2007

A Cue-Utilization Approach to Cognitive Monitoring and Performance: The Effect of Strategy Differences on Monitoring Accuracy

Ainsley Linn Mitchum

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

A CUE-UTILIZATION APPROACH TO COGNITIVE MONITORING AND PERFORMANCE: THE EFFECT OF STRATEGY DIFFERENCES ON MONITORING ACCURACY

By

AINSLEY LINN MITCHUM

A Thesis submitted to the
Department of Psychology
in partial fulfillment of the
requirements for the degree of
Master of Science

Degree Awarded:
Summer Semester, 2007

The members of the Committee approve the thesis of Ainsley L. Mitchum defended on June 8th, 2007.

Colleen M. Kelley
Professor Directing Thesis

K. Anders Ericsson
Committee Member

Joyce Ehrlinger
Committee Member

The Office of Graduate Studies has verified and approved the above named committee members.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Colleen Kelley, who has been an exceptional mentor in many respects. She has contributed greatly to both my professional and personal development. Her unwavering support, encouragement and enthusiasm for research have been an inspiration throughout this project. I would also like to thank my committee members, K. Anders Ericsson and Joyce Ehrlinger for their many thoughtful comments and suggestions.

In addition, I would like to thank my friends, Cari Zimmerman, Tres Roring, Mark Fox, Mike Tuffiash, Katy Nandagopal, and many others who offered their guidance, advice, and moral support. I feel fortunate to be surrounded by so many who are both loyal friends and exceptional thinkers. In particular, I would like to thank Edward Cokely. His patience, guidance, and constant support have been invaluable personally and professionally.

As research is rarely a solitary venture, I would like to extend a heartfelt thanks to my research assistants, Chantel Cooper, Tabata Arvelo, Shayla Jones, John Hartzog, and Tracey Longley who assisted in data collection for this project. Their dedication, attention to detail, and hours of hard work greatly contributed to the quality of this research.

Finally, I would like to thank my family. They have always encouraged me and unconditionally believed in my potential to achieve my goals, no matter how ambitious. Their support throughout my lengthy academic career has been fundamental to my success. I hope my future endeavors continue to make them proud.

TABLE OF CONTENTS

List of Tables vi
List of Figures vii
Abstract viii

INTRODUCTION 1

The Cue-Utilization Approach 2
Strategy Use on Inductive Reasoning Tasks 4

EXPERIMENT 1 6

Method 6
Participants 6
Materials, Design, and Procedure 6
Results 8
Overall Performance on Raven's Matrices 8
Strategy Measures 8
Measures of Monitoring Accuracy 9
Global Performance Judgments 10
Discussion 11

EXPERIMENT 2 13

Method 13
Participants 13
Materials, Design, and Procedure 14
Results 14
Overall Performance 14
Monitoring Accuracy 14
Discussion 15

EXPERIMENT 3 17

Method 17
Participants 17
Materials, Design, and Procedure 17
Results 17
Overall Performance 17
Monitoring Accuracy 17
Discussion 19

GENERAL DISCUSSION 20

APPENDIX: Informed Consent Form..... 33
REFERENCES 34
BIOGRAPHICAL SKETCH 40

LIST OF TABLES

Table 1: Summary of regression predicting calibration in Experiment 1 23

Table 2: Summary of regression predicting estimated percentile in Experiment 1 24

Table 3: Summary of descriptive statistics..... 25

LIST OF FIGURES

Figure 1: Sample Raven's Advanced Progressive Matrices Item 26

Figure 2: Distribution of proportion time on matrix in Experiment 1 27

Figure 3: Graph of mean estimated and actual percentile rank by quartile 28

Figure 4: Distribution of proportion time on matrix in Experiment 2 29

Figure 5: Scatterplot showing cue-utilization by strategy in Experiment 3 30

Figure 6: Calibration plot for control group in Experiment 3 31

Figure 7: Calibration plot for strategy instructed group in Experiment 3 32

ABSTRACT

The ability to accurately monitor and regulate one's cognitive performance is essential to success in a number of settings. What distinguishes between those who can accurately monitor their own performance and those who cannot? Inferential, cue-based approaches to monitoring suggest that monitoring accuracy is influenced by the amount and quality of information available during monitoring. Differences in task strategy may influence monitoring accuracy by bringing about differences in the type and quality of cues available for assessment. A series of experiments explores the complex relationship between task performance, strategies, and monitoring accuracy on a nonverbal inductive reasoning task, the Raven's Advanced Progressive Matrices (RAPM). Results suggest that qualitative individual differences in task strategy influence monitoring accuracy by bringing about differences in the type and quality of cues available during monitoring. Moreover, these differences play a role in the accuracy of participants' subjective confidence. Implications for self-regulation and adaptive cognitive monitoring and control are discussed.

INTRODUCTION

The study of metacognitive monitoring and control examines meta-level processes that guide cognition. Cognitive control refers to processes that organize and sequence mental operations in the service of a specific goal, while monitoring generally focuses on individuals' meta-level knowledge of object level cognitive processes (Nelson, 1996; Nelson & Narens, 1994). Metacognitive monitoring ability has been shown to develop during childhood and adolescence and is considered to be a crucial component of intelligent and adaptive, goal-directed behavior (Sternberg, 1998; Flavell, Friedrichs, & Hoyt, 1970; Borkowski, Carr, & Pressley, 1987; but see also Carr, Alexander, & Schwanenflugel, 1996). Indeed, research has demonstrated numerous examples of peoples' reasonably accurate and effective monitoring performance (Ackerman, Beier, & Bowen, 2002; Arbuckle & Cuddy, 1969; Dunlosky & Nelson, 1994; Koriat, 1997; Koriat & Goldsmith, 1996; Koriat, 1993), as well as cases in which monitoring performance is inaccurate or impaired (Kelley & Sahakyan, 2003; Kruger & Dunning, 1999; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Koriat, 1995; Koren, Seidman, Goldsmith, & Harvey, 2004; Koren et al., 2006).

What distinguishes those who make accurate judgments about their performance on cognitive tasks from those who are poor judges of the accuracy of their own performance? According to some research, monitoring accuracy is determined by and related to overall task performance. Across a number of tasks and domains, poor performers tend to show poorer monitoring accuracy, as well as marked overconfidence (Kruger and Dunning, 1999; Dunning, Johnson, Ehrlinger, and Kruger, 2003; Maki, Shields, Wheeler, and Zacchilli, 2005; Hodges, Regehr, and Martin, 2001; Haun, Zeringue, Leach, and Foley, 2000). However, there are instances where cognitive performance and monitoring are unrelated or where they are differentially affected by the same factors (Chandler, 1994; Kelley & Lindsay, 1993; Mazzoni & Nelson, 1995). In addition, there are also cases where knowledge and skill advantages do not necessarily lead to more accurate monitoring (Glenberg & Epstein, 1987; Lynn, Holzer, & O'Neill, 2006; Gordon, 1991).

An alternative to the view that cognitive and metacognitive performance tap the same set of skills and knowledge is that while cognitive and metacognitive processes are often closely related, they are functionally distinct and rely on different mechanisms. Rather than being based on direct monitoring of the strength of a memory trace, monitoring judgments are seen as being

based on a broad range of inferential cues (Schwartz, Benjamin, & Bjork, 1997; Koriat, 1997; Koriat, 1993). Thus, monitoring accuracy is related to the type and quality of cues available during monitoring, which in turn may depend on the cognitive strategies employed in task performance. Therefore monitoring accuracy need not always be associated with overall performance on tasks. There are many cases in the literature where qualitative and quantitative differences in the amount and quality of information available during monitoring significantly influence monitoring accuracy (e.g. Kelley & Lindsay, 1993; Kelley & Sahakyan, 2003; Nelson & Dunlosky, 1991; Dunlosky & Nelson, 1992; Koriat & Bjork, 2006). As well, individuals differ not only in their overall level of performance on tasks, but also in terms of the specific strategies they use to complete tasks (Snow, 1980; Bethell-Fox, Lohman, & Snow, 1984; Vigneau, Caissie, and Bors, 2006). These qualitative differences in performance strategies should influence the type and quality of cues available during monitoring and thus may be an additional source of individual differences in monitoring accuracy on cognitive tasks. The present experiments examine this possibility by applying a framework originally presented in the metamemory literature, the cue-utilization approach (Koriat, 1997), to a novel domain - retrospective confidence judgments in problem solving.

The Cue-Utilization Approach

Research in metamemory has made progress toward a theoretical framework for understanding the informational basis of monitoring judgments, as well as factors influencing their accuracy. For this reason, much of the research reviewed in the following sections will come from the metamemory literature dealing with judgments of learning (JOLs), which are subjective estimates of the probability that a studied item will be recalled later, and feeling of knowing judgments (FOK), which are typically elicited after a recall failure and indicate participants' confidence that information is stored in memory but cannot be accessed. In addition, I will also discuss the rationale for extending these findings to retrospective confidence judgments in nonverbal reasoning tasks.

There are two main theoretical accounts of how monitoring judgments are made: the direct-access approach (King, Zechmeister, & Shaughnessy, 1980), which describes metamemory judgments as being based on direct monitoring of the strength of a memory trace, and the inferential approach, which suggests that monitoring judgments are based on inferences from cues associated with the likelihood that an item will be recalled (e.g. Koriat, 1997; Schwartz,

Benjamin, & Bjork, 1997; Schwartz, 1994; Reder & Ritter, 1992). Examples of such cues include retrieval fluency (Kelley & Lindsay, 1993), cue-familiarity (Metcalfe, Schwartz & Joaquim, 1993), and beliefs about the memorial consequences of different encoding strategies (Zechmeister & Shaughnessy, 1980).

Koriat's (1997) cue-utilization approach extends the inferential approach to the metamemory domain as a theoretical framework for understanding the effects of different factors on judgments of learning (JOLs). According to the cue-utilization approach, individuals make judgments of learning through access to cues that are generally diagnostic of future recall. Koriat (1997) organizes cues that affect monitoring judgments into three broad classes: Intrinsic, extrinsic, and mnemonic. Intrinsic cues refer to characteristics of study or test items themselves that are generally diagnostic of an item's difficulty. For example, in paired associate learning, the associative relatedness of paired words serves as a cue to the likelihood that an item will be recalled later (Koriat & Bjork, 2005; Dunlosky & Matvey, 2001). The second class of cues, extrinsic cues, consists of external factors related to the conditions of learning or to mental operations applied by an individual. In memory tasks, the number of times an item is studied, distribution of practice (spaced or massed), time constraints (self-paced vs. experimenter paced), encoding operations, incentive conditions, or task strategy would all be considered types of extrinsic cues (Zechmeister & Shaughnessy, 1980; Dunlosky & Nelson, 1994; Mazzoni & Nelson, 1995; Koriat et al., 2006). The third class of cues, mnemonic cues, are defined as internal indicators that signal the extent to which an item is learned (Koriat, 1997). These internal, subjective cues may include the accessibility of related information, cue familiarity, ease of processing of retrieval cues, encoding fluency, and fluency of retrieval are also sensitive to both intrinsic and extrinsic factors (e.g. Hertzog, Dunlosky, Robinson, & Kidder, 2003; Kelley & Lindsay, 1993; Kelley & Jacoby, 1996; Koriat, 1993; Koriat & Ma'ayan, 2005; Morris, 1990; Mazzoni & Nelson, 1995).

Although the cue-utilization approach was originally proposed as a framework for understanding the basis of JOLs, it could be more generally applied to other types of monitoring judgments, such as retrospective confidence in reasoning and problem solving. As in the case of JOLs, there is evidence that retrospective confidence judgments are also cue-based. For example, the perceived difficulty of an item, as well as a person's *a priori* beliefs about their skill with a particular type of problem (e.g. Nelson & Leonesio, 1988; Glenberg & Epstein, 1987; Hertzog,

Dixon, & Hultsch, 1990; Costermans, Lories, & Ansay, 1992; Ehrlinger & Dunning, 2003) could be classified as intrinsic cues contributing to retrospective confidence. Time limits imposed on tasks, test format (free response or multiple choice), or item presentation order (e.g. easiest items presented first) would be examples of extrinsic cues influencing retrospective confidence. Finally, other cues might be based on feedback from one's own cognitive processes, such as how quickly and fluently responses are generated (Kelley & Lindsay, 1993), and strength of supporting and contradicting evidence when evaluating one's chosen response (Koriat, Lichtenstein, Fischhoff, 1980).

Feedback from cognitive operations leading up to the production of a response has been shown to be an important source of information contributing to subjective confidence (Koriat, Ma'ayan, Nussinson, 2006). While part of this feedback may include latency to produce a response, in more complex reasoning tasks additional feedback may be generated from one's specific task strategy, which may include differences in the specific set of mental operations used to complete a task goal or the sequence in which mental operations are carried out (Kyllonen, Lohman, & Woltz, 1984). Thus, different task strategies may also give rise to differences in the quality and amount of information available during monitoring, as well as individuals' sensitivity to this feedback. The present experiments explore the complex relationship between task performance, strategies, and monitoring accuracy on a nonverbal inductive reasoning task, the Raven's Advanced Progressive Matrices (RAPM).

Strategy Use on Inductive Reasoning Tasks

Research examining performance on inductive, analogical, and spatial/geometric reasoning tasks has shown a surprising degree of both inter and intra-individual flexibility in performance strategies (Bethell-Fox, Lohman, & Snow, 1984; Kyllonen, Lohman, & Woltz, 1984; Egan & Grimes-Farrow, 1982). Snow (1980) outlines two primary strategies that participants tend to use on multiple-choice nonverbal reasoning tasks, constructive matching and response elimination. Constructive matching, which is more likely to be favored by high performing participants (Snow, 1980; Bethell-Fox et al., 1984; Vigneau, Caissie, & Bors, 2006) is characterized by a tendency to spend proportionally more time examining each problem before inspecting available answer choices. Converging evidence from verbal reports from subjects and eye-movement analyses suggest that participants spend this extra time constructing a potential answer that is then compared to the presented response options (Snow, 1980; Vigneau, Cassie, & Bors, 2006).

Response elimination, which is more likely to be favored by poor performers, is characterized by a more trial-and-error approach to solving items. Rather than predicting what the correct answer would look like beforehand, those using response elimination tend to compare features of the stimulus items with features of response options hoping to eliminate incorrect responses, in essence, "reasoning backwards" from each potential response (Snow, 1980; Bethell-Fox et al., 1984). Snow (1980) and Bethell-Fox et al. (1984) note that individual participants often don't exclusively use one strategy or the other. Within tasks, some participants initially used constructive matching but switched to response elimination as problems became more difficult. However, the poorest performing participants tended to rely almost exclusively on response elimination or switched to that strategy fairly early in the task.

More recently, Vigneau, Caissie, and Bors (2006), have confirmed the use of constructive matching and response elimination strategies in geometric inductive reasoning tasks and their relationship to performance through an eye-movement analysis of performance on a short form of the Raven's Advanced Progressive Matrices, a test of nonverbal inductive reasoning (Raven, Raven, & Court, 1998; Vigneau & Bors, 2001). The test is made up of 36 items consisting of a 3x3 matrix with the bottom right section removed and 8 response options (see *Figure* 1). Participants are instructed to select the response option that best completes the pattern. Consistent with results reported by Snow (1980) and Bethell-Fox et al. (1984), Vigneau and colleagues report that participants who spend proportionally more time examining the matrix portion of items before examining the response options (taken as an indication of greater reliance on the constructive matching strategy) tend to earn higher scores than those who spend less time examining the matrix. In contrast, participants who show a bias toward inspection of response options (which was taken as indicating heavy reliance on the response elimination strategy) tend to earn lower scores.

Using qualitatively different strategies on a task, in addition to affecting overall performance, may also affect monitoring accuracy through differences in the type and quality of cues available for monitoring. Compared to participants who rely heavily on response elimination, participants who use constructive matching may have a qualitatively different and more diagnostic set of cues to draw from when making monitoring judgments, allowing them to more accurately evaluate their performance. For example, if participants using constructive matching generate a candidate response before viewing the available answer choices, whether or not they find that

response among multiple-choice options provides useful and diagnostic information about the accuracy of the response. This information would not be available to participants relying primarily on response elimination.

EXPERIMENT 1

Experiment 1 evaluates the relationship between spontaneous task strategy, performance, and monitoring accuracy in a nonverbal inductive reasoning task, the Raven's Advanced Progressive Matrices (RAPM). I predict that individuals relying most heavily on constructive matching, when compared to those using response elimination, will show better monitoring accuracy and less overconfidence in item-by-item judgments, as well as in retrospective estimates of their overall performance. The increased monitoring accuracy of participants favoring constructive matching will come from differences in the diagnostic value of available cues. In addition, I predict that differential strategy will explain unique variance in monitoring accuracy, even after controlling for overall task performance.

Experiment 1 also examines the effect of differential strategy use on the *unskilled and unaware* phenomenon (Kruger & Dunning, 1999; Dunning, Johnson, Ehrlinger, & Kruger, 2003), which can be broadly defined as the tendency for those performing poorly on tasks to be overconfident when retrospectively evaluating their own performance. Poor performance is often associated with the use of less adaptive task strategies (e.g. Snow, 1980; Bethell-Fox et al., 1984). Poor strategies provide less diagnostic information on the item-by-item level, thus the use of poor strategies may contribute, at least in part, to poor performers' inaccurate self-assessments. Therefore, I predict that participants using response elimination will show more overconfidence in retrospective judgments of overall performance, as well as in comparative judgments, such as percentile rank estimates.

Method

Participants

Participants were 55 Florida State University undergraduates recruited from the general psychology participant pool. Participants received course credit in exchange for their participation. Data from 5 participants were excluded from all analyses because they indicated that they had previously participated in another experiment using the RAPM.

Materials, Design, and Procedure

Participants were tested individually in a single session lasting about 1 hour, during which they completed the Raven's Advanced Progressive Matrices, followed by a post-experimental questionnaire assessing strategy use and motivation.

Raven's Advanced Progressive Matrices. Participants first completed a computer-administered version of RAPM, set II (Raven, Court, & Raven, 1988) that was modified to assess strategy use. The standard paper-and-pencil version of the task includes 36 items that are presented in ascending order of normative difficulty. Each item consists of a 3x3 matrix of geometric figures or patterns in which the bottom right cell is empty. Participants are instructed to select the piece that completes the pattern, both down the column and across the rows, from the 8 response options presented below the matrix. In the present experiment, rather than presenting the matrix and response choices simultaneously, as is typically done, each item was displayed in two separate sections. The first display consisted of only the 3x3 matrix. Participants were free to view the matrix for as long as they wished and were instructed to press the space bar to display a second screen showing both the 3x3 matrix and the 8 response choices. Participants entered their responses while the second screen was displayed. Reaction time for the matrix screen was collected, starting as soon as the matrix was displayed until the key press that initiated the full item screen. Reaction time for the duration of the full item screen was also collected, starting immediately when the full item screen was displayed and terminating when a response was entered. The reaction times from these two screens were used to determine strategy use, as noted below.

Monitoring Judgments. Following each RAPM item, participants were asked to rate their confidence in the likelihood that their response was correct on a 12% to 100% scale. After entering confidence judgments, participants were free to move on to the next test item. After completing all 36 items, similar to Kruger and Dunning (1999), participants were asked to estimate their total raw score, the raw score of the average Florida State University introductory psychology student, as well as their own percentile rank as compared to other FSU introductory psychology students. Percentile ranks were explained as the percentage of students that would earn a score lower than the participant's own score and could range from 0% (very bottom) to 100% (very top).

Post-Experimental Questionnaire. After completing the RAPM, participants completed a brief questionnaire asking about their strategy use on the task. With respect to strategy use, participants were asked which of the following potential strategies they used most often on the task: a.) Looked at each response choice until I found one that seemed to fit, b.) Tried to predict what the correct answer should be and then searched for it among the response options given, c.)

A little of both, d.) Other. Participants were asked to provide additional detail about their problem solving approach if they selected either of the latter two options. In addition, participants were also asked to rate their motivation to perform well, as well as their overall effort on the task using a 1 (not motivated at all) to 7 (very motivated) scale. Two additional free-response questions were included that asked participants if there were any special circumstances that may have affected their performance (e.g. not enough rest, hungry, ill, etc…), as well as whether they have participated in any other experiments using similar matrix-reasoning tasks. Participants indicating that they had participated in similar experiments were excluded from all data analyses.

Results

Overall Performance on RAPM

Overall performance on the RAPM was similar to the normative sample of 506 first year university students collected by Bors and Stokes (1998). In the current sample, the average score was 21.44 (SD = 5.57) correct responses out of a possible 36, whereas Bors and Stokes (1998) reported a mean of 22.17 (SD = 5.6). Percentile ranks presented for the current sample were calculated based on the Bors and Stokes (1998) norms.

Strategy Measures

Participants' strategy use was operationalized as the average proportion of total time per problem spent examining the matrix portion before displaying response choices:

$$S_u = \frac{\Sigma(\text{Matrix RT} / \text{Problem RT})}{36}$$

Following Vigneau et al. (2006), spending a larger proportion time viewing the matrix alone was taken as an indication of greater reliance on constructive matching. Participants varied greatly in the proportion of time spent examining the matrix (M = .26, SD = .21), ranging from as little as 3% to as much as 72% of total time. The distribution of strategy use in Experiment 1 was not normal, as confirmed by significant Shapiro-Wilk test, $W = .88$, $p < .001$ and was slightly positively skewed (skewness = .61). As well, the distribution appeared to be bimodal (see *Figure 2*), which offers additional support for the hypothesis that participants were indeed using two qualitatively different types of strategies. Consistent with past studies (Snow, 1980; Bethell-Fox et al., 1984; Vigneau et al., 2006), strategy use was related to overall performance, $r(48) = .33$, $p = .02$, indicating that higher performing participants also tended to favor the constructive matching strategy.

Participants' responses on the post-experimental strategy questionnaire generally matched the objective measure of average proportion time on matrix as a measure of strategy use. Participants who reported using response elimination alone (N = 6) spent, on average, only 6% of their total time per problem examining the matrix alone before displaying answer choices. Those reporting using constructive matching (N = 20) spent an average of 33% of their total time examining the matrix alone. Participants who reported using a combination of the two strategies (N = 22) spent, on average, 26% of their total time examining the matrix alone.

Measures of Monitoring Accuracy

Relative Accuracy. Relative accuracy or resolution measures the degree to which confidence judgments for individual items distinguish between correct and incorrect responses. The Goodman-Kruskal gamma coefficient, a nonparametric within-subjects correlation, is the most widely used measure of the relative accuracy for item-by-item judgments (Nelson, 1984). Values for gamma range from –1 to +1, with higher values indicating better relative accuracy. It is important to note that gammas need not be related to overall performance on the RAPM (Nelson, 1984).

The Goodman-Kruskal gamma coefficient between confidence and problem accuracy was calculated for each participant as a measure of relative accuracy. Participants showed considerable variability in relative accuracy with gammas ranging from .14 to .98 (M = .70, SD = .19). As predicted, strategies were related to gamma, r (48) = .41, p = .003, indicating that participants relying more heavily on constructive matching had better relative accuracy. However, gamma was unrelated to overall performance, r (48) = .17, p =.29.

Absolute Accuracy. Absolute accuracy or calibration (also called bias), examines the magnitude of the difference between one's level of subjective confidence and actual performance, indicating overconfidence, underconfidence, or perfect calibration. This is typically done by plotting a calibration curve that displays actual performance as a function of subjective confidence ratings. For example, perfect calibration would be indicated when items to which an individual assigns a 50% probability of being correct are correct 50% of the time, and so forth across the scale.

Participant confidence judgments for each of the 36 items on the RAPM were divided into 11 discrete categories (0-12, 13-20…91-99, 100). Calibration error scores (Oskamp, 1962) were

calculated for each participant as the weighted mean of the absolute differences between the mean confidence and actual proportion correct for each confidence grouping:

$$C = \frac{\Sigma(n\ |p\text{-}c|)}{N}$$

Where n is the total number of observations at each confidence level, p is the assessed confidence level, p is the actual proportion correct, and N is the total number of observations.

As in past studies (e.g. Stankov, 1998), participants were fairly well calibrated (M = .20, SD = .10), but showed considerable variability. Calibration error scores ranged from .06 to .51, with higher scores indicating poorer calibration. As predicted, calibration error scores were related to overall performance such that those earning higher scores on the RAPM tended to have smaller calibration error scores, r (48) = -.56, $p < .001$. More importantly, calibration error scores were also related to strategy use such that those relying more heavily on constructive matching had smaller calibration error scores, r (48) = -.51, $p < .001$.

Hierarchical regression analysis was used to test the hypothesis that strategy use accounts for unique variance in calibration, even after controlling for performance (see *Table 1* for a summary). Calibration error scores were regressed on overall RAPM performance and strategy use. Together, these predictors accounted for a significant amount of variance in calibration, F (2,47) = 15.43, $p < .001$, adjusted $R^2 = .41$. Performance alone accounted for a significant proportion of the overall variance in calibration error scores, r (48) = -.38, $p = .003$. More importantly and as predicted, strategy use also accounted for significant unique variance in calibration error, r (48) = -.40, $p = .002$.

Global Performance Judgments

Following Kruger and Dunning (1999), a series of analyses examined the accuracy of participants' global assessments of their own performance, as well as the accuracy of comparative judgments (percentile rank estimates). As described above, percentile ranks for each participant were calculated using the norms reported by Bors and Stokes (1998) and compared with participants' estimates. Consistent with past research (e.g. Kruger & Dunning, 1999; Ehrlinger & Dunning, 2003), though to a lesser degree, participants tended to overestimate their own performance relative to other introductory psychology students. On average, participants believed their performance fell in the 52nd percentile (estimates ranged from 0 to 80, SD = 15.07), which was significantly higher than the true average (49th percentile), t (49) = 2.99, $p = .004$ (see *Figure* 3). However, participants did not overestimate their raw score on the RAPM. On

average, participants estimated that they had answered 20.82 (SD = 6.21) problems correctly, which was not significantly higher than the actual average of 21.44 ($t < 1$). As in past research (e.g. Kruger & Dunning, 1999; Ehrlinger & Dunning, 2003) the relationship between performance estimates and actual performance depended on the measure used. As expected, there was a strong relationship between predicted and actual raw scores, r (48) = .61, $p < .001$. However, the relationship between actual and predicted percentile rank reached only marginal significance, r (48) = .27, p = .06.

The *unskilled and unaware* hypothesis suggests that the accuracy of comparative self-assessments (self-estimates of percentile rank) is strongly related to overall task performance, that is, participants performing poorly on tasks tend to be grossly overconfident in their estimates while those earning the highest scores show slight to moderate underconfidence. According to Kruger and Dunning (1999) this relationship between monitoring accuracy and performance arises because the same skills required to perform well on a task are also required to accurately assess one's own performance. However, in addition to possessing different levels of skill, the highest and lowest performing individuals may also differ in their task strategies. As the results presented above suggest, differential strategy use may affect monitoring accuracy independent of task performance. Thus, strategy differences may also contribute to the *unskilled and unaware* phenomenon. To test this hypothesis, overall performance and strategy use were used to predict participants' predicted percentile rank in a hierarchical regression analysis. The overall model was marginally significant, F (2, 47) = 2.61, p = .08 (see *Table 2* for a complete summary). Participants' estimated percentile rank was significantly related to their actual percentile rank, r (48) □ = .32,p = .03, but not to strategy use, r (48) □ =-.18, p = .22.

Discussion

The results of Experiment 1 demonstrate that there is a relationship between spontaneous task strategy and monitoring performance. Participants who used the constructive matching strategy consistently showed advantages both in task performance and monitoring accuracy. Advantages in monitoring accuracy were found for both relative and absolute accuracy measures. Additionally, the relationship between strategy use and monitoring was stronger than the relationship between strategy use and performance. More importantly, strategies accounted for unique variance in monitoring accuracy, even after accounting for performance differences. Taken together, these results suggest that individuals who differ in monitoring accuracy differ

from each other not only in terms of their skills and knowledge, but also in the strategies they use to perform cognitive tasks.

Differences in strategy are predicted to influence monitoring accuracy by changing the quality and diagnostic value of the cues available when making confidence judgments. For example, participants using constructive matching generate candidate responses that are compared to presented response options. If one's generated candidate answer isn't found among the given options, this is a very salient cue that one's generated response is likely incorrect. Having such information would give participants using constructive matching an additional opportunity (one not available to those using response elimination) to revise incorrect responses and adjust confidence ratings accordingly. This cue would help reduce overconfidence at the item-by-item level, which would also improve one's global monitoring performance. However, the hypothesis that differential strategy use contributes to the *unskilled and unaware* phenomenon was not supported. Participants' actual performance was the only significant predictor of global estimates of performance, as measured by percentile rank estimates, and that factor predicted only about 7% of the total variance. This finding suggests that other factors, such as self-views of skill, underlie individuals' percentile rank estimates (Ehrlinger & Dunning, 2003).

EXPERIMENT 2

Although the results of Experiment 1 demonstrate a clear relationship between spontaneous task strategy and monitoring accuracy across several measures, they do not allow causal inferences regarding the role of strategies in determining monitoring accuracy. The purpose of Experiment 2 was to manipulate task strategy between-subjects on the RAPM to determine if differential strategy use can causally influence monitoring performance. If qualitative differences in task strategy causally influence the type and quality of cues available, then instructing participants to use constructive matching should lead to better absolute and relative monitoring accuracy, as well as more accurate global assessments of performance, regardless of performance differences.

Perhaps the most salient and diagnostic cue available to participants using constructive matching is the absence of one's constructed response among presented response options. In the present experiment, one indication that participants are using this cue would be the within subjects relationship between time to enter a response (subtracting out time-on-matrix) and confidence. Participants using constructive matching tend to spend a greater proportion of their overall time looking at the screen containing the matrix alone. This suggests that, when the matrix and answers are both displayed, participants using constructive matching are searching for their constructive answers. By this reasoning, when the constructed answer is present among the presented options, those using constructive matching should have relatively quick response times on the full problem screen and should be highly confident in their responses. On the other hand, when the constructed response is not found among the presented options, participants should take longer to enter a response and should be less confident in their response. Therefore, I predict that the strategy instructed group will show a moderate to strong negative correlation between response times on the full matrix screen and confidence ratings. In the uninstructed condition, this relationship should be significantly weaker and related to measures of strategy use (proportional time on matrix).

Method

Participants

Participants were 72 Florida State University undergraduates recruited from the general psychology participant pool. Participants received course credit in exchange for their participation. A total of 7 participants (4 had participated in similar experiments, 3 failed to

follow directions) were excluded from all analyses. All 3 participants who failed to follow directions were in the experimental condition, of these participants 2 did not draw candidate answers on 3 or more problems and one participant changed his or her drawings after viewing the response options.

Materials, Design, and Procedure

The basic procedure for Experiment 2 was the same as in Experiment 1. Participants were randomly assigned to either the strategy instructed group (N = 32) or the control group (N = 33). Participants in the experimental group were instructed to use the constructive matching strategy to solve all items. In order to ensure compliance with the strategy instructions, participants in the strategy condition were instructed to draw their candidate response on an answer sheet while viewing the matrix portion of each problem. An experimenter checked that participants had drawn a candidate answer before they were allowed to view the answer choices. Participants who did not draw candidate responses for 3 or more items were excluded from all analyses. As in Experiment 1, participants made confidence judgments after completing each problem. After completing all 36 items, participants completed a brief questionnaire asking about their strategy use on the task. In addition, they were also asked about their level of motivation during the task, if there were any special circumstances that may have affected their performance (e.g. not enough rest, hungry, etc…), and whether they have participated in any similar experiments. Those who had participated in other experiments using the RAPM were excluded from all analyses.

Results

Overall Performance

For the entire group, performance on the RAPM was similar to Experiment 1 (M = 22.45, SD = 5.6) and did not differ between the two experimental groups, $F < 1$. See *Table 3* for a summary of descriptive statistics for all key measures.

Monitoring Accuracy

As in Experiment 1, measures of absolute and relative accuracy were computed for all participants. Contrary to predictions, the constructive matching group did not show better monitoring accuracy than the uninstructed group. Participants in both groups showed good relative accuracy, as measured by gamma, but did not differ significantly from one another, $F(1,63) = 1.01$, $p = .32$. As well, participants showed good calibration overall, as measured by calibration error scores. Calibration error scores also did not differ significantly between

experimental groups, $F < 1$. See *Table 3* for a summary of descriptive statistics for all key measures.

The within subjects correlation between time to enter a response on screen two (with the matrix and responses) was calculated for all participants. To test the hypothesis that participants in the strategy instructed condition relied on time to enter a response as a cue to a greater extent than participants in the control group, a one-way ANOVA was used to compare the strength of this correlation between groups. Contrary to predictions, there were no significant differences between groups, $F < 1$.

Discussion

The results of Experiment 2 did not support the hypothesis that differential strategy use causally influences monitoring accuracy. However, the lack of a strategy effect could be a result of two potential factors. First, when compared to participants in Experiment 1, participants in Experiment 2 showed better relative accuracy overall, as average gammas differed significantly across the two experiments, F (1, 113) = 4.45, p = .04. When calibration error scores were compared across experiments, participants in Experiment 2 showed a slight advantage. However, this difference trended toward, but did not reach significance, F (1,113) = 1.92, p = .17. Given these findings, it is possible that the failure to find a significant difference in monitoring between groups could, in part, be a result of a ceiling effect for gammas in Experiment 2.

This improved monitoring for participants in Experiment 2 could be a result of sampling error. First, because the data for Experiment 1 was collected across an entire semester, it is more likely that the sample was more diverse, both in terms of strategy use and performance. Experiment 2 took place during the first half of a single semester, during which more motivated students tend to register for experiments. These more motivated students may have tended to select more effective strategies. In addition, because the control condition in Experiment 2 was a smaller sample overall compared to Experiment 1 (N = 33 in the control condition for Experiment 2 vs. N = 50 in Experiment 1), it is possible that the distribution of strategy use could differ between experiments. Although the mean strategy use (as measured by proportion time on matrix) did not differ between Experiment 1 and the control condition in Experiment 2, $F < 1$, the distribution of strategy use did differ between the two experiments. In Experiment 2, the distribution of the strategy variable, proportion time on matrix, was normally distributed (see *Figure 4*), as confirmed by a nonsignificant Shapiro-Wilks test, W = .96, p = .23, suggesting that a

larger proportion of participants in Experiment 2 were relying on constructive matching. In contrast, the distribution of the proportion time on matrix variable in Experiment 1 was non-normal and positively skewed, suggesting that most participants did not use constructive matching. A final indicator that more participants in the control group were spontaneously using constructive matching comes from the average magnitude of the correlation between time to enter a response and confidence. It was predicted that participants using constructive matching would show a moderate to strong negative correlation between these two variables. In Experiment 2, both groups showed at least a moderate negative correlation between time to enter a response and confidence. For the experimental group, the average correlation was -.39, while the average correlation in the control group was -.43. In summary, the smaller sample size of the control group for Experiment 2, possible sampling error, and differences in the distribution of the key strategy variable could all have contributed to the nonsignificant findings in Experiment 2.

Items on the RAPM are typically administered in ascending order of difficulty. This would serve as an extremely salient, and diagnostic cue for participants' confidence judgments. It is possible that participants relied heavily on this cue and did not need to attend to feedback generated from their own mental effort when making confidence judgments. In order to test this possibility, gamma coefficients were calculated between item difficulty, as measured by normative pass rates reported for each item in the RAPM manual (Raven, Court, & Raven, 1998), and participant confidence. Indeed, participants in Experiment 2 did appear to rely heavily on item gradient as a cue (M = -.64, SD = .18), with difficult items that were given later in the test receiving lower confidence ratings. There were no significant differences between the two experimental groups, $F < 1$. Participants in Experiment 1 (M = -.58, SD = .22) also showed a strong reliance on item gradient as a cue, but to a somewhat lesser extent than participants in Experiment 2. However, when both experiments were compared the difference reached only marginal significance, $F(1, 113) = 2.75$, $p = .10$.

EXPERIMENT 3

The purpose of Experiment 3 was to manipulate strategy to test the hypothesis that strategies causally affect monitoring. Because the influence of item gradient as a cue may have overshadowed an effect of strategy use in Experiment 2, RAPM items were administered in a random order rather than in ascending order according to difficulty, as is typically done. This manipulation was intended to both reduce the influence of item gradient as a cue and make monitoring more difficult, particularly for participants in the control group.

Method

Participants

Participants were 69 Florida State University undergraduates recruited from the general psychology participant pool. Participants received course credit in exchange for their participation. In Experiment 3, a total of 12 participants were excluded from final analyses (8 had participated in similar experiments, 4 failed to follow directions), leaving 28 participants in the experimental condition and 29 in the control condition.

Materials, Design, and Procedure

The basic procedure for Experiment 3 was identical to Experiment 2, with the exception that RAPM items were presented to participants in a random order. As in the previous experiments, participants who indicated that they had participated in another experiment using the RAPM, as well as those who failed to follow given directions, were excluded from all analyses.

Results

Overall Performance

For the entire group, the mean RAPM score was 18.97 (SD = 6.42) and did not differ between the two experimental groups, $F < 1$. See *Table 3* for a summary of descriptive statistics for all key measures.

Monitoring Accuracy

As in the previous experiments, gamma coefficients were calculated for each participant as a measure of relative accuracy. As predicted, participants in the strategy instructed condition (M = .79, SD = .12) showed better relative accuracy than controls (M = .68, SD = .23), F (1, 55) = 5.77, p = .02, d = .70. In addition, calibration error scores were calculated for each participant as a measure of absolute accuracy. Participants in the strategy instructed condition showed significantly better absolute accuracy than controls, F (1, 55) = 4.68, p = .04, d = .60. Because

calibration error scores do not indicate whether participants were over or underconfident, calibration curves were plotted for the two experimental groups. Consistent with predictions, the constructive matching instructed group showed less overconfidence, particularly at the highest levels of confidence (see *Figure 6* and *Figure 7*).

Participants in the strategy condition were hypothesized to have better monitoring accuracy through differences in the quality of monitoring cues available. The within-subjects correlation between time to enter a response (on the screen displaying the matrix and answers) and confidence was computed for all participants. The magnitude of this relationship was compared between experimental groups. Participants using constructive matching were predicted to show a moderate to strong negative correlation between time to enter a response and confidence. Short response times, which would be an indication that one's generated answer was found among the presented options, should be associated with high confidence. On the other hand, longer response times, which would be an indication that they did not find their generated answer among response choices and are trying to revise their response, should be associated with lower confidence. Because participants using response elimination do not generate candidate responses before displaying the full matrix screen, they should show a weaker relationship between time to enter a response and confidence. Indeed, as predicted, these correlations differed significantly between the two groups, $F\,(1,55) = 36.93$, $p < .001$. In the strategy group, participants taking longer to enter their responses tended to be less confident in their responses, (Mean correlation, $r\,(26) = -.38$, SD = .18). Participants in the uninstructed group did not show any relationship between time to enter their response choice and confidence ratings, (Mean correlation, $r\,(27) = -.08$, SD = .19).

It is possible that the between groups difference in the relationship between time to enter a response and confidence is primarily a result of the changed structure of the task for participants in the constructive matching group (who had to draw responses on paper). For example, the process of searching for a drawn response among visually presented options might be qualitatively different from searching for an internally generated candidate answer. However, it is likely that some participants in the uninstructed group spontaneously used constructive matching. If this is the case, then there should be a significant relationship between the strategy measure used in Experiment 1 (proportion time spent on matrix) and the relationship between time to enter a response and confidence. Indeed, within the uninstructed group, there was a significant negative

relationship between the strategy measure, proportion time on matrix, and the within subjects correlation between time to enter a response and confidence, r (27) = -.54, p = .003. That is, participants spending proportionally longer times on the matrix section of problems also showed a strong negative relationship between time to enter a response on the full problem screen and confidence. This finding suggests that participants in the control condition who spontaneously used constructive matching, similar to participants in the strategy instructed condition, also showed evidence that they were comparing a generated response to presented response choices (see *Figure 5*). On items where the generated response option was found among the given options, participants using constructive matching entered their responses quickly and were also more confident.

In order to determine if presenting test items in a random order did indeed reduce participants' reliance on item gradient as a confidence cue, the gamma coefficient between item difficulty and confidence was again calculated for each participant. As predicted, participants in Experiment 3 showed a significantly weaker relationship between item normative difficulty and confidence ratings (M = -.59, SD = .18) compared to participants in Experiment 2 (M = -.64, SD = .18), F (1,120) = 9.15, p = .003. There were no between-group differences in the use of item gradient as a cue in Experiment 3, $F < 1$.

Discussion

Experiment 3 demonstrates that strategy use can causally affect monitoring accuracy. However, this effect depends on the absence of more salient and diagnostic cues, such as an item gradient on the test. When test items were presented in a random order, participants instructed to use constructive matching showed superior monitoring accuracy across measures of absolute and relative accuracy. This advantage in monitoring accuracy occurred in the absence of performance differences, suggesting that the observed differences can be primarily attributed to differential strategy use.

GENERAL DISCUSSION

Accurate monitoring is essential to the effective control of cognition and behavior. For this reason, a large volume of research has focused on understanding the informational basis of monitoring judgments, as well as factors affecting the accuracy of such judgments. Previous research has shown that monitoring accuracy is frequently associated with ability or overall performance on tasks (e.g. Kruger and Dunning, 1999; Dunning, Johnson, Ehrlinger, and Kruger, 2003; Maki, Shields, Wheeler, and Zacchilli, 2005). However, there are also cases in the literature where monitoring and performance are dissociated or differentially affected by the same factors (Chandler, 1994; Kelley & Lindsay, 1993; Mazzoni & Nelson, 1995). Direct access approaches to monitoring are unable to account for such results. As a result, cue-utilization approaches to monitoring (such as Koriat, 1997) have become more widespread, receiving strong support in the metamemory literature. Such approaches have the advantage of flexibly specifying the informational basis of monitoring judgments and being broadly applicable to a wide range of cognitive tasks. Recent studies have demonstrated that cue-utilization approaches are useful, not only for memory tasks but for also for other types of complex reasoning tasks (Koriat, et al., 2006, Experiment 7; Koren et al., 2006). The present experiments present further support for the application of the cue-utilization approach to complex reasoning tasks, as well as for the idea that feedback from one's own cognitive operations can serve as a monitoring cue (Koriat et al., 2006). In addition, these experiments also offer unique evidence that differences in task strategy can lead to differences in cue availability and can causally affect monitoring accuracy, independent of overall task performance.

Experiment 1 established a link between spontaneous task strategy, overall task performance, and monitoring accuracy. Higher performing participants tended to select more effective task strategies. However, strategies were found to account for unique variance in monitoring accuracy, even after differences in overall performance were taken into account. These findings suggest that the superior monitoring accuracy of those performing well on tasks is not only due to differences in ability, but is also a result of the selection of more effective and adaptive strategies.

Strategy use was not found to be a significant factor in the *unskilled and unaware* phenomenon. Global assessments of performance (as measured by percentile rank estimates) were unrelated to strategy use, suggesting that other factors exert a larger effect on global self-

assessments of skill. However, there could be a link between self-estimates of skill and strategy selection. Exploring this possibility could be a direction for future research.

In Experiment 2 task strategy was manipulated between subjects to determine whether strategy differences could causally affect monitoring performance. Although Experiment 2 failed to find an effect of strategy use on monitoring accuracy, the results did provide evidence that a strategy effect might be overshadowed if a more salient cue is available. Participants in Experiment 2, rather than relying on feedback from their own solution effort, relied more heavily on another cue, that test items were presented in ascending order of difficulty.

Experiment 3 demonstrated that that differential strategy use causally affects monitoring accuracy, independent of performance differences. In order to reduce the effect of item gradient as a cue, test items were presented in a random order. Across several different measures of monitoring, participants instructed to use constructive matching consistently demonstrated better monitoring accuracy and were less overconfident. Further analyses established that these between-groups differences in monitoring accuracy were related to differential cue use between the constructive matching and uninstructed groups. Within the uninstructed group, participants spontaneously using constructive matching also showed similar patterns of cue use.

Performance on the RAPM, and on intelligence tests in general, predicts a number of real-world behaviors ranging occupational and educational performance to health outcomes (Hunter & Schmidt, 1996; Neisser, Boodoo, Bouchard, Boykin, Brody, Ceci, et al., 1996). Research by Koren et al. (2004) has demonstrated that monitoring accuracy on an executive functioning task, the Wisconsin Card Sorting Task, was a better predictor poor insight (which is associated with treatment outcomes) in patients with schizophrenia than conventional WCST scores. Future studies could investigate whether monitoring performance on intelligence tests, such as the RAPM, may also predict additional variance in real-world performance indicators (e.g. school performance).

The present studies measured only one indicator of strategy use, the average proportion of time spent viewing the matrix section of each item. However, it is likely that participants draw from a much more diverse set of cues than can be measured through reaction times alone. Future research could apply process-tracing methods, such as verbal protocol analysis (Ericsson & Simon, 1980) to further identify the cues individuals attend to during monitoring.

In conclusion, the present results offer support for extending cue-utilization approaches to complex cognitive tasks and highlight the important role of individual differences in task strategy on monitoring performance. The broad application of cue-utilization approaches to complex cognitive tasks offers an additional tool for understanding the informational basis of monitoring judgments, as well as the factors underlying accurate monitoring in a wide range of real-world and laboratory tasks.

Table 1. *Summary of Regression Results for Experiment 1 predicting calibration error scores.*

	Beta	*t*	*Sig.*	*Correlations*			*Tolerance*
				Zero Order	*Part*	*Partial*	
Overall Model (R^2=.44, Adj. R^2=.41)							
Mean Accuracy on RAPM	-.38	-3.15	.003	-.51	-.42	-.36	.89
Proportion Time on Matrix	-.40	-3.29	.002	-.52	-.43	-.37	.89

Table 2. *Summary of regression results for Experiment 1 predicting participants' estimated percentile rank.*

	Beta	*t*	*Sig.*	*Correlations*			*Tolerance*
				Zero Order	*Part*	*Partial*	
Overall Model (R^2=.10, Adj. R^2=.06)							
Actual Percentile Rank	-.32	2.21	.03	.27	.31	.31	.91
Proportion Time on Matrix	-.18	-1.25	.22	-.08	-.18	-.17	.91

Table 3. *Summary of Descriptive Statistics*

	Raven's Mean (SD)	*Gamma* Mean (SD)	*Calibration Error* Mean (SD)
Experiment 1 (N=50)	21.44 (5.57)	.70 (.19)	.20 (.10)
Experiment 2			
Strategy (N=32)	23.10 (5.62)	.78 (.15)	.17 (.08)
Control (N=33)	21.82 (5.59)	.75 (.13)	.18 (.08)
Experiment 3			
Strategy (N=28)	19.00 (7.55)	.79 (.12)	.16 (.07)
Control (N=29)	18.93 (5.24)	.68 (.23)	.21 (.09)

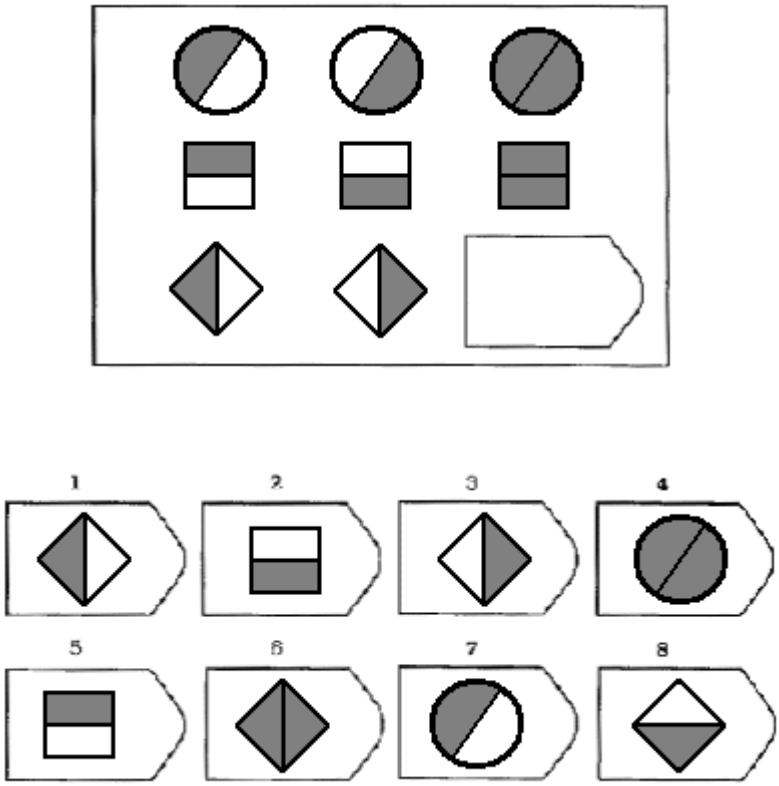

Figure 1. *Sample RAPM item. The participant is asked to select the response option that best completes the pattern both down the columns and across the rows, but not diagonally.*

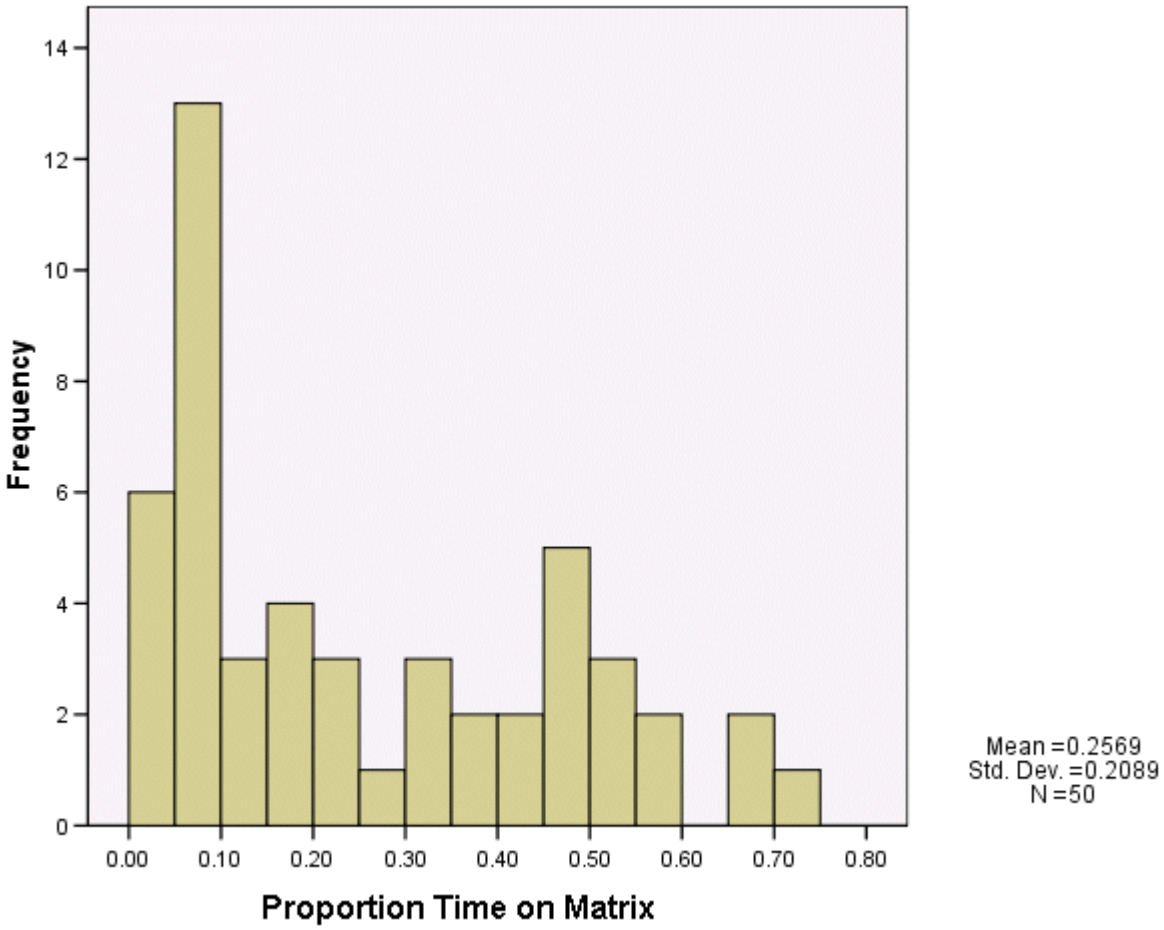

Figure 2. *Distribution of proportion time on matrix for Experiment 1.*

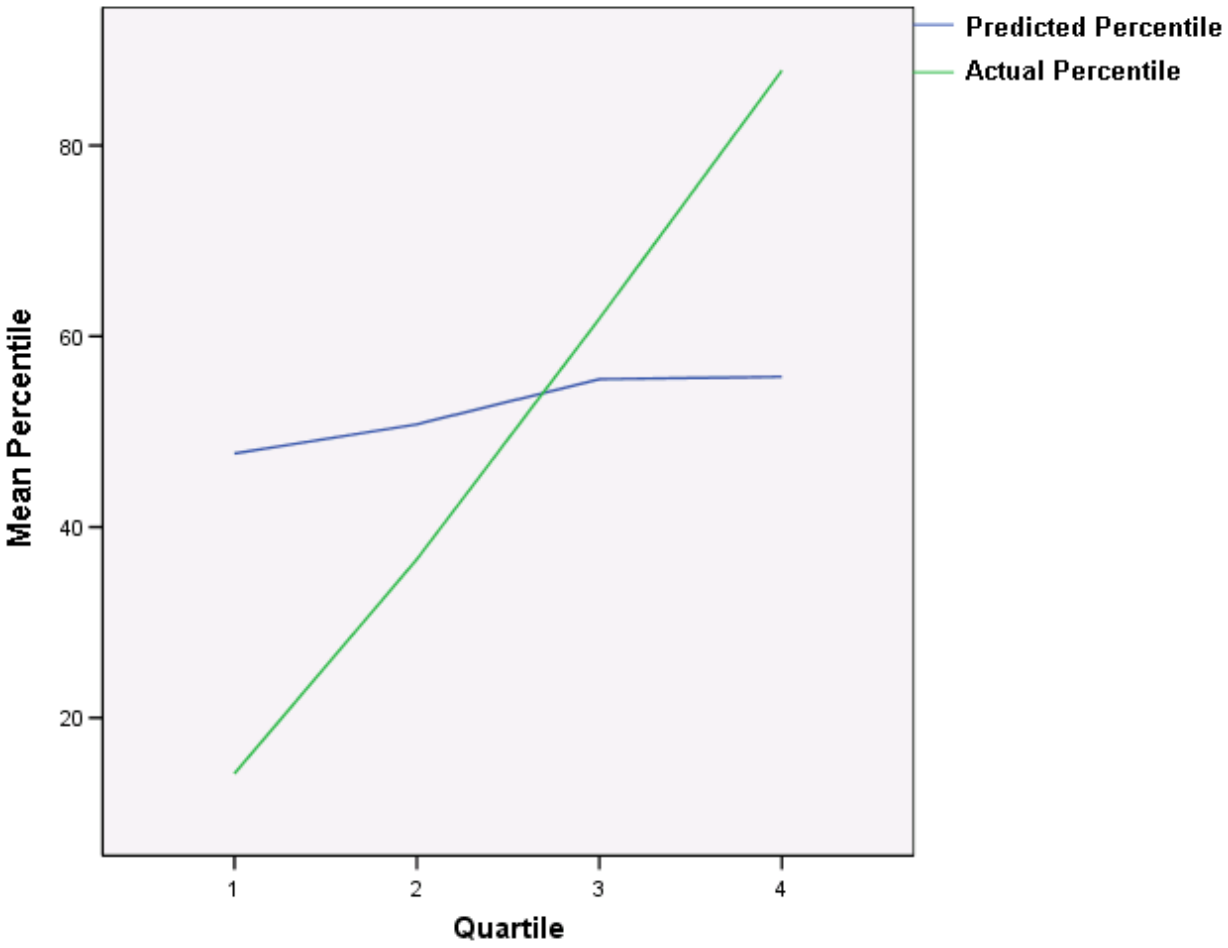

Figure 3. *Graph of mean estimated and actual percentile rank by quartile. The "Unskilled and Unaware" pattern of results was replicated in Experiment 1.*

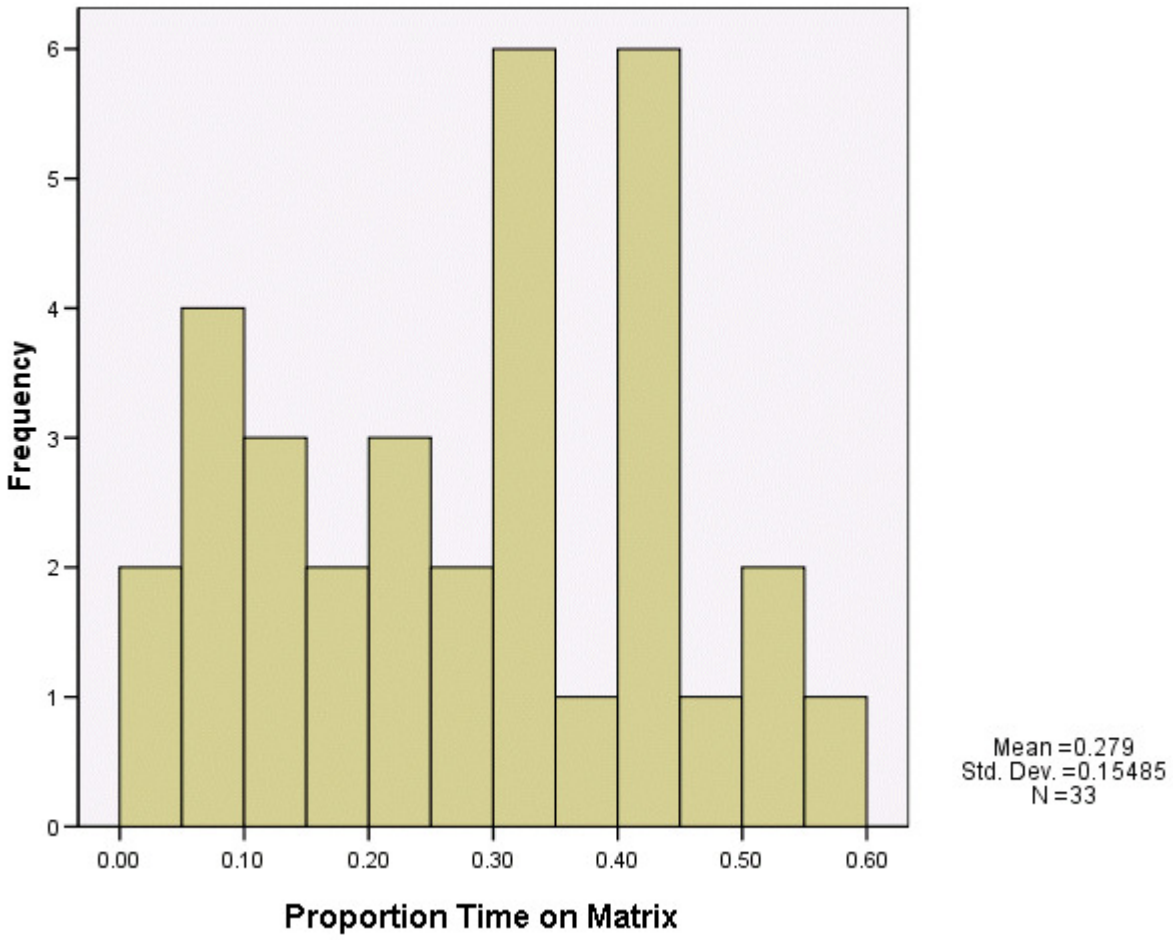

Figure 4. *Distribution of proportion time on matrix for the control group in Experiment 2.*

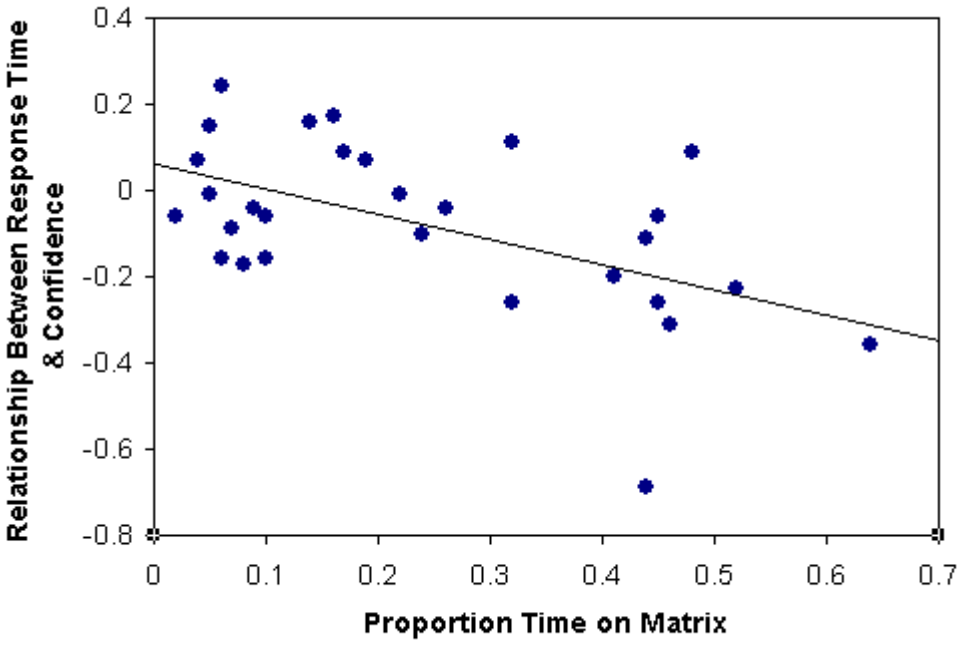

Figure 5. *Scatterplot of cue-utilization by strategy. Participants in the control condition in Experiment 3 who spontaneously used constructive matching were less confident when they took longer to enter responses.*

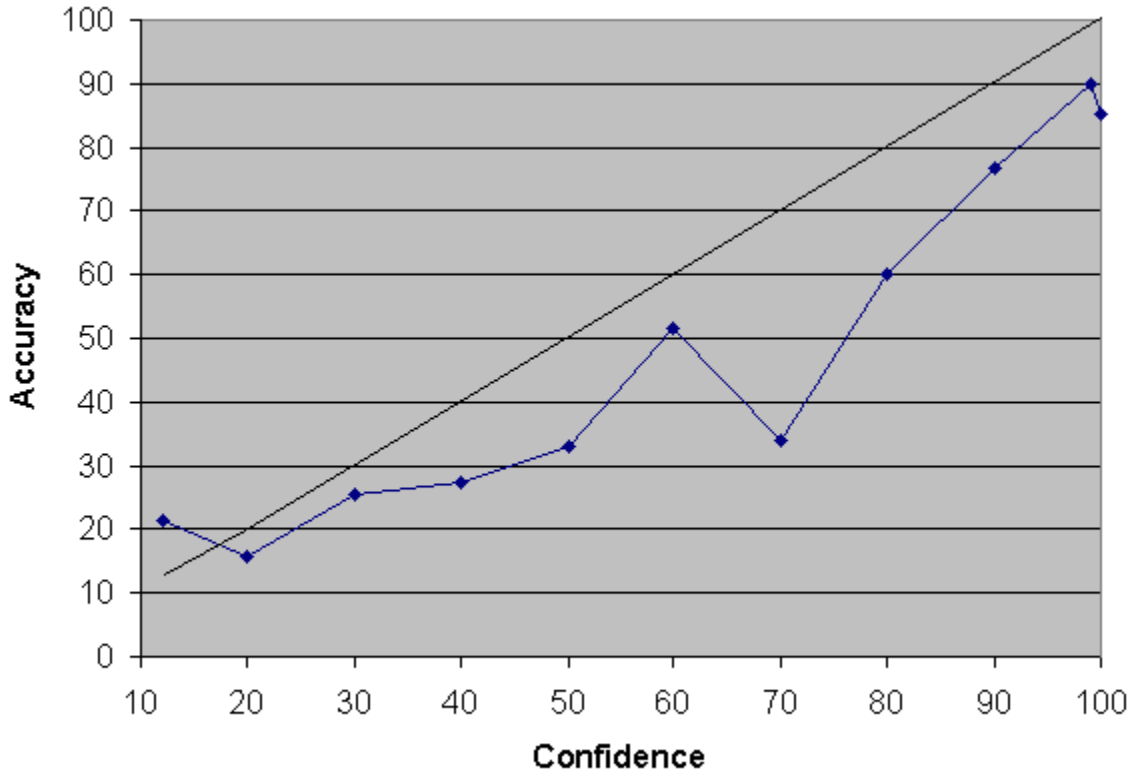

Figure 6. *Calibration plot for the control group in Experiment 3.*

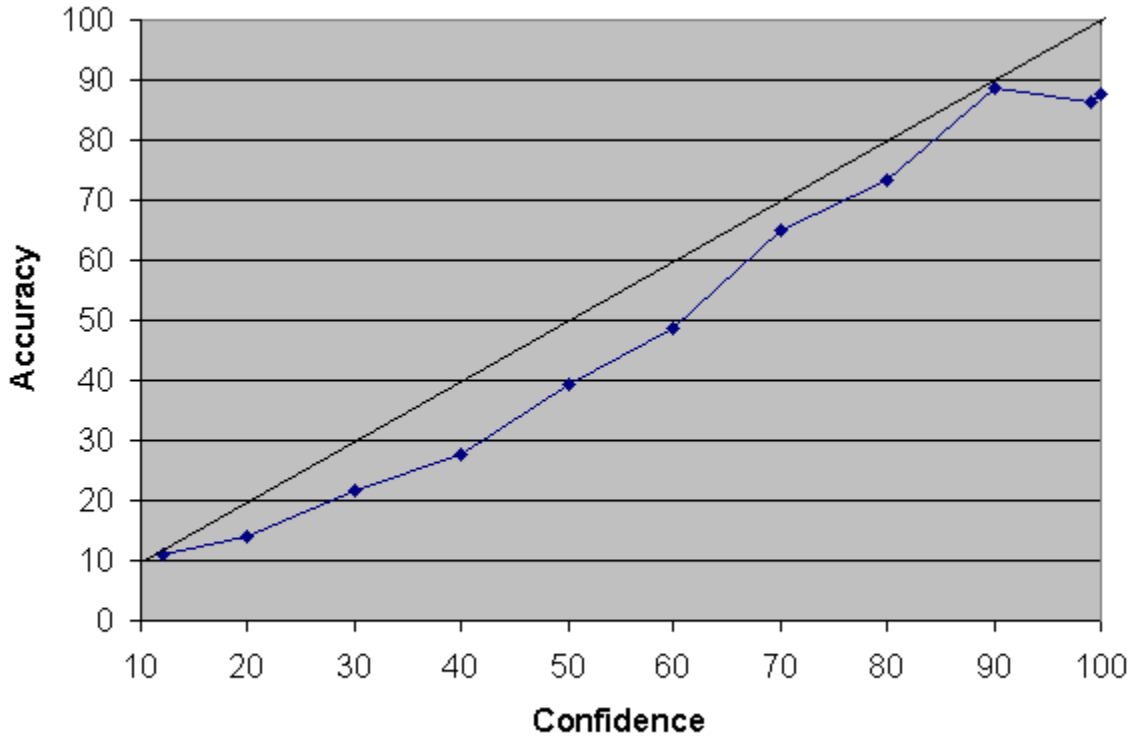

Figure 7. *Calibration plot for the strategy instructed group in Experiment 3.*

APPENDIX: INFORMED CONSENT FORM

Informed Consent Form

For Undergraduate Participants

I freely and voluntarily consent to be a participant in the research project entitled "Metacognition and Inductive Reasoning". The principal investigators will be Ainsley L. Mitchum and Dr. Colleen M. Kelley.

The experiment does not in any way constitute a risk to me. I will receive course credit for this experiment, .5 experimental credits per half hour spent participating. The entire experiment will take approximately 1 hour to complete.

I understand that I will be given tests measuring different cognitive abilities and general knowledge. In addition, I understand that I may be observed during a typical session and that some parts of this session could be audiotaped.

I understand that the records of this research which refer to my data will be given a code so that no one except the investigators and their designated assistants will have access to the data, and that no identifiable data, including handwritten information that I have supplied, will be used for publication. In addition, the records of this research, which refer to my performance, will be kept confidential to the extent allowed by law. I understand that any information, including written records, computer files, and audio tapes used in this project will be retained in the Kelley memory lab suite at the Florida State University Department of Psychology, and that the tapes will be erased or destroyed within ten years (no later than August 31, 2016).

My consent may be withdrawn at any time without prejudice, penalty, or loss of benefits to which I am otherwise entitled. That is, my grade in the course will not be affected if I choose to withdraw from the experiment, nor will I receive an experiment credit penalty. However, I will still be obliged to fulfill my experiment participation obligation for the General Psychology course.

I have been given the right to ask and have answered any inquiry concerning this consent form. Questions, if any, have been answered to my satisfaction. I understand that I may contact Ainsley L. Mitchum, Department of Psychology, Florida State University, Tallahassee, FL 32306, phone: (850) 644-9873, or Dr. Colleen M. Kelley, phone: (850) 644-3816, for answers to pertinent questions about this research.

If I have questions about my rights as a subject/participant in this research, or if I feel that I have been placed at risk, I can contact the Chair of the Human Subjects Committee, Institutional Review Board, through the Office of the Vice President for Research at (850) 644-8633.

I have read and understand this consent form, and I am 18 years or older.

_____ _____

(Participant) (date)

REFERENCES

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33*, 587-605.

Alexander, J. M., Carr, M., & Schwanenflugel, P. J. (1995). Development of metacognition in gifted children: Directions for future research. *Developmental Review, 15*(1), 1-37.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81*(1), 126-131.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology-General, 127*(1), 55-68.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye-movement analysis of geometric analogy performance. *Intelligence, 8*(3), 205-238.

Borkowski, J. G., Carr, M., & Pressley, M. (1987). "Spontaneous" strategy use: Perspectives from metacognitive theory. *Intelligence. Special Issue: A symposium: Why are the mentally retarded strategically deficient?, 11*(1), 61-75.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97*(3), 404-431.

Carr, M., Alexander, J., & Schwanenflugel, P. (1996). Where gifted children do and do not excel on metacognitive tasks. *Roeper Review, 18*(3), 212-217.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence in a modified recognition test. *Memory & Cognition, 22*(3), 273-280.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology-Learning Memory and Cognition, 18*(1), 142-150.

Dunlosky, J. & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology-Learning Memory and Cognition, 27*(5), 1180-1191.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOLs) and the delayed-JOL effect. *Memory & Cognition, 20*(4), 374-380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effect of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*(4), 545-565.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83-87.

Ehrlinger, J. & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*(1), 5-16.

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology, 1*(4), 324-340.

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

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine, 66*(12), 762-769.

Haun, D. E., Zeringue, A., Leach, A., & Foley, A. (2000). Assessing the competence of specimen-processing personnel. *Laboratory Medicine, 31*(11), 633-637.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging, 5*(2), 215-227.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*(1), 22-34.

Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine, 76*(10), S87-S89.

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology Public Policy and Law, 2*(3-4), 447-472.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*(2), 157-175.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1-24.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48*(4), 704-721.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The importance of retrieval practice. *American Journal of Psychology, 93*(2), 329-343.

Knouse, L. E., Paradise, M. J., & Dunlosky, J. (2006). Does ADHD in adults affect the relative accuracy of metamemory judgments? *Journal of Attention Disorders, 10*(2), 160-170.

Koren, D., Seidman, L. J., Goldsmith, M., & Harvey, P. D. (2006). Real-world cognitive - and metacognitive - dysfunction in schizophrenia: A new approach for measuring (and remediating) more "right stuff". *Schizophrenia Bulletin, 32*(2), 310-326.

Koren, D., Seidman, L. J., Poyurovsky, M., Goldsmith, M., Viksman, P., Zichel, S., et al. (2004). The neuropsychological basis of insight in first-episode schizophrenia: a pilot metacognitive study. *Schizophrenia Research, 70*(2-3), 195-202.

Koriat, A. (1993). How do we know that we know: The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609-639.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology-General, 124*(3), 311-333.

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., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(2), 187-194.

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology-Learning Memory and Cognition, 32*(5), 1133-1145.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology-General, 133*(4), 643-656.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490-517.

Koriat, A., Levy-Sadot, R., Edry, E., & de Marcas, S. (2003). What do we know about what we cannot remember? Accessing the semantic attributes of words that cannot be recalled. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*(6), 1095-1105.

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

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*(4), 478-492.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology-General, 135*(1), 36-69.

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.

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.

Kyllonen, P. C., Lohman, D. F., & Woltz, D. J. (1984). Componential modeling of alternative strategies for performing spatial tasks. *Journal of Educational Psychology, 76*(6), 1325-1345.

Lynn, D. J., Holzer, C., & O'Neill, P. (2006). Relationships between self-assessment skills, test performance, and demographic variables in psychiatry residents. *Advances in Health Sciences Education, 11*(1), 51-60.

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.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology-Learning Memory and Cognition, 21*(5), 1263-1274.

McClain, L. (1983). Behavior during examinations: A comparison of A, C, and F students. *Teaching of Psychology, 10*(2), 69-71.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113*(2), 123-132.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology-Learning Memory and Cognition, 19*(4), 851-861.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology-Learning Memory and Cognition, 16*(2), 223-232.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77-101.

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. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology-General, 122*(2), 269-273.

Nelson, T. O., & Dunlosky, J. (1991). When peoples judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The delayed JOL effect. *Psychological Science, 2*(4), 267-270.

Nelson, T. O., Dunlosky, K. J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*(4), 207-213.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the labor-in-vain effect. *Journal of Experimental Psychology-Learning Memory and Cognition, 14*(4), 676-686.

Nelson, T. O. & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing.* (pp. 1-25). Cambridge, MA, US: The MIT Press.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs, 76*(28), 1-21.

Oskamp, S. (1965). Overconfidence in case study judgments. *Journal of Consulting Psychology, 29*(3), 261-265.

Raven, J. C., Court, J. H., Raven, J. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: Advanced Progressive Matrices, Sets I and II (1998 edition).* Oxford: Oxford Psychologists Press.

Rhodes, M. G., & Kelley, C. M. (2005). Executive processes, memory accuracy, and memory monitoring: An aging and individual difference analysis. *Journal of Memory and Language, 52*(4), 578-594.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review, 1*(3), 357-375.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science, 6*(5), 132-137.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 1074-1083.

Snow, R. E. (1980). Aptitude processes. In R.E. Snow, P.A. Federico, & W.E. Montague (Eds.), *Aptitude, learning, and instruction: Cognitive process analyses of aptitude, vol. 1* (27-63). Hillsdale NJ: Erlbaum.

Stankov, L. (1998). Calibration curves, scatterplots, and the distinction between general knowledge and perceptual tasks. *Learning and Individual Differences, 10*(1), 29-50.

Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science, 26*(1-2), 127-140.

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence, 34*(3), 261-272.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*(3), 273-281.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*(1), 41-44.

BIOGRAPHICAL SKETCH

Ainsley Mitchum was born in Charleston, South Carolina on February 17th, 1978. She earned her Bachelor of Science in Psychology from the College of Charleston, in 2000. In 2004, she moved to Tallahassee, Florida to begin work on her Master of Science degree under the direction of Colleen Kelley. Her main research interests are in metacognition, memory, cognitive control, and individual differences.

In May of 2007, Ainsley accepted a position as a pre-doctoral fellow in the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World. Ainsley will be relocating to Berlin, Germany and completing work on her Ph.D. under the direction of Lael Schooler and Joerg Rieskamp at the Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development.

In her very limited free time, Ainsley enjoys playing a variety of classic and contemporary video games. Although her primary interests are in role-playing and adventure games, she has recently become a fan of the rhythm and music genre. Her skills in Guitar Hero and Guitar Hero II are both intimidating and legendary. Ainsley counts herself among the most skilled players in Tallahassee.