study's findings on metacognitive monitoring across different domains, further replication will be critical to extend our understanding of the involved processing and metacognitive mechanisms and their relative contributions throughout processing stages (Ackerman &Thompson, 2017).

Second, described methodological issues relating to both cross-domain and cross-index comparisons, such as considerations of performance variability and task difficulty, should be

empirically explored in future studies. A better understanding of how task format and exposure can impact indices of monitoring accuracy is also warranted.

Third, to help understand how to best support the emergence of monitoring accuracy abilities, future research grounded in a developmental perspective would be of value. In particular, given that calibration demonstrates domain-generality and can be likened to a trait-based ability, an in-depth understanding of the overconfidence bias in children and whether such a bias improves with early intervention and metacognitive training would be of value.

Fourth, proposed theoretical implications of this study should be further tested and explored. The current study largely focused on the middle column of Ackerman & Thompson's (2017) Meta-Reasoning framework model; a step further would be to empirically test how monitoring accuracy indexes subsequently relate and impact the right column of the model, which represent processes of metacognitive control.

Last, study implications suggest that poor calibration and resolution require different channels of intervention. It would be of importance to empirically test these hypotheses and whether the suggested domain-general versus domain-specific interventions lead to expected improvements in corresponding accuracy indices.

Concluding Remarks

The purpose of the current study was to extend the study of calibration and resolution to diverse every day domains, where people must make implicit judgments about their accuracy before deciding on an action. In addition to general knowledge, financial and probability calculations were specifically selected, given the relevance of these domains for personal financial management (Lusardi & Mitchell, 2014) and the fact that we are presented with probability information in so many facets of our lives (Gigerenzer, Gaissmaier, Kurz-Milcke,

Schwartz, & Woloshin, 2007). Overall, more consistency was found across tasks in predicting calibration than in predicting resolution. The degree of domain generality has practical implications for learning, education, and other applications based on monitoring accuracy, like medical and financial decisions. The findings highlight the critical difference between monitoring indices, such that a pivotal implication of this study is that calibration and resolution represent distinct and separable measures of monitoring accuracy, both of importance for effective monitoring and subsequent effort regulation. To our knowledge, this is the first study to explore resolution across a diverse set of domains and this distinction between both these levels of monitoring accuracy is of great importance. From the perspective of trainability, different interventions may be required for each index and across domains. If resolution indices are more sensitive to domain differences, training of unique aspects of a domain may be more likely to positively impact this index. However, given the generality of the overconfidence index, it is worth considering intervention strategies that might facilitate better calibration across domains.

References

Ackerman, R. (2019). Heuristic Cues for Meta-Reasoning Judgments: Review and Methodology. *Psychological Topics, 28*(1), 1-20.

Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgements during reasoning and memorisation. *Thinking & Reasoning, 23*(4), 376-408.

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18-32

Ackerman, R., Parush, A., Nassar, F., & Shtub, A. (2016). Metacognition and system usability: Incorporating metacognitive research paradigm into usability testing. *Computers in Human Behavior*, *54*, 101-113.

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607-617.

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review*, *19*(6), 1187-1192.

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition, 146*, 377-386.

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*(1), 110-118.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417-444.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*, 938-956.

Cardell-Elawar, M. (1995). Effects of metacognitive instruction on low achievers in mathematics problems. *Teaching and Teacher Education,* 11, 81-95.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270-295.

Chen, P. P. (2002). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*(1), 77-90.

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

Dornic, S., Ekehammar, B., & Laaksonen, T. (1991). Tolerance for mental effort: Self-ratings related to perception, performance and personality. *Personality and Individual Differences, 12*(3), 313-319

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. purchase. *De Economist, 164*(1), 19-39.

Erickson, S., & Heit, E. (2015). Metacognition and confidence: Comparing math to other academic subjects. *Frontiers in Psychology, 6*, 10.

Everson, H. T., Smodlaka, I., & Tobias, S. (1994). Exploring the relationship of test anxiety and metacognition on reading test performance: A cognitive analysis. *Anxiety, Stress & Coping: An International Journal, 7*(1), 85-96.

Fernandes, D., Lynch, J. G., Jr., & Netemeyer, R. G. (2014). Financial literacy, financial education, and downstream financial behaviors. *Management Science, 60*(8), 1861-1883.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19-34.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906-911.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443-451.

Fitzgerald, L. M., Arvaneh, M., & Dockree, P. M. (2017). Domain-specific and domain-general processes underlying metacognitive judgments. *Consciousness and Cognition: An International Journal, 49*, 264-277.

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning, 8*(1), 1-18.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in The Public Interest, 8*, 53-96.

Glaser, R. (1991). The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and Instruction*, *1*(2), 129-144.

Grainger, C., Williams, D. M., & Lind, S. E. (2016). Metacognitive monitoring and control processes in children with autism spectrum disorder: Diminished judgement of confidence accuracy. *Consciousness and Cognition: An International Journal, 42*, 65-74.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of metamemory and memory; handbook of metamemory and memory* (pp. 429-455, Chapter xiv, 487 Pages) Psychology Press, New York, NY.

Hilgert, M. A., Hogarth, J. M., & Beverly, S. G. (2003). Household financial management: The connection between knowledge and behavior. *Federal Reserve Bulletin, 89*(7), 309-322

Hsu, C. F., Eastwood, J. D., & Toplak, M. E. (2017). Differences in Perceived Mental Effort Required and Discomfort during a Working Memory Task between Individuals At-risk And Not At-risk for ADHD. *Frontiers in Psychology*, *8*, 407-415.

Hsu, C. F., Propp, L., Panetta, L., Martin, S., Dentakos, S., Toplak, M. E., & Eastwood, J. D. (2018). Mental effort and discomfort: Testing the peak-end effect during a cognitively demanding task. *PloS One*, *13*(2), e0191479.

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*(1), 25-49.

Jackson, S. A., Kleitman, S., Howie, P., & Stankov, L. (2016). Cognitive abilities, monitoring confidence, and control thresholds explain individual differences in heuristics and biases. *Frontiers in Psychology, 7*, 14.

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

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, *107*(2), 384-396.

Klayman, J., & Soll, J. (1999). Overconfidence: It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247.

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*(1), 92-107.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217-273.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216-247.

Kleitman, S. (2008). *Metacognition in the Rationality Debate. Self-confidence and its Calibration*. Saarbrucken, Germany: VDM Verlag Dr Muller.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*(3), 321-341.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*, 161-173.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed. ed.) Guilford Press, New York, NY.

Komori, M. (2016). Effects of working memory capacity on metacognitive monitoring: A study of group differences using a listening span test. *Frontiers in Psychology, 7*, 9.

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.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition, 36*(2), 416-428.

Koriat, A. (2012a). The relationships between monitoring, regulation and performance. *Learning and Instruction*, *22*(4), 296-298.

Koriat, A. (2012b). The self-consistency model of subjective confidence. *Psychological Review, 119*(1), 80.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107-118.

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy: Insights from the processes underlying judgments of learning in children. In M. Izaute, P. Chambres, & P.-J. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 1–17). New York: Kluwer.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*(2), 147-162.

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept - towards a model of response confidence. *Intelligence, 35*(6), 580-590.

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

Laija-Rodriguez, W., Grites, K., Bouman, D., Pohlman, C., & Goldman, R. L. (2013). Leveraging strengths assessment and intervention model (LeStAIM): A theoretical strength-based assessment framework. *Contemporary School Psychology, 17*(1), 81-91.

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, *35*, 455-463.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences, 19*(3), 355-365.

Legg, A. M., & Locker, L., Jr. (2009). Math performance and its relationship to math anxiety and metacognition. *North American Journal of Psychology, 11*(3), 471-486.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?. *Organizational Behavior and Human Performance*, *20*(2), 159-183.

Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgments. *International Journal of Behavioral Development, 26*(4), 327-333

Lusardi, A., & Mitchell, O. S. (2007). Financial literacy and retirement preparedness: Evidence and implications for financial education. *Business Economics, 42*(1), 35-44.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1), 5-44.

Lusardi, A., Mitchell, O. S., & Curto, V. (2009). *Financial literacy among the young: Evidence and implications for consumer policy*. St. Louis: Federal Reserve Bank of St Louis.

Lusardi A., & Tufano, P. (2015). Debt literacy, financial experiences, and overindebtedness. *Journal of Pension Economics & Finance, 14*(4), 332-368.

Macher, D., Papousek, I., Ruggeri, K., & Paechter, M. (2015). Statistics anxiety and performance: Blessings in disguise. *Frontiers in Psychology, 6*, 4.

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.

Malmendier, U., & Tate, G. (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. *Journal of Financial Economics*, *89*(1), 20-43.

Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 509-527.

Metcalfe, J., & Shimamura, A. P. (1995). Metacognition: Knowing about knowing. *Information*

Processing and Management, 31(2), 261-262.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 502-506.

Miron-Shatz, T., & Hanoch, Y., Doniger, G., Omer, B., & Ozanne, E. (2014). Subjective but not objective numeracy influences willingness to pay for BRCA1/2 genetic testing. *Judgment and Decision Making. 9.* 152-158.

Morsanyi, K., Cheallaigh, N. N., & Ackerman, R. (2019). Mathematics Anxiety and Metacognitive Processes: Proposal for a New Line of Inquiry. *Psychological Topics, 28*(1), 147-169.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109-133.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 338-368.

Pallier, G., Wilkinson, R., Danthir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology, 129*(3), 257-299.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual differences approach. *Journal of Behavioral Decision Making, 18*, 1-27.

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgments in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *18*(2), 157-168.

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology, 108*(5), 802.

Pressley, M. (2002). Metacognition and self-regulated comprehension. In A. E. Farstrup & S. J. Samuels, (Eds.), What research has to say about reading instruction (3rd ed., pp. 291 -309). Newark, DE: International Reading Association.

Pugalee, D. K. (2001). Writing, mathematics, and metacognition: Looking for connections through students' work in mathematical problem solving. *School Science and Mathematics, 101*(5), 236-245.

Reynolds, W. M., Ramírez, M. P., Magriña, A., & Allen, J. E. (1980). Initial development and validation of the academic self-concept scale. *Educational and Psychological Measurement, 40*(4), 1013-1016.

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS One, 9*(7).

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*(2), 193-218.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521-562.

Schrand, C. M., & Zechman, S. L. (2012). Executive overconfidence and the slippery slope to financial misreporting. *Journal of Accounting and Economics*, *53*(1-2), 311-329.

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.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist?. *Journal of Educational Psychology, 87(3), 433-444.*

Scott, B. M., & Berman, A. F. (2013). Examining the domain-specificity of metacognition using academic domains and task-specific individual differences. *Australian Journal of Educational & Developmental Psychology, 13*, 28-43.

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). Shipley-2. *Los Angeles, CA: Western Psychological Services*.

Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction, 51*, 61-73.

Soderstrom, N. C., Yue, C. L., & Bjork, E. L. (2015). Metamemory and education. In *The Oxford Handbook of Metamemory*.

Stankov, L., Kleitman, S., & Jackson, S. A. (2014). Measures of the trait of confidence. In G. J. Boyle, H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs*. Academic Press (pp. 158-189).

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? Learning and Individual Differences, 22(6), 747-758.

Stanovich, K. (2011). *Rationality and the reflective mind*. New York, NY: Oxford University Press.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161-188.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The Rationality Quotient: Toward a test of rational thinking*. MIT Press.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66-73.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1267-1280.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, *47*(4), 331-362.

Thiede, KW, Muller, ML, & Dunlosky, J. (2015) Methodology for investigating human metamemory: Problems and pitfalls: The Oxford handbook of metamemory.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*(2), 264-273.

Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In J. ST. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond. Oxford University Press, Oxford*.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language*, *28*(2), 127-154.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods*, *45*(4), 1115-1143.

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences, 15*(2), 159-176.

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3-14.

Veenman, M. V. J., & Verheij, J. (2003). Technical students' metacognitive skills: Relating general vs. specific metacognitive skills to study success. *Learning and Individual Differences, 13*(3), 259-272.

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin & Review*, *4*(3), 387-392.

Yates, J. F. (1990). *Judgment and decision making* Prentice-Hall, Inc, Englewood Cliffs, NJ.

Yates, J. F., Lee, J., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and "reality". *Organizational Behavior and Human Decision Processes, 70*, 87-94.

Zabrucky, K. M. (2010). Knowing what we know and do not know: Educational and real-world implications. *Procedia-Social and Behavioral Sciences*, *2*(2), 1266-1269.

Zacharakis, A. L., & Shepherd, D. A. (2001). The nature of information and overconfidence on venture capitalists' decision making. Journal of Business Venturing, 16(4), 311-332.

APPENDIX A:
DEMOGRAPHICS FORM

Demographic Information

Age: _____

Gender:
___ Male
___ Female
___ Other

_Year in University:
___ 1st year undergrad
___ 2nd year undergrad
___ 3rd year undergrad
___ 4th year undergrad
___ 5th year undergrad
___ Post-BA Continuing
___ Other

Please indicate your ethnicity (Check one)
___White/European
___ Black
___Asian
___Aboriginal
___South-Asian
___Arab
___ Latino-Hispanic
___ Other (please specify)_____

Is English your first language? ___Yes ___ No
If No, how long have you been speaking English?
_____ Years

Financially, do you consider your family to be:
___Well below average income
___Below average income
___Average income
___Above average income
___Well above average income

Current academic average grade:
___Below 49%
___50 – 59%
___60 – 69%
___70 – 79%
___80 – 100%

Overall, do you think that you have difficulties in one or more of the following areas: emotions, behavior or being able to get along with other people?
____No
____Yes – minor difficulties
____Yes – definite difficulties
____Yes – severe difficulties

Overall, do you think that you have difficulties in learning or academics?
____No
____Yes – minor difficulties
____Yes – definite difficulties
____Yes – severe difficulties

	Mother	**Father**
Less than 7th Grade	☐	☐
Junior high / Middle school (9th grade)	☐	☐
Partial high school (10th or 11th grade)	☐	☐
High school graduate	☐	☐
Partial college/university (at least one year)	☐	☐
College/university education	☐	☐
Graduate/professional degree	☐	☐

APPENDIX B:
EXPERIMENTAL TASKS PILOT DATA

General Knowledge Test

Item	Correct Response	Success Rate (%)
1. What is the longest river in South America?	Amazon	63.16
2. For which country is the yen the monetary unit?	Japan	57.89
3. What is the last name of the man who first studied genetic inheritance in plants?	Mendel	52.63
4. What is the proper name for a badminton bird?	Shuttlecock	36.84
5. What is the last name of the author who wrote "Oliver Twist"?	Dickens	63.16
6. What is the last name of the man who invented the phonograph?	Edison	33.33
7. What is the name of an inability to sleep?	Insomnia	94.74
8. What is the name of the lizard that changes its colour to match the surroundings?	Chameleon	100
9. What is the name of the largest ocean on earth?	Pacific	36.84
10. What is the capital of Australia?	Canberra	26.32
11. What animal runs the fastest?	Cheetah	100
12. What is the term for hitting a volleyball down hard into the opponents court?	Spike	89.47
13. What is the name of the brightest star in the sky excluding the sun?	Sirius	52.63
14. What is the name of a dried grape?	Raisin	84.21
15. What is the name of the largest desert on earth?	Antarctica	5.26
16. What is the name of the mountain range that separates Asia from Europe?	Ural	15.79
17. What is the name for a medical doctor who specializes in diseases of the skin?	Dermatologist	47.37
18. What is the unit of sound intensity?	Decibel	63.16
19. What is the name of deer meat?	Venison	57.89
20. What is the name of the organ that produces insulin?	Pancreas	68.42
21. What is the name of the automobile instrument that measures mileage?	Odometer	63.16
22. What is the name of the bird that cannot fly and is the largest bird on earth?	Ostrich	94.74
23. What is the name for a cyclone that occurs over land?	Tornado	78.95
24. What is the largest planet in the solar system?	Jupiter	63.16

Note. $N = 18$; Total success rate = 60.38%

Financial Calculation Test

Item	Correct Response	Success Rate (%)
1. Imagine you are planning a trip to the USA, and the current exchange rate is $1 Canadian dollar = $0.75 US dollars. That is, for every Canadian dollar, you will receive 75 cents in US funds. If you have $50 Canadian to spend, how much will you get in US dollars?	$37.50 US	89.47
2. Janice has a job where she earns $2000 per month. She spends $900 for rent and $150 for groceries each month. She also spends $250 per month on transportation. If she budgets $100 each month for clothing, $200 for restaurants and $250 for everything else, how long will it take her to save $750?	5 months	89.47
3. Marlon decided to buy a new couch for his apartment. He bought a new couch for $800 and used his credit to pay for this purchase. His credit card charges a 26% annual interest rate and Marlon expects that after his bill arrives, he won't be able to pay for any of his balance and he will need an additional 12 months to pay for his new couch. How much will his monthly payment be for his new couch?	$84.00	63.16
4. Which of the following is a correct calculation of simple interest on $1000 at the end of one year?	4% = $40 or $1040	73.68
5. Imagine you are planning a trip to Mexico where Mexican Pesos are the currency, and the current exchange rate is $1 Canadian dollar = 16 Mexican pesos. That is, for every Canadian dollar, you will receive 16 Mexican pesos. If you have $25 Canadian to spend, how much will you get in Mexican pesos?	400 Mexican pesos	100
6. If Frederick deposits 25% of $130 into a savings account, what is the amount of his deposit?	$32.50	84.21
7. Stefan bought a used car. The car cost $12000, and he put a $7000 deposit on it from his savings. He put the remainder on his credit card. His credit card charges a 21% annual interest rate and Stefan expects that after his bill arrives, he won't be able to pay for any of his balance and he will need an additional 12 months to pay for his car. How much interest total interest will he pay for his car?	$1050	78.95
8. If Bilal gets a simple interest rate of 2.6% on his savings account, how much interest will he receive on $5000 that he has had in the bank for a year?	$130	89.47
9. Imagine you are planning a trip to Switzerland where Euros are the currency, and the current exchange rate is $1 Canadian	140 Euros	100

dollar = $0.70 Euros. That is, for every Canadian dollar, you will receive 70 cents in Euros. If you have $200 Canadian to spend, how much will you get in US dollars?		
10. Drummond is trying to figure out if he will need to get a part-time job over the school year. He has gotten a student loan for $15000 this year for 12 months. His rent is $850 per month, his grocery bill is $200 per month, his cell phone bill is $55 per month, and his transit pass is $120 per month. He expects to need $250 per month for additional expenses. Which of the following situations will Drummond be in?	Drummond will have to earn $225 each month in a part-time job	68.42
11. Sonja unexpected had to get a new muffler and brakes for her car. The total cost was $1500. She used her credit to pay for this purchase. Her credit card charges a 19% annual interest rate and Sonja expects that after her bill arrives, she won't be able to pay for any of her balance and she will need an additional 12 months to pay for this expense. How much will her monthly payment be?	$148.75	63.16
12. Bank A provides 0.5% interest for the first year in a savings account and 3.9% in the second year. Bank B provides an interest rate of 1.9% for savings accounts. Mohamed has $5000 and he decides to put his money into Bank B for the first year and then transfer his money into Bank A for the second year. If he uses this strategy, how much money will he have after 2 years?	$5220.98	72.22
13. Imagine you are planning a trip to Brazil, and the current exchange rate is $1 Canadian dollar = 2.37 Brazilian Reals. That is, for every Canadian dollar, you will receive 2.37 Brazilian Reals. If you have $1000 Canadian to spend, how much will you get in Brazilian Reals?	2370 Brazilian Reals	89.47
14. If Lena withdraws 35% of the $4000 that she has in her savings account, how much money did she withdraw?	$1400	94.74
15. Raman decided to take a business training year abroad in India. The cost of this training opportunity was $9500. She used her credit card to pay for this opportunity. Her credit card charges a 20% annual interest rate and Raman expects that after her bill arrives, she won't be able to pay for any of her balance and she will need an additional 12 months to pay for her training opportunity. How much total interest will she pay for this opportunity?	$1900	94.74
16. Denoja had $5000 saved up in her bank account. After one year, she got 2.0% interest on this money. She then took out $1000 to buy some books. She saved the rest of her money in this account for another year at the same interest rate. How much money did she have after 2 years?	$4182	89.47

17. Imagine you are planning a trip to Costa Rica where Costa Rican Colons are the currency, and the current exchange rate is $1 Canadian dollar = 140 Costa Rican Colons. That is, for every Canadian dollar, you will receive 140 Costa Rican Colons. If you have $20 Canadian to spend, how much will you get in Costa Rican Colons?	2800 Costa Rican Colons	84.21
18. For his first year, Noel's parents provided him with $25000 to cover his university expenses. After paying $7200 for his tuition, $12000 for his residence, how much will Noel have left over to spend each month if he must budget for 8 months?	$725	89.47
19. Dana used her credit card to pay $875 for her books. Her credit card charges a 22% annual interest rate and Dana expects that after her bill arrives, she won't be able to pay for any of her balance and she will need an additional 12 months to pay for her books. How much total interest will she pay for her books?	$192.50	89.47
20. Dina inherited $2500 from her grandmother. Dina wants to save this money for university, and she has put it into savings account for two years. She will receive an annual interest rate of 3.5%. How much will her money be worth in two years?	$2678.06	73.68
21. Imagine you and your partner are planning a trip to Vietnam, and the current exchange rate is $1 Canadian dollar = $1750 Vietnamese Dong. That is, for every Canadian dollar, you will receive 1750 Vietnamese Dong. If you have $200 Canadian to spend, how much will you get in Vietnamese Dongs	350000 Vietnamese Dong	78.95
22. Maika has two jobs. She earns $400 a week at her first job and $600 a month at her second job. Each month, she spends $1000 for rent, $200 for groceries, $125 on transportation, and $75 on her cell phone. If she budgets $200 each month for spending and $125 for clothing, how long will it take her to save $950?	2 months	73.68
23. Grant decided to get a new laptop for school. He bought a new MacBook for $1600 and used his credit to pay for this purchase. His credit card charges a 22% annual interest rate and Grant expects that after his bill arrives, he won't be able to pay for any of his balance and he will need an additional 12 months to pay for his new laptop. How much will his monthly payment be for his new laptop?	$162.67	63.16
24. Which of the following is a correct calculation of simple interest on $2000 at the end of one year?	3% = $30 or $2030	78.95

Note. $N = 18$; Total success rate = 82.18%

Probability Calculation Test

Item	Correct Response	Success Rate (%)
1. Which of the following represents the biggest risk for getting a disease?	1 in 10	100
2. If you flipped a fair coin three times, what is the probability that it will land "heads" on all three flips?	12.5%	72.22
3. Treatments W, X, Y and Z have the same effects. […] Which treatment will most likely cause headaches?	Treatment Z	88.89
4. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). Imagine that the die is rolled 100 times. Out of 100 rolls, how many times do you think the die would come up as an even number (2, 4, or 6)?	50 times	83.33
5. If there are 5 red, 3 green and 2 yellow marbles in a paper bag, what are the chances of choosing a green marble?	30%	94.44
6. Which player is most likely to score a goal in hockey?	Player C who scores 31% of the shots on net	77.78
7. A chance of miscarriage is approximately 15% during the first 20 weeks of pregnancy. Drug use can triple the incidence of miscarriage during this time. If a pregnant woman uses drugs during her first 20 weeks, what chance does she have of having a miscarriage?	45%*	77.78
8. The chances that a sewing machine will break the thread in a clothing factory is 12%. Out of the 6000 clothing items produced in one day, how many items will be affected by the broken thread?	720 clothing items	83.33
9. Which of the following represents the best chance for winning the lottery?	4/10	66.67
10. There are currently four main products available to stimulate hair growth products for middle-aged men. Sam wants to select the product that will give him the best outcome. Which product should he select?	Product IV has a 4 out of 20 chance of a positive outcome in 6 weeks.	83.33
11. If you flipped two fair coins at the same time, what is the probability that they both come up tails?	33%	77.78
12. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). Imagine that the die is rolled. What is the probability of getting a 3 or a 5 on your throw?	33%	88.89

13. Which of the following represents the lowest risk for getting a disease?	1 in 100	94.44
14. A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?	5/7	61.11
15. If 4 people out of 10000 ticket holders wins a prize, what is the probability of winning a prize?	0.0004%	55.56
16. Which of the following represents the biggest risk for getting a disease?	1 in 5	94.44
17. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). Imagine that two dice are rolled. What is the likelihood that the dice will both land on a one?	1/36	55.56
18. Imagine that you a fair coin three times. What is the probability that it will come up "heads" on the fourth toss	50%	77.78
19. Which of the following represents the lowest risk for getting a disease?	431 in 10000	72.22
20. The risk of a heart attack is about 9 in 100 people. Taking aspirin can reduce the risk of a heart attack by two-thirds. What are the chances of getting a heart attack if you take aspirin?	3%	66.67
21. Your family tells you that there is between a 1/100 to a 1/5 chance that your prescription will cause hives. What is the probability range that you will get hives from your prescription?	1% to 20%	61.11
22. There are four new stain removers for rugs available on the market. Nelson wants to select the product that will give him the best outcome on his stained rugs. Which product should he select?	Product I has a 4 out of 25 chance of removing the stain	72.22
23. If 386 people out of every 100000 get a disease, what is the probability of getting this disease?	0.00386%	38.89
24. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). What is the probability of getting a sum 9 from two throws of a dice?	1/9	61.11

Note. $N = 18$; Total success rate = 71.06%

APPENDIX C:
EXPERIMENTAL TASKS STUDY DATA

General Knowledge Test

Item	Correct Response	Success Rate (%)	Mean Confidence
1. What is the longest river in South America?	Amazon	54.4	61%
2. For which country is the yen the monetary unit?	Japan	37.5	71%
3. What is the last name of the man who first studied genetic inheritance in plants?	Mendel	59.6	71%
4. What is the proper name for a badminton bird?	Shuttlecock	34.6	81%
5. What is the last name of the author who wrote "Oliver Twist"?	Dickens	46.3	54%
6. What is the last name of the man who invented the phonograph?	Edison	25.7	57%
7. What is the name of an inability to sleep?[a]	Insomnia	91.2	92%
8. What is the name of the lizard that changes its colour to match the surroundings? [a]	Chameleon	96.3	94%
9. What is the name of the largest ocean on earth?	Pacific	58.8	76%
10. What is the capital of Australia? [a]	Canberra	9.6	77%
11. What animal runs the fastest? [a]	Cheetah	94.9	92%
12. What is the term for hitting a volleyball down hard into the opponents court? [a]	Spike	93.4	88%
13. What is the name of the brightest star in the sky excluding the sun? **	Sirius	19.4	68%
14. What is the name of a dried grape? **	Raisin	87.8	91%
15. What is the name of the largest desert on earth? [a]	Antarctica	3.7	80%
16. What is the name of the mountain range that separates Asia from Europe? [a]	Ural	6.6	63%
17. What is the name for a medical doctor who specializes in diseases of the skin? [a]	Dermatologist	94.9	94%
18. What is the unit of sound intensity?	Decibel	54.4	79%
19. What is the name of deer meat?	Venison	47.1	61%
20. What is the name of the organ that produces insulin?	Pancreas	68.4	73%
21. What is the name of the automobile instrument that measures mileage?	Odometer	40.4	76%
22. What is the name of the bird that cannot fly and is the largest bird on earth?	Ostrich	86.0	88%
23. What is the name for a cyclone that occurs over land?	Tornado	75.0	79%
24. What is the largest planet in the solar system?	Jupiter	51.5	74%

Note. $N = 136$; ** $N = 98$; Total success rate = 53.19%
[a]Item removed due to overall success of less than 10% or greater than 90%.

Financial Calculation Test

Item	Correct Response	Success Rate (%)	Mean Confidence
1. Imagine you are planning a trip to the USA, and the current exchange rate is $1 Canadian dollar = $0.75 US dollars. That is, for every Canadian dollar, you will receive 75 cents in US funds. If you have $50 Canadian to spend, how much will you get in US dollars? [a]	$37.50 US	91.2	92%
2. Janice has a job where she earns $2000 per month. She spends $900 for rent and $150 for groceries each month. She also spends $250 per month on transportation. If she budgets $100 each month for clothing, $200 for restaurants and $250 for everything else, how long will it take her to save $750?	5 months	72.1	86%
3. Marlon decided to buy a new couch for his apartment. He bought a new couch for $800 and used his credit to pay for this purchase. His credit card charges a 26% annual interest rate and Marlon expects that after his bill arrives, he won't be able to pay for any of his balance and he will need an additional 12 months to pay for his new couch. How much will his monthly payment be for his new couch?	$84.00	39.0	76%
4. Which of the following is a correct calculation of simple interest on $1000 at the end of one year?	4% = $40 or $1040	54.4	59%
5. Imagine you are planning a trip to Mexico where Mexican Pesos are the currency, and the current exchange rate is $1 Canadian dollar = 16 Mexican pesos. That is, for every Canadian dollar, you will receive 16 Mexican pesos. If you have $25 Canadian to spend, how much will you get in Mexican pesos? [a]	400 Mexican pesos	94.9	91%
6. If Frederick deposits 25% of $130 into a savings account, what is the amount of his deposit?	$32.50	85.3	89%
7. Stefan bought a used car. The car cost $12000, and he put a $7000 deposit on it from his savings. He put the remainder on his credit card. His credit card charges a 21% annual interest rate and Stefan expects that after his bill arrives, he won't be able to pay for any of his balance and he will need an additional 12 months to pay for his car. How much interest total interest will he pay for his car?	$1050	70.6	72%

8. If Bilal gets a simple interest rate of 2.6% on his savings account, how much interest will he receive on $5000 that he has had in the bank for a year?	$130	64.7	73%
9. Imagine you are planning a trip to Switzerland where Euros are the currency, and the current exchange rate is $1 Canadian dollar = $0.70 Euros. That is, for every Canadian dollar, you will receive 70 cents in Euros. If you have $200 Canadian to spend, how much will you get in US dollars? [a]	140 Euros	91.9	89%
10. Drummond is trying to figure out if he will need to get a part-time job over the school year. He has gotten a student loan for $15000 this year for 12 months. His rent is $850 per month, his grocery bill is $200 per month, his cell phone bill is $55 per month, and his transit pass is $120 per month. He expects to need $250 per month for additional expenses. Which of the following situations will Drummond be in?	Drummond will have to earn $225 each month in a part-time job	65.4	73%
11. Sonja unexpected had to get a new muffler and brakes for her car. The total cost was $1500. She used her credit to pay for this purchase. Her credit card charges a 19% annual interest rate and Sonja expects that after her bill arrives, she won't be able to pay for any of her balance and she will need an additional 12 months to pay for this expense. How much will her monthly payment be?	$148.75	51.5	72%
12. Bank A provides 0.5% interest for the first year in a savings account and 3.9% in the second year. Bank B provides an interest rate of 1.9% for savings accounts. Mohamed has $5000 and he decides to put his money into Bank B for the first year and then transfer his money into Bank A for the second year. If he uses this strategy, how much money will he have after 2 years?	$5220.98	50.0	55%
13. Imagine you are planning a trip to Brazil, and the current exchange rate is $1 Canadian dollar = 2.37 Brazilian Reals. That is, for every Canadian dollar, you will receive 2.37 Brazilian Reals. If you have $1000 Canadian to spend, how much will you get in Brazilian Reals? [a]	2370 Brazilian Reals	91.2	87%
14. If Lena withdraws 35% of the $4000 that she has in her savings account, how much money did she withdraw? [a]	$1400	93.4	86%
15. Raman decided to take a business training year abroad in India. The cost of this training opportunity was $9500. She used her credit card to pay for this	$1900	77.2	73%

opportunity. Her credit card charges a 20% annual interest rate and Raman expects that after her bill arrives, she won't be able to pay for any of her balance and she will need an additional 12 months to pay for her training opportunity. How much total interest will she pay for this opportunity?			
16. Denoja had $5000 saved up in her bank account. After one year, she got 2.0% interest on this money. She then took out $1000 to buy some books. She saved the rest of her money in this account for another year at the same interest rate. How much money did she have after 2 years?	$4182	58.8	66%
17. Imagine you are planning a trip to Costa Rica where Costa Rican Colons are the currency, and the current exchange rate is $1 Canadian dollar = 140 Costa Rican Colons. That is, for every Canadian dollar, you will receive 140 Costa Rican Colons. If you have $20 Canadian to spend, how much will you get in Costa Rican Colons? [a]	2800 Costa Rican Colons	91.2	89%
18. For his first year, Noel's parents provided him with $25000 to cover his university expenses. After paying $7200 for his tuition, $12000 for his residence, how much will Noel have left over to spend each month if he must budget for 8 months?	$725	84.6	85%
19. Dana used her credit card to pay $875 for her books. Her credit card charges a 22% annual interest rate and Dana expects that after her bill arrives, she won't be able to pay for any of her balance and she will need an additional 12 months to pay for her books. How much total interest will she pay for her books?	$192.50	68.4	77%
20. Dina inherited $2500 from her grandmother. Dina wants to save this money for university, and she has put it into savings account for two years. She will receive an annual interest rate of 3.5%. How much will her money be worth in two years?	$2678.06	51.5	70%
21. Imagine you and your partner are planning a trip to Vietnam, and the current exchange rate is $1 Canadian dollar = $1750 Vietnamese Dong. That is, for every Canadian dollar, you will receive 1750 Vietnamese Dong. If you have $200 Canadian to spend, how much will you get in Vietnamese Dongs	350000 Vietnamese Dong	85.3	88%
22. Maika has two jobs. She earns $400 a week at her first job and $600 a month at her second job. Each month, she spends $1000 for rent, $200 for groceries,	2 months	69.9	78%

$125 on transportation, and $75 on her cell phone. If she budgets $200 each month for spending and $125 for clothing, how long will it take her to save $950?			
23. Grant decided to get a new laptop for school. He bought a new MacBook for $1600 and used his credit to pay for this purchase. His credit card charges a 22% annual interest rate and Grant expects that after his bill arrives, he won't be able to pay for any of his balance and he will need an additional 12 months to pay for his new laptop. How much will his monthly payment be for his new laptop?	$162.67	52.9	75%
24. Which of the following is a correct calculation of simple interest on $2000 at the end of one year?	3% = $30 or $2030	61.8	63%

Note. $N = 136$; Total success rate = 62.99%.

[a] Item removed due to overall success rate of less than 10% or greater than 90%.

Probability Calculation Test

Item	Correct Response	Success Rate (%)	Mean Confidence
1. Which of the following represents the biggest risk for getting a disease? [a]	1 in 10	91.2	90%
2. If you flipped a fair coin three times, what is the probability that it will land "heads" on all three flips?	12.5%	43.4	68%
3. Treatments W, X, Y and Z have the same effects. […] Which treatment will most likely cause headaches?	Treatment Z	77.9	86%
4. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). Imagine that the die is rolled 100 times. Out of 100 rolls, how many times do you think the die would come up as an even number (2, 4, or 6)?	50 times	87.5	76%
5. If there are 5 red, 3 green and 2 yellow marbles in a paper bag, what are the chances of choosing a green marble? [a]	30%	90.4	83%
6. Which player is most likely to score a goal in hockey?	Player C who scores 31% of the shots on net	69.9	79%
7. A chance of miscarriage is approximately 15% during the first 20 weeks of pregnancy. Drug use can triple the incidence of miscarriage during this time. If a pregnant woman uses drugs during her first 20 weeks, what chance does she have of having a miscarriage?	45%	73.5	83%
8. The chances that a sewing machine will break the thread in a clothing factory is 12%. Out of the 6000 clothing items produced in one day, how many items will be affected by the broken thread?	720 clothing items	69.9	78%
9 Which of the following represents the best chance for winning the lottery?	4/10	50.7	76%
10. There are currently four main products available to stimulate hair growth products for middle-aged men. Sam wants to select the product that will give him the best outcome. Which product should he select?	Product IV has a 4 out of 20 chance of a positive outcome in 6 weeks.	67.6	78%

11. If you flipped two fair coins at the same time, what is the probability that they both come up tails?	33%	70.6	64%
12. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). Imagine that the die is rolled. What is the probability of getting a 3 or a 5 on your throw?	33%	62.5	71%
13. Which of the following represents the lowest risk for getting a disease? [a]	1 in 100	90.4	91%
14. A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?	5/7	48.5	72%
15. If 4 people out of 10000 ticket holders wins a prize, what is the probability of winning a prize?	0.0004%	55.9	80%
16. Which of the following represents the biggest risk for getting a disease?	1 in 5	72.8	81%
17. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). Imagine that two dice are rolled. What is the likelihood that the dice will both land on a one?	1/36	32.4	74%
18. Imagine that you a fair coin three times. What is the probability that it will come up "heads" on the fourth toss	50%	47.8	73%
19. Which of the following represents the lowest risk for getting a disease?	431 in 10000	66.9	79%
20. The risk of a heart attack is about 9 in 100 people. Taking aspirin can reduce the risk of a heart attack by two-thirds. What are the chances of getting a heart attack if you take aspirin?	3%	47.8	71%
21. Your family tells you that there is between a 1/100 to a 1/5 chance that your prescription will cause hives. What is the probability range that you will get hives from your prescription?	1% to 20%	41.9	79%
22. There are four new stain removers for rugs available on the market. Nelson wants to select the product that will give him the best outcome on his stained rugs. Which product should he select?	Product I has a 4 out of 25 chance of removing the stain	65.4	76%
23. If 386 people out of every 100000 get a disease, what is the probability of getting this disease?	0.00386%	48.5	81%

24. A six-sided die has an odd number on three sides (1, 3, 5) and an even number on three sides (2, 4, 6). What is the probability of getting a sum 9 from two throws of a dice?	1/9	35.3	57%

Note. $N = 136$; Total success rate = 58.89%.

[a] Item removed due to overall success rate of less than 10% or greater than 90%.

.

APPENDIX D:
WORKING MEMORY TASK

Instructions:

You will see a sentence on the screen and hear it being said aloud. Your job is to read the sentence out loud along with me, and then as soon as you have finished reading the sentence, decide if the sentence is True or False by checking off either True or False on your sheet of paper that you have in front of you.

After we read some sentences, you will see a screen that says, "What was the last word in each sentence that you read from set #1, 2, 3 etc." When you see this screen, you will be prompted to write down the last word of each sentence that we read from that specific set of words. Don't worry about spelling!

Please put your pencils down as soon as you have finished writing down the words.

Let's give it a try…

Practice Question

A) Jackets have a **zipper**.
TRUE FALSE
B) Shoes go on your **hand**.
TRUE FALSE
What was the last word in each sentence you read? Please put pencils down after writing down the words.

_____ _____

Let's do some more.

Set #1
1) The sun rises in the morning.
TRUE FALSE
2) Trees lose their leaves in spring.
TRUE FALSE
What was the last word of each sentence you read for Set #1?

_____ _____

Set #2
1) A race car is fast.
TRUE FALSE

2) Peas are vegetables.
TRUE FALSE

What was the last word of each sentence you read for Set #2?

_____ _____

Set #3

1) Dogs have six legs.
TRUE FALSE
2) Giraffes are tall.
TRUE FALSE

What was the last word of each sentence you read for Set #3?

_____ _____

Set #4

1) Cars have four wheels.
TRUE FALSE
2) Cows eat meat.
TRUE FALSE
3) A red traffic light means "STOP".
TRUE FALSE

What was the last word of each sentence you read for Set #4?

_____ _____ _____

Set #5

1) Hens lay eggs.
TRUE FALSE
2) Elephants have purple spots.
TRUE FALSE

3) Stars are in the sky.
TRUE FALSE

What was the last word of each sentence you read for Set #5?

_____ _____ _____

Set #6

1) Horses sleep in barns.
TRUE FALSE
2) Boiling water is hot.
TRUE FALSE
3) Strawberries are a fruit.

TRUE FALSE

What was the last word of each sentence you read for Set #6?

_____ _____ _____

Set #7

1) We get milk from cows.
TRUE FALSE
2) Plants need water to grow.
TRUE FALSE
3) It is warm in Winter.
TRUE FALSE
4) Carrots are orange.
TRUE FALSE

What was the last word of each sentence you read for Set #7?

_____ _____ _____
_____

Set #8

1) Birds have wings.
TRUE FALSE
2) Whales live in the ocean.
TRUE FALSE
3) An apple is a fruit.
TRUE FALSE
4) Fish swim in the sky.
TRUE FALSE

What was the last word of each sentence you read for Set #8?

_____ _____ _____
_____

Set #9

1) A soccer ball is round.
TRUE FALSE
2) We sleep at night.
TRUE FALSE
3) Bees make honey.
TRUE FALSE
4) A feather is heavy.
TRUE FALSE

What was the last word of each sentence you read for Set #9?

_____ _____ _____
_____

Set #10

1) Birds fly south for the Winter.
TRUE FALSE
2) The earth travels around the sun.
TRUE FALSE
3) Purple is a colour.
TRUE FALSE
4) A car is a vegetable.
TRUE FALSE
5) Tadpoles become frogs.
TRUE FALSE

What was the last word of each sentence you read for Set #10?

_____ _____ _____
_____ _____

Set #11

1) Grass is green.
TRUE FALSE
2) Monkeys eat bananas.
TRUE FALSE
3) Pizza is a plant
TRUE FALSE
4) Ice is hot.
TRUE FALSE
5) Basketball is a sport.
TRUE FALSE

What was the last word of each sentence you read for Set #11?

_____ _____ _____
_____ _____

Set #12

1) Ants are insects.
TRUE FALSE
2) Lions live on farms.
TRUE FALSE
3) Dogs can bark.
TRUE FALSE

4) Spiders have two legs.
TRUE FALSE
5) A beach has sand.
TRUE FALSE
What was the last word of each sentence you read for Set #12?

_____ _____ _____
_____ _____