Please cite as:

Dentakos, S., Saoud, W., Ackerman, R., & **Toplak, M. E.** (in press). Does Domain Matter? Monitoring Accuracy across Domains. *Metacognition and Learning*.

Does Domain Matter? Monitoring Accuracy across Domains

Stella Dentakos*

LaMarsh Centre for Child and Youth Research

Department of Psychology

York University

Wafa Saoud*

LaMarsh Centre for Child and Youth Research

Department of Psychology

York University

Rakefet Ackerman

Faculty of Industrial Engineering & Management

Technion—Israel Institute of Technology

Maggie E. Toplak

LaMarsh Centre for Child and Youth Research

Department of Psychology

York University

*These two authors made equal contributions to the submission

Address of corresponding author:

Maggie E. Toplak
126 BSB, Department of Psychology
York University
4700 Keele St.
Toronto, Ontario, Canada M3J 1P3
Email: mtoplak@yorku.ca
Phone: 416-736-2100, ext. 33710; Fax: 416-736-5814

Running head: MONITORING ACCURACY ACROSS DOMAINS

Running head: MONITORING ACCURACY ACROSS DOMAINS

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. In this study, we examined whether monitoring accuracy, both calibration (resistance to overconfidence) and resolution (discrimination) indices, are stable within individuals across tasks that represent highly diverse domains. We examined the well-established domain of general knowledge and three understudied applied domains of financial calculation, probability calculation, and the social skill of emotion recognition. In addition, we examined 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 performance, task difficulty, and effort required) as well as the classic postdictive itemized confidence ratings. We found that resistance to overconfidence (calibration) was significantly positively correlated across tasks, reflecting a confidence trait, but not resolution. We also found that cognitive abilities were more consistently predictive of calibration than of resolution. Aggregated judgments and postdictive confidence were significant predictors of both calibration and resolution, but associations were task specific. Emotion recognition displayed the most unique profile of findings relative to other tasks. We conclude that when considering a wide range of domains, calibration displays domain generality, but resolution may display specificity across tasks.

Keywords: calibration, overconfidence, resolution, monitoring accuracy, individual differences

The discrepancy between knowledge and metacognitive monitoring of one's knowledge has been examined extensively in both the metacognition field (Kleitman & Stankov, 2007; Koriat, 2008, 2012a; Stankov, Kleitman, & Jackson, 2014) and in the field of judgment and decision making (Bruine de Bruin, Parker, & Fischhoff, 2007; Parker & Fischhoff, 2005; Stanovich, West, & Toplak, 2016; Yates, Lee, & Bush, 1997; West & Stanovich, 1997). 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 monitoring accuracy (Erickson & Heit, 2015; West & Stanovich, 1997). In this study, we examined calibration and resolution in a classic domain of general knowledge and compared it to additional domains that have been scarcely studied in this context despite being common in real-life scenarios: probability calculation, financial calculation and the social skill of emotion recognition. We examined whether calibration and resolution show domain generality across these diverse domains. We also examined cognitive abilities (intellectual abilities and working memory) and additional aggregated metacognitive judgments (predictive and postdictive ratings, task difficulty ratings and effort required ratings) as correlates of monitoring accuracy in each of these domains.

Assessing Monitoring Accuracy: Calibration and Resolution

Metacognitive monitoring represents a subjective assessment of the chance of one's own answers to be correct (Bjork, Dunlosky, & Kornell, 2013; Nelson & Narens, 1980). Calibration is one measure to assess monitoring accuracy. This index represents the degree of fit between subjective confidence judgments and the objective accuracy of knowledge or performance (Keren, 1991). The more closely overall level of confidence judgments match success rates, the better calibrated the individual is considered. Poor calibration can happen in both directions, including overconfidence or 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 et al., 2002; Rinne & Mazzocco, 2014). In contrast, underconfidence occurs when confidence judgments are lower than actual performance and reflect 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. Overconfidence can 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. However, studies have demonstrated that individuals 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 been associated with several real-world outcomes, such as several risk behaviors, externalizing behavior 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).

There are different ways to index calibration. The most commonly used score is the bias index, which is the difference between average subjective confidence estimates and objective accuracy (Jackson & Kleitman, 2014). The bias index ranges from -1 to +1 with -1 indicating underconfidence, zero indicating perfect calibration and +1 indicating overconfidence. 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. Thus we needed a measure to index the magnitude of the bias score to measure the raw degree of miscalibration. As the tendency for participants is to display overconfidence rather than under confidence, and to be consistent with the decision-making literature (Stanovich, West & Toplak, 2016), we called this variable the overconfidence index, but we acknowledge that it does assess the magnitude of the miscalibration rather than the direction of the miscalibration. Consequently, we subtracted overconfidence from one to measure resistance to overconfidence, so that higher scores indicate better calibration; this is also consistent with our analyses of the resolution index where a higher score indicates better resolution.

Resolution, also known as discrimination or relative accuracy, is another metacognitive index used to examine monitoring accuracy. Resolution indicators can be used to assess whether a person discriminates between correct versus incorrect performance (Koriat, 2012a). We used the Goodman Kruskal Gamma correlation (Nelson, 1984; but see Masson & Rotello, 2009; Schraw, 2009) which provides a within-subjects measure of the relationship between confidence judgments and the correctness of each response. 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).

Importantly, these two aspects of monitoring accuracy, calibration and resolution, have different functions. Their measures are dissociable—that is, in the same experimental paradigm, calibration might be high, while resolution might be low, or vice versa. For instance, Koriat, Sheffer, and Ma'ayan (2002) showed that, with practice over multiple study-test trials, calibration worsens whereas resolution improves (see also Maki, Shields, Wheeler, & Zacchilli, 2005; see Thiede, Mueller, & Dunlosky, 2015, for a review).

Monitoring Accuracy Across Different Domains

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; see 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 for individual differences in isolation and have rarely been examined in parallel in the same study (see Erickson & Heit, 2015).

Monitoring accuracy has been studied as an issue 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), and some studies have yielded inconsistent findings. 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, studies have yielded a consistent picture of a calibration bias as a characteristic of individuals—people who tend to be overconfident on one type of task, tend to be overconfident on other types of tasks relative to the

others (e.g., Schraw, Dunkle, Bendixen, & Roedel, 1995; West & Stanovich, 1997). For example, Jackson and Kleitman (2014) found that people tend to show a robust bias across tasks (specifically, confidence ratings on a medical decision task and cognitive ability indicators measured by solving Raven Matrices and a vocabulary questionnaire). In contrast, the domain-specific hypothesis suggests that calibration reflects the ability to assess specific content knowledge and that these abilities vary across domains (e.g., Glaser, 1991). For example, Perfect (2003) reported that confidence judgments were more predictive of general knowledge than of eye-witness memory, but notably this study did not specifically examine monitoring accuracy indices.

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 suggest that researchers should use tasks with similar difficulty level for controlling for these differences when comparing monitoring accuracy across domains (e.g., Erickson & Heit, 2015).

To contribute to the understanding of domain similarities and differences, we examined calibration and resolution in the well-studied domain of general knowledge and compared it to three understudied domains in this context which are particularly diverse: financial calculations, probability calculation, and emotion recognition. For each of these domains, the research has been focused either on calibration or on resolution, but no study we know of had both measures and none compared these measures within individuals across these domains.

General knowledge refers to accumulated knowledge across several topics, as opposed to in-depth knowledge about one topic. General knowledge questions are the most well-known and widely-used means of measuring calibration bias (Lichtenstein, Fischhoff, & Phillips, 1977; Yates, Lee, & Bush, 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, Lee, & Bush, 1997).

Financial literacy represents the ability to use numeric information and to make decisions regarding financial planning, wealth accumulation, debt, and pensions (Lusardi & Mitchell, 2014). In real-world financial problems, both specific knowledge about financial concepts and numeracy skills are simultaneously required (e.g., calculating an interest rate). In general, it has been reported that individuals tend to overestimate their ability to apply numeracy skills in 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). For this reason, we chose to study financial calculation literacy 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 have generally 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 & Volpe, 1998; Mandell, 2008). To assess this domain, we developed financial calculation items to assess conversion of currency rates, costs and savings calculations, credit card interest rate calculations, and calculating bank interest rates.

A second numeracy area that we included involves 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, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). Individuals have been reported to be overconfident in their probabilistic numeracy skills (Miron-Shatz, Hanoch, Doniger, Omer, & Ozanne, 2014), which is why we selected this additional domain for this study.

To the best of our knowledge, all metacognitive comparisons across domains that were done so far involved cognitive tasks. We wanted to examine domain-generality even beyond the real-life calculations detailed above. Thus, we included a social skill of emotion recognition. Emotion recognition has typically been assessed using a laboratory paradigm that involves asking participants to identify the emotion displayed in images of various facial expressions (e.g., happy, disgusted). Emotion recognition has been found to be an automatic, universal process, but some people have more difficulty identifying another's emotions than others (e.g., Ekman, 1972; Kohler, Turner, Bilker, Brensinger, Siegal, Kanes et al. 2003; Elfenbein & Ambady, 2002; Adolphs, Sears, & Piven, 2001). Research on confidence judgments for emotion recognition has

been relatively sparse. Kelly and Metcalfe (2011) examined emotion recognition through tasks that required participants to identify emotional expressions and obtained self-ratings of confidence on an item per item basis. They demonstrated strong resolution, suggesting that individuals are good at judging when they know or do not know an emotional expression. It was concluded that people possess good self-awareness in their ability to understand another's emotional expressions. We included emotion recognition as an additional domain for examining both calibration and resolution across domains.

Correlates of Monitoring Accuracy: Individual Differences in Cognitive Abilities and Aggregated Monitoring

Relatively little is known about the role of person-related factors on confidence judgments in reasoning and problem solving contexts, including the role of individual and dispositional differences (Thompson, 2009; West & Stanovich, 1997) and metareasoning processes (see Ackerman & Thompson, 2017, for a review). Performance on knowledge calibration paradigms, such as the overconfidence index, has 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). That is, better calibration was 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 we expected to replicate using our three cognitive tasks (Bruine de Bruin et al., 2007; Stanovich & West, 1998).

While calibration and resolution have been described as conceptually dissociable measures, resolution has also been associated with depth of processing, particularly on reading

comprehension tasks (Thiede et al., 2003). Studies that have examined methods to encourage greater depth of processing in learners have been shown to improve 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, to further understand domain generality and specificity across indices and tasks, we examined whether calibration and resolution are intercorrelated in a similar manner in cognitive tasks and in emotion recognition.

We also collected aggregated judgments, beyond the classic postdictive confidence ratings per item. In particular, we were interested in predictive judgments which are hypothesized to guide allocation and regulation of mental resources for a given task (see Ackerman & Thompson, 2017). Such aggregated predictive and postdictive confidence judgments of performance have been shown to be positively associated with success rates across different domains (Erickson & Heit, 2015). For pre-ratings, individuals would have to estimate based on stored memory structures relevant to a given domain, whereas for post ratings individuals can use their perceived performance on each task as a reference point for their rating, which is related to the distinction between theory-based and experience-based metacognitive judgments (Koriat, 1997). In addition to judgments of performance, we asked participants to rate task difficulty, effort required and their affective reaction to engaging in mental effort, as aggregated judgments provided following each experimental task (Finn, 2010; Hsu, Eastwood, & Toplak, 2017; Hsu, Propp, Panetta, Martin, Dentakos, Toplak, & Eastwood, 2018). In our study, participants provided a separate rating of these dimensions for each task. It was important that these ratings were associated with their respective task so that participants had a specific reference point for their subjective ratings in each domain. We examined correlations between monitoring accuracy indices (calibration and resolution) and these additional aggregated judgments for each task.

In summary, we examined calibration and resolution in the domains of general knowledge, financial calculation, probability calculation, and emotion recognition. We examined whether calibration and resolution would be intercorrelated across individuals between tasks and whether these indices would be correlated with cognitive abilities and aggregated judgments

despite the diverse nature of the tasks. Notably, Jackson and Kleitman (2014) found the strength of intercorrelations of calibration across the medical tasks they used to be stronger than the intercorrelations in resolution, although most correlations were significant. We were interested to see how this pattern is affected by having more diverse tasks.

Method

Participants

A total of 136 participants (*M* age = 20.43 years, *SD* = 2.77; age range 18 – 30 years of age; 34 males and 102 females) took part in the study. The participants were recruited from an undergraduate participant pool of students in psychology. Participants received course credit for participating in this study.

Most participants were in their first year of the undergraduate program (n = 77, 56.5%); 28 were in their second year (20.6%), 18 were in their third year (13.2%), eight were in their fourth year (5.9%), and five were post fourth year students (3.7%). In terms of ethnicity, 43 participants reported White/European (31.6%), 33 reported South-Asian (24.3%), 23 reported Asian (16.9%), 12 reported Black (8.8%), 5 reported Arab (3.7%), six reported Latino-Hispanic (4.4%) and 14 reported Other (10.3%). Twenty-seven participants reported that their current academic average was 80-100% (19.9%), 62 participants that their current academic average was 70-79% (45.6%), 36 participants reported that their current academic average was 60-69% (26.5%), five reported that their current academic average was 50-59% (3.7%) and six participants did not provide a response (4.4%).

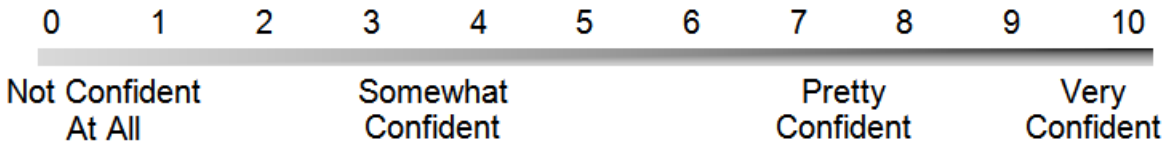
The initial sample included 153 participants, but 17 participants were excluded from data analyses (i.e., four participants reported English as a second language and identified language-related difficulties in understanding and/or completing study tasks; 11 participants did not complete the full task battery due to time constraints; one participant did not complete the tasks correctly; and one participant did not have adequate variability to calculate a resolution score). Thus, the final sample included in the analyses was 136 participants.

Materials, Procedure, and Measures

We developed four monitoring accuracy tasks in different domains, including general knowledge, financial calculation knowledge, probability calculation, and emotion recognition. All four 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 to ensure that the items were not too easy or too difficult.

Prior to completing each task, participants were asked for the aggregated predictive judgments regarding their ability to answer general knowledge questions (such as, "What is the capital of Greece?"), their ability to solve financial calculation problems ("How many Canadian dollars is 30 US dollars?"), their ability to solve probability questions ("Two coins are tossed, what is the probability that two heads are obtained?"), and their ability to read the emotional state of a person (based on only an image of the person's eyes) on the following scale:



Following each experimental task, participants were asked to rate their performance on each task based on the same scale ("How confident are you in your ability to correctly solve general knowledge/real-world calculations/probability questions?" or "How confident are you in your ability to read the emotional state of a person?"; 0=Not Confident at All; 10=Very Confident). In addition, at the end of each experimental task, participants were asked to rate task difficulty ("How difficult did you find this task?"; 0=Not Difficult at All; 10=Very Difficult), effort required ("How much effort was required of you to complete this task?"; 0=No Effort at All; 10=Extreme Effort) and the feeling associated with effort ("How did using that level of effort make you feel?"; 0=Extremely Pleasant; 10=Extremely Unpleasant).

General Knowledge. 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:

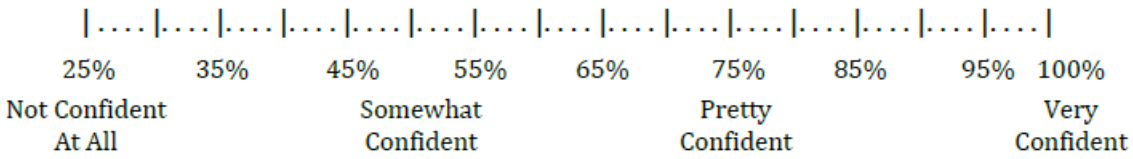
In this survey, you will be asked to answer general knowledge questions. There will be four options for each question (A, B, C, and D). Please choose the option that you think is correct for each question.
Following each question, you will be asked to indicate how sure you are about your answer on a 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

How confident are you that your response is correct? Please circle the number that best represents how sure you are about the answer you just provided.



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 (Lamb) was the correct response. Using the response of "Lamb" (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. Participants then completed 24 additional general knowledge questions in this task.

Mean success rate and mean confidence across all the items were dependent measures. To assess calibration, we used an overconfidence index which is the absolute difference between mean confidence and percentage correct across items. For descriptive statistics in Table 1, we reported the overconfidence index to indicate the magnitude of overconfidence obtained in this sample. For the correlation and regression analyses, we used a resistance to overconfidence index so that a higher score indicates better calibration, measured by subtracting the overconfidence index from one (Bruine de Bruin, Parker, & Fischhoff, 2007). The resolution

index was measured with the use of the Goodman–Kruskal Gamma correlation and involves a within-subjects measure of the relationship between confidence judgments and the correctness of each individual item (e.g., Koriat et al., 2009a). A higher Gamma correlation indicates better resolution.

The general knowledge task originally had 24 items. 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; we selected items to reflect a wide range of difficulty and confidence ratings, while avoiding floor and ceiling effects. In order to ensure adequate variability to calculate the resolution indices, items with overall accuracy of less than 10% or greater than 90% were removed. Eight items (three below 10% accuracy and five above 90% accuracy) were removed, leaving 16 items on this task for analysis[1].

Financial Calculation. This task was developed to assess the financial calculation skill that reflects the types of problems that emerging adults (such as undergraduate students) would need to solve in their everyday lives. We developed eight simple interest problems, four budgeting problems, six credit card problems, and six currency conversion problems. The following is a sample of a budgeting problem:

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

Option C (5 months) was the correct response on this sample item. The procedure and question selection criteria for this task were the same as for the general knowledge task. Six items (five currency conversion items and one simple interest item) were excluded due to floor or ceiling effects, leaving 18 items for analysis.

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

Probability Calculation. This task was developed to assess participants' probability calculation skill. We developed nine items that require calculating the likelihood of an event (such as a coin toss, dice toss or pulling random balls out of a bag), 11 items that require calculating decimals from fractions to identify the correct response and three items involving comparing probabilities in decimal format. The following is a sample of a problem that requires generating decimals to identify the correct response:

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.
Product II has a 3 out of 20 chance of removing the stain.
Product III has a 2 out of 15 chance of removing the stain.
Product IV has a 1 out of 10 chance of removing the stain.

Option A (Product I, 4 out of 25 chance) was the correct response for this sample item. The procedure and item selection criteria were the same as for the previous tasks. Three items (one likelihood problem and two generating decimals from fractions problems) were excluded for inadequate variability in success rates, leaving 21 items for analysis.

Emotion Recognition. The items for this task were taken from the Reading The Mind in the Eyes Task developed by Baron-Cohen (2001). We selected facial expressions reflecting positive (friendly, confident, playful), negative (regretful, hostile, panicked) and neutral (thoughtful, regretful, insisting) emotions. The following is a sample of an item:

Apologetic
Friendly
Uneasy
Dispirited

Option C (Uneasy) was the correct response for this sample item. The procedure and item selection criteria were the same as for the previous tasks. One item was excluded for inadequate variability in accuracy (facial expression was "suspicious"), leaving 23 items for analysis.

The four monitoring accuracy tasks displayed adequate variability in performance (general knowledge: .21 to .94; calculation: .17 to .94; probability: .14 to .95; eyes: .35 to .91) and in confidence ratings (general knowledge: 27.9% to 96.4%; calculation: 29.2% to 100%; probability: 25% to 100%; eyes: 43.7% to 100%). Mean accuracy and confidence scores for each task are displayed in Table 1.

Cognitive Ability Tasks

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. This task included 12 blocks, each consisting of two, three, four, or five sentences. There were a total of 42 items (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, which was the dependent measure on this task. The mean score on this task was 32.10 (SD=4.08), with a range of 20-42. Cronbach's alpha was .66 on this task. Higher scores on this task reflect better working memory abilities.

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 was group administered. On the Vocabulary subtest, participants were asked to choose amongst four alternatives of the definition that most closely matches the target word. On the

Block Patterns subtest, participants were asked to choose amongst four alternatives the block that best completes the design. Scores on the Vocabulary subtest provide a measure of crystallized ability and the Block Patterns subtest provides a measure of fluid reasoning ability. Raw scores (not age corrected) were standardized and summed to create a composite score of general intelligence, called the Intelligence Raw Score Composite. In this study, mean performance on the Vocabulary subtest was 25.92 (SD=4.35), with a range of 16-37, and mean performance on the Block Patterns subtest was 16.18 (SD=4.92), with a range of 5 to 26. 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 indicate better intellectual abilities.

Session structure

Testing sessions were conducted in group format and were approximately 120 minutes in length. A maximum of 10 participants were tested at one time and each testing session included two examiners. Participants first completed the informed consent and demographic forms. Participants then completed the group-administered working memory task first, followed by the measure of intellectual ability. Participants were then each provided with a calculator and asked to independently complete the experimental tasks. The general knowledge task was presented first, followed by either the financial calculation, probability calculation, or emotion recognition tasks (Order 1) or the emotion recognition, financial calculation or probability calculation tasks (Order 2). There were no significant differences in accuracy between the three tasks based on presentation order.

Results

The means (and SDs) of all measured variables are presented in Table 1. As designed to be, mean success rates (global mean = 60.0%) and confidence ratings (global mean = 74.3%) in all tasks were in an intermediate level of difficulty, which allowed confidence variability above and below actual success rates. The small differences in accuracy and confidence among the tasks allow comparisons across the tasks. Interestingly, confidence order was quite similar to the order of success rates, with both highest in emotion recognition and lowest in general knowledge.

The significance of one-sample t-tests for differences of calibration and resolution from zero are marked by asterisks near the means in the table. The overconfidence index in the study reflects the magnitude of miscalibration but does not take into account underconfidence in

responses. In a separate analysis, we subtracted accuracy from confidence for each task and subjected these mean differences to a one sample t-test. Even when underconfident responses were included, we still obtained a significant overconfidence effect across all tasks, $t(135)=8.61$ to 16.22, $p<.0001$. Mean resolution, measured by Gamma correlations[2], was significantly positive in all tasks, indicating that participants discriminated successfully between correct and incorrect responses. The means on the aggregated confidence ratings ranged in the 4-6 range on the scales, indicating that none of the tasks were rated as extremely difficult, requiring extreme effort, or was extremely unpleasant. Table 2 displays the correlations between the pre-confidence, post-confidence, mean accuracy and mean confidence for each of the experimental tasks. In general, these variables tend to be positively intercorrelated, but as might be expected, the magnitude of correlations tends to be larger among the confidence ratings than correlations between confidence and accuracy.

Table 3 displays intercorrelations between the calibration and resolution across the experimental tasks. All the resistance to overconfidence indices were significantly and positively intercorrelated. These correlations indicate that better calibration on one task was associated with better calibration on the other tasks, generalizing previous findings in other domains (e.g., Jackson & Kleitman, 2014). As for resolution, none of the tasks were significantly correlated.

Table 4 displays correlations between the task-based measures, cognitive abilities, and pre- and post-task aggregated ratings. The predictive rating in each domain was consistently correlated with item-by-item confidence. However, these prediction ratings with the other task-based measures varied across the tasks.

The intelligence raw score composite was consistently positively correlated with success rate (accuracy), but not consistently with the other task-based measures. Emotion recognition was the least correlated with this ability measure. Working memory was positively correlated with success rates in most tasks, except for probability calculation. Working memory was also not consistently correlated with mean confidence, calibration and resolution in any task. The calibration index was significantly correlated with either the intelligence raw score composite (for financial calculation and probability calculation) or the working memory score (general

[2] We also examined 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. Across all of the analyses, we found parallel findings using the Gamma and CAQ indices.

knowledge and emotion recognition). Working memory capacity and fluid intelligence, in particular, have been shown to be highly related traits that reflect complementary processes for facilitating complex cognition (Engle, 2018; Shipstead, Harrison, & Engle, 2016).

In general, post confidence ratings and difficulty ratings displayed similar patterns to predictive ratings with success rate, confidence, calibration and resolution. However, while post confidence tended to show stronger associations than predictions, this pattern was less consistent with task difficulty ratings. In general, task difficulty was negatively associated with accuracy and confidence; specifically, higher accuracy and higher confidence were correlated with less perceived difficulty. As mentioned above, confidence and task difficulty seem to differ in their basis, although both could be based on experience with the task. Required effort ratings and feeling of effort (where a higher score indicates less pleasant) showed similar associations in most tasks. Higher accuracy and higher confidence were correlated with lower perceived effort and this effort was perceived as less unpleasant. Both effort-related ratings were uncorrelated with calibration and resolution, with probability calculations being exceptional in this respect (better calibration correlated with a higher rating of unpleasant affect).

Due to their centrality in our research's focus, we delved further into predictors of calibration and resolution by using simultaneous regression analyses to identify unique predictors of monitoring accuracy. We examined whether the cognitive ability indices (either the intelligence raw score composite or working memory total score) and the various aggregated judgments would enter as unique predictors of calibration and resolution on each of the monitoring accuracy tasks.

The results of these regression analyses appear in Tables 5 and 6. All of the regression analyses were statistically significant except for the financial calculation task (for calibration and resolution) and the probability calculation task (for resolution); thus these regression analyses are not included in Tables 5 and 6. The most consistent predictors of both calibration and resolution were cognitive abilities (either the intelligence raw score composite or working memory total score). Cognitive abilities consistently positively predicted resistance to overconfidence in the expected direction, but in the case of resolution, cognitive abilities positively predicted resolution in general knowledge, but negatively predicted resolution in emotion recognition. All other predictors rarely had unique explanatory power. Specifically on the probability calculation task, lower ratings of effort required and the unpleasantness of the effort required entered as unique

predictors of resistance to overconfidence. On the emotion recognition task, lower post confidence and pleasantness of the feeling of effort required uniquely predicted resistance to overconfidence on this task. Only the predictive rating of confidence entered as a unique predictor of resolution on the emotion regulation task: higher confidence predicted better resolution.

In the case of the emotion recognition task, cognitive ability was a positive predictor of calibration and postdictive confidence was a negative predictor. To delve further into this finding, we conducted additional analyses to further elucidate these relationships. In the correlational analyses in Table 4, the working memory total score was positively correlated with calibration, but the predictive rating and postdictive confidence ratings were negatively correlated with calibration. As both pre and post ratings displayed this correlation, we examined the adjustment from pre to post in the confidence ratings in a further regression analysis. We subtracted post task confidence from predictive confidence, and we obtained a mean score of 0.94 (*SD*=2.25), suggesting that participants tended to adjust their confidence lower after completing the task. When we entered the working memory total score, the difference score between the pre and post confidence ratings and task difficulty ratings, the regression analysis was significant, $F(3, 132)=5.77$, $p<.001$, and both the working memory and difference score entered as unique predictors, $t(132)=2.81$, $p=.006$; $t(132)=2.50$, $p=.014$, respectively. Specifically, better working memory and an adjustment to lower aggregate confidence (from prior to after completing the task) were unique predictors of calibration, explaining 5% and 4% of the unique variance, respectively. We conducted a parallel analysis with resolution on the emotion recognition task, but the regression analysis did not reach statistical significance.

Finally, we also compared the pre and post judgments for each monitoring accuracy task. Means for these ratings are presented in Table 1. Participants predicted their probability calculation skills to be higher than after experiencing the task, $t(135) = 4.12$, $p < .0001$. Similarly, participants predicted their ability to recognize emotions as higher before performing the task than in retrospect, $t(135) = 4.83$, $p < .0001$. No differences were obtained on the general knowledge and financial calculation ratings, $t < 1$; and $t < 1$, respectively. The attenuation of the judgments after taking a detailed exam is similar to findings of Illusion of Explanatory Depth (Rosenblit & Keil, 2002).

Discussion

In this study, we examined individuals' confidence accuracy and other monitoring types across a particularly diverse collection of four tasks. In addition to general knowledge, three of the tasks we used have been rarely examined from a metacognitive perspective: financial calculation, probability calculation, and emotion recognition. The association between them and the well-studied general knowledge domain allows connecting our findings to the existing literature for understanding the stability of monitoring accuracy within people across domains.
Overall, calibration was significantly and positively correlated across the experimental tasks, which suggests domain generality. In contrast, we found a lack of correlations among the tasks within individuals in resolution. Our results, in comparison to previous studies with more homogenous tasks, suggest that when the tasks are diverse, resolution might be domain specific. Thus, our results reinforce 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). These findings support the conceptualization of these two indices as complementary but both critical for effective effort regulation (see Ackerman, Parush, Nassar, & Shtub, 2016, for a review).

Across the four experimental tasks, participants displayed overconfidence, which is consistent with what has been 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). In addition, we found reliable resolution across all four tasks, indicating that participants reported higher confidence for correct than incorrect responses. Our collection of tasks does not include tasks which are known to show low resolution, in particular, reading comprehension, as mentioned above (Thiede et al., 2010). The finding that each task in isolation showed reliable resolution strengthens the inconsistency in resolution correlations across the tasks and of the sporadic association between the various cognitive ability and aggregate rating measures with resolution.

Our consistent positive correlations between confidence ratings on our monitoring accuracy tasks are in line with what has been called a self-confidence trait (Jackson & Kleitman, 2014; Kleitman, 2008; Kleitman & Stankov, 2001). All of our tasks required some degree of knowledge, especially our general knowledge, financial calculation, and probability calculation tasks. Lichtenstain and Fischoff (1997) have suggested that tasks where participants lack some

knowledge are characterized by a consistent pattern of overconfidence and no discrimination or resolution, but with increasing knowledge, participants display less overconfidence until accuracy becomes medium-high (e.g., 80%). However, there is evidence that knowledge does not always impact resolution scores (Lichtenstein & Fischhoff, 1997). Prior knowledge would certainly impact performance in the current study, and our findings showing differences between calibration and resolution are consistent with the knowledge conditions studied by Lichtenstein and Fischhoff (1997).

In the current study, the emotion recognition task was clearly unique. It displayed the lowest resolution among the experimental tasks, but individuals were generally quite accurate in recognizing emotions. In addition to specific knowledge required for each domain, other factors that may impact these monitoring accuracy indices are the very nature of these domains. For example, content in some domains may be more "fuzzy" than discrete, which may impact resolution more than calibration. For example, determining the difference between the emotions of shyness and embarrassment may seem a bit fuzzy compared to boundaries in general knowledge, where a capital city of a country is a discrete and absolute fact, which may impact how discriminative our confidence is relative to our accuracy. This fuzziness may characterize domains whereby calibration and resolution are particularly separable.

In addition to comparing these monitoring accuracy indices empirically, methodological and scoring considerations should also be taken into account. For example, if there is inadequate variability in success rates, then resolution scores cannot be derived, but this is less problematic or often not taken into account when deriving the overconfidence index. Indeed, for using the gamma index, it is critical to use items that allow variability (not extremely easy or difficult; Fleming & Lau, 2014). In the current study, we eliminated such items so that we used the same set of responses and participants in order to directly compare findings across accuracy monitoring indices. 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. It is important to highlight these methodological differences in understanding and interpreting these indices. We used what we considered the most conservative approach in order to directly compare these indices, but these issues of scoring criteria and variability may have important implications that warrant further study. Certainly efforts to control for item difficulty

and consistency in inclusion/exclusion of items will be critical for future comparisons of these indices.

Across the regression analyses, higher cognitive abilities (as indexed by either the intelligence raw score composite or working memory total score) predicted better calibration across three of the four tasks, except for financial calculation. 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). A positive association between cognitive abilities and metacognitive monitoring accuracy was also found for one of the four tasks using the resolution index (general knowledge). However, but a negative association was obtained for the resolution index on the emotion recognition task. Behavioral regulation by emotions has been described as an autonomous process characterized by rapid, mandatory, execution when the triggering stimuli do not put a heavy load on central processing capacity (Stanovich, 2009; Stanovich, West & Toplak, 2011). In general, continuous individual differences in autonomous processes are few, which may explain the findings with the emotion recognition task. Suppressing overconfidence may require some input form higher level control systems, but accurate discrimination of emotions may be a more autonomous process. It is more likely that cognitive abilities are unrelated than negatively related to resolution, but further studies and replication will be useful to further elucidate the relationship between monitoring accuracy and cognitive abilities, especially in the unique domain of emotion recognition.

In a few instances, the metacognitive monitoring ratings and postdictive ratings also entered as significant predictors of calibration and resolution. In particular, on the emotion recognition task, we found that participants who adjusted their confidence (indicating less confidence from pre to post ratings) displayed better calibration (resistance to overconfidence). These results suggest that our subjective judgments in some domains may be more amenable to adjustments in the moment, and that such adjustments are importantly related to calibration. In the case of judging faces, people may be particularly sensitive to tracking their success in reading faces, which may partly explain why this is such a well-developed competency (Kelly & Metcalfe, 2011). Given the limited number of studies for confidence in emotion recognition, future work will be needed to replicate and further elucidate these findings. The ratings of effort required and feeling of effort were unique predictors of calibration on the probability calculation and emotion recognition tasks. These findings suggest that the 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 in future studies.

Demonstrating generality in miscalibration across a diverse set of tasks is a novel contribution of this study. We also acknowledge the preliminary nature of our findings on metacognitive monitoring across different domains, especially with respect to our exploratory analyses of the emotion recognition task. Further replication will be critical to extend our understanding of the involved processing and metacognitive mechanisms for parsing their relative contributions throughout processing stages (Ackerman &Thompson, 2017). In particular, including predictions of success in a domain and aggregate post ratings may help us to understand the underlying processes that contribute to monitoring accuracy across different domains.

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, we specifically selected financial calculations and probability judgment given the relevance of these domains for personal financial management (Lusardi & Mitchell, 2014) and the fact that we are presented with probabilistic information in so many facets of our lives (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). Emotion recognition from facial images is an additional real-world domain, which has rarely been studied from a metacognitive perspective (Kelly & Metcalfe, 2011). Overall, we found more consistency 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 eyewitness testimony, medical decisions, and financial decisions. The findings highlight the critical difference between monitoring indices. 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 our overconfidence index, it is worth considering intervention strategies that might facilitate better calibration across domains.

Compliance with Ethical Standards

The research reported in this study involving human participants was approved by the Research Ethics Board at York University. The authors declare that they have no conflicts of interest.

References

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.

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience*, *13*(2), 232-240.

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.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241-251.

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.

Chen, H., & Volpe, R. P. (1998). An analysis of personal financial literacy among college students. *Financial Services Review, 7*(2), 107-128.

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

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.

Ekman, P. (1972).Universals and cultural differences in facial expressions of emotion. In J.Cole (Ed.), Nebraska Symposium on Motivation (pp.207-283). Lincoln: University of Nebraska Press.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin, 128*(2), 203-235.

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science, 13*(2), 190-193.

Erickson, S., & Heit, E. (2015). Metacognition and confidence: comparing math to other academic subjects. *Frontiers in Psychology, 6*, 742.

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

Finn, B. (2010). Ending on a high note: Adding a better end to effortful study. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 36*(6), 1548-1553. doi:10.1037/a0020605

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

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

García, T., Rodríguez, C., González-Castro, P., González-Piendo, J., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning, 11*,139-170. https://doi.org/10.1007/s11409-015-9139-1

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.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, *15*(1), 84-93.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research *Advances in psychology* (Vol. 52, pp. 139-183): Elsevier.

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.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1441-1451.

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.

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.

Kelly, K. J., & Metcalfe, J. (2011). Metacognition of emotional face recognition. *Emotion*, *11*(4), 896-906

Keren, G. (1991). Calibration and probability judgements: 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.

Kohler, C. G., Turner, T. H., Bilker, W. B., Brensinger, C. M., Siegel, S. J., Kanes, S. J., ... & Gur, R. C. (2003). Facial emotion recognition in schizophrenia: intensity effects and error pattern. *American Journal of Psychiatry*, *160*(10), 1768-1774.

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., Ackerman, R., Lockl, K., & Schneider, W. (2009). The memorizing effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology, 102*(3), 265-279.

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

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.

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

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.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In *Decision making and change in human affairs* (pp. 275-324). Springer, Dordrecht.

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*(4), 345-391.

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

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.

Mandell, L. (2008). Financial literacy of high school students. In *Handbook of consumer finance research* (pp. 163-183). Springer, New York, NY.

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.

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. M., Omer, Z. B., & Ozanne, E. M. (2014). Subjective but not objective numeracy influences willingness to pay for BRCA1/2 genetic testing. *Judgment and Decision Making*, *9*(2), 152-158.

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

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 judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *18*(2), 157-168.

Rinne, L. F., & Mazzocco, M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PloS One*, *9*(7), e98663.

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.

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., 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*.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science*, *11*(6), 771-799.

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).

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., & Toplak, M. E. (2011). Intelligence and rationality. In R. J. Sternberg & S. B. Kaufman (Eds.), *Cambridge Handbook of Intelligence* (pp. 784-826). New York: Cambridge 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.

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.

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. 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*.

Touron, D. R., Oransky, N., Meier, M. E., & Hines, J. C. (2010). Metacognitive monitoring and strategic behaviour in working memory performance. *Quarterly Journal of Experimental Psychology*, *63*(8), 1533-1551.

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

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., 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.

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.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111-125.

Table 1
Mean (SD) judgment, success rate, and monitoring accuracy indices across tasks (N=136)

Dependent Measure	General Knowledge	Financial Calculation	Probability Calculation	Emotion Recognition
Pre Question Ratings				
Predictive aggregated judgment (10=very confident)	5.80 (2.09)	5.35 (2.24)	6.24 (2.06)	7.15 (1.74)
Task Performance, Confidence Ratings, Calibration and Resolution				
Success rate	0.53 (0.13)	0.63 (0.20)	0.59 (0.17)	0.65 (0.11)
Confidence	0.72 (0.12)	0.74 (0.18)	0.75 (0.16)	0.76 (0.12)
Calibration (overconfidence)[3]	0.20 (0.12)*	0.14 (0.12)*	0.18(0.12)*	0.15 (0.11)*
Resolution (Gamma)	0.35 (0.28)*	0.43 (0.32)*	0.37 (0.28)*	0.29 (0.23)*
Post Aggregate Ratings				
Post-Task Confidence (10=very confident)	5.87 (1.88)	5.54 (2.95)	5.31 (2.65)	6.21 (2.38)
Task Difficulty (10=very difficult)	3.90 (2.10)	5.62 (2.46)	5.47 (2.17)	4.91 (2.58)
Effort Required (10=extreme effort)	4.62 (2.07)	6.54 (2.47)	6.15 (2.12)	5.32 (2.09)
Feeling of Effort (10=extremely unpleasant)	4.60 (1.74)	5.93 (2.22)	5.69 (1.97)	4.69 (1.83)

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

[3] Higher score indicates larger discrepancy between confidence and accuracy

Table 2

Correlations between pre-confidence, post-confidence, mean accuracy and mean item-by-item confidence across the experimental tasks

	Predictive Confidence General Knowledge	Post-Task Confidence General Knowledge	Mean Accuracy General Knowledge	Mean Confidence General Knowledge		Predictive Confidence Financial Calculation	Post-Task Confidence Financial Calculation	Mean Accuracy Financial Calculation	Mean Confidence Financial Calculation
Predictive Confidence General Knowledge	--				Predictive Confidence Financial Calculation	--			
Post-Task Confidence General Knowledge	.58***	--			Post-Task Confidence Financial Calculation	.51***	--		
Mean Accuracy General Knowledge	.07	.25**	--		Mean Accuracy Financial Calculation	.28**	.59***	--	
Mean Confidence General Knowledge	.31***	.57***	.41***	--	Mean Confidence Financial Calculation	.31***	.69***	.71***	--
	Predictive Confidence Probability Calculation	Post Confidence Probability Calculation	Mean Accuracy Probability Calculation	Mean Confidence Probability Calculation		Predictive Confidence Emotion Recognition	Post Confidence Emotion Recognition	Mean Accuracy Emotion Recognition	Mean Confidence Emotion Recognition
Predictive Confidence	--				Predictive Confidence	--			

Probability Calculation					Emotion Recognition				
Post-Task Confidence Probability Calculation	.39***	--			Post-Task Confidence Emotion Recognition	.44***	--		
Mean Accuracy Probability Calculation	.22*	.42***	--		Mean Accuracy Emotion Recognition	.11	.02	--	
Mean Confidence Probability Calculation	.30***	.60***	.60***	--	Mean Confidence Emotion Recognition	.35***	.55***	.23**	--

*p<.05; **p<.01; ***p<.001

Table 3

Intercorrelations of calibration and resolution across experimental tasks: Calibration (resistance to overconfidence) correlations appear below diagonal and resolution (gamma correlations) correlations appear above the diagonal

	1.	2.	3.	4.
1. General Knowledge	--	-.17	.16	-.15
2. Financial Calculation	.34***	--	.08	<.01
3. Probability Calculation	.22*	.49***	--	-.11
4. Emotion Recognition	.30***	.32***	.28**	--

*p < .05; **p < .01; ***p < .001

Table 4
Intercorrelations between task-based measures, cognitive abilities, and pre/post aggregated ratings across tasks

	Predictive Rating[4]	Intelligence Raw Score Composite	Working Memory Total Raw Score	Post-Confidence	Task Difficulty[2]	Effort Required[2]	Feeling of Effort[2]
General Knowledge Task							
Success rate	.07	.48***	.38***	.25**	-.15	-.12	-.04
Confidence	.31***	.36***	.11	.57***	-.48***	-.25*	-.17*
Calibration[5]	-.24***	.14	.22*	-.21*	.21*	.10	.13
Resolution	.10	.26**	.13	.19*	-.21*	-.16	-.07
Financial Calculation Task							
Success rate	.28**	.46***	.17*	.59***	-.47***	-.46***	-.36***
Confidence	.31***	.36***	.05	.69***	-.46***	-.46***	-.43***
Calibration[3]	.08	.21*	.10	.11	-.12	-.13	-.01
Resolution	-.04	-.02	-.20*	-.16	.12	.13	.04
Probability Calculation Task							
Success rate	.22*	.52***	.12	.42***	-.10	-.24**	-.06
Confidence	.30***	.28**	.01	.60***	-.27**	-.26**	-.25**
Calibration[3]	-.01	.30**	.09	-.10	.10	-.06	.20*
Resolution	.08	.21*	-.02	.08	-.08	-.16	-.06

[4] Ratings specific to each task.
[5] Resistance to overconfidence index, where a higher score indicates better calibration.

Emotion Recognition Task							
Success rate	.11	.28***	.25**	.02	-.09	-.19*	-.13
Confidence	.35***	-.02	-.06	.55***	-.34***	-.22*	-.19*
Calibration[3]	-.21*	.15	.23*	-.37***	.17*	-.05	-.07
Resolution	.21**	-.21*	<.01	.07	-.06	.05	-.10

*$p < .05$; **$p < .01$; ***$p < .001$

Table 5

Simultaneous Regression Results for Resistance to Overconfidence

	Standardized Beta	*t*	Unique Variance Explained
Criterion Variable = Resistance to Overconfidence Index on General Knowledge Task			
Predictive Rating	-0.12	-1.18	1%
Working Memory Raw Score	0.24	2.82**	5%
Intelligence Raw Composite z-Score	0.15	1.74	2%
Post-Confidence	-0.07	-0.64	<1%
Task Difficulty Rating	0.21	1.96	2%
Effort Required	-0.03	-0.29	<1%
Feeling of Effort	0.08	-.96	<1%
Overall Regression: $F(7, 128)=3.71$***			
Multiple $R = 0.41$			
Multiple $R^2 = 0.17$			
Criterion Variable = Resistance to Overconfidence Index on Probability Calculation Task			
Predictive Rating	-0.04	-0.43	<1%
Working Memory Raw Score	-0.01	-0.15	<1%
Intelligence Raw Composite z-Score	0.30	3.60***	8%
Post-Confidence	-0.07	-0.74	<1%
Task Difficulty Rating	0.14	1.21	<1%
Effort Required	-0.26	-2.22*	3%
Feeling of Effort	0.24	2.67**	5%
Overall Regression: $F(7, 128)=3.87$***			
Multiple $R = 0.42$			
Multiple $R^2 = 0.18$			
Criterion variable = Resistance to Overconfidence Index on Emotion Recognition Task			
Predictive rating	-0.03	-0.29	<1%
Working Memory Raw Score	0.18	2.29*	3%

Intelligence Raw Score Composite	0.08	0.94	<1%
Post-Confidence	-0.37	-3.88***	9%
Task Difficulty Rating	0.16	1.57	1%
Effort Required	-0.14	-1.52	1%
Feeling of Effort	-0.22	-2.46*	4%
Overall Regression: $F(7, 128)=5.83$***			
Multiple $R = 0.49$			
Multiple $R^2 = 0.24$			

**$p < .01$; ***p<.001

Table 6

Simultaneous Regression Results for Resolution

	Standardized Beta	*t*	Unique Variance Explained
Criterion Variable = Resolution on General Knowledge Task			
Predictive rating	<0.01	<0.01	<1%
Working Memory Raw Score	0.07	0.75	<1%
Intelligence Raw Composite z-Score	0.22	2.50*	4%
Post-Confidence	0.13	1.13	<1%
Task Difficulty Rating	-0.05	-0.44	<1%
Effort Required	-0.10	-1.07	<1%
Feeling of Effort	0.01	0.16	<1%
Overall Regression: $F(7, 128)=2.37$*			
Multiple $R = 0.34$			
Multiple $R^2 = 0.12$			
Criterion Variable = Resolution on Face Recognition Task			
Predictive rating	0.27	2.81**	6%
Working Memory Raw Score	0.07	0.85	<1%
Intelligence Raw Composite z-Score	-0.24	-2.68**	5%
Post-Confidence	-0.11	-1.06	<1%
Task Difficulty Rating	-0.03	-0.23	<1%
Effort Required	0.10	0.93	<1%
Feeling of Effort	-0.10	-1.03	<1%
Overall Regression: $F(7, 126)=2.29$*			
Multiple $R = 0.34$			
Multiple $R^2 = 0.11$			

*p<.05; **p < .01