THE EFFECTS OF OBSERVABILITY AND EVALUATIVENESS ON METACOGNITIVE SELF- AND OTHER-JUDGMENTS

A Dissertation

by

ROBERT MICHAEL TIRSO

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Heather Lench
Committee Members,	Lisa Geraci
	Steve Smith
	Hart Blanton
Head of Department,	Heather Lench

August 2021

Major Subject: Psychology

ABSTRACT

Metacognition is defined as awareness and beliefs about one's own cognitive processes and abilities. Research on metacognition suggests that the accuracy of metacognitive self- and other-judgments is largely determined by two broad factors: the information available to judges and biases stemming from motivations and desires. The goal of this dissertation was to test whether the Self-Other Knowledge Asymmetry (SOKA) model can explain differences in the accuracy of self- and other-judgments of cognitive abilities. According to the SOKA model, the information available to judges (i.e., observability or how easily outside observers can see a trait) and motivational biases (i.e., evaluativeness or how important a trait is to the judge) together can be used to predict self- and informant-report accuracy, at least for judgments of personality. Working memory, prospective memory, creativity, and visuospatial ability were identified as cognitive abilities that are relatively high or low in their evaluativeness and observability by examining participants' average ratings of importance (Study 1A) and interrater reliability, a measure of observability (Study 1B). The accuracy of participants' and informants' judgments of these abilities was investigated using a multilevel modeling approach. Results were somewhat mixed—observability did not reliably moderate informants' metacognitive judgment accuracy contrary to the model, however evaluativeness did moderate participants' metacognitive judgment accuracy, which is consistent with the SOKA model (Study 2). Attempts to manipulate evaluativeness by extolling the importance (or lack thereof) of creativity had no effect on participants' and informants' metacognitive judgments or their accuracy (Study 3). A novel observability manipulation proved successful at increasing the observability of creativity (Study 4), and will be used in future research to determine if there might be a causal relationship between observability and metacognitive judgment accuracy. Overall, results suggest that evaluativeness does affect metacognitive self-judgment accuracy in a manner consistent with the SOKA model, but additional research is needed to determine if this is a causal relationship and to determine the extent to which observability moderates metacognitive other-judgment accuracy.

ACKNOWLEDGEMENTS

My journey through graduate school has been a long and difficult one, but ultimately rewarding one. It has been a journey that I am certain I never would have finished or possibly even started were it not for my family and the extraordinary people I met along the way. To those of you that are responsible for getting me this far, I would like to take a moment to express my sincere and heartfelt thanks and appreciation.

I would like to thank my committee chair, Dr. Heather Lench, and my committee members, Drs. Lisa Geraci, Steve Smith, and Hart Blanton, for their guidance and support as I completed my dissertation. Lisa, I will always be immeasurably grateful to you for believing in me and accepting me as a graduate student, and for the enormous amount of time and energy you have invested in me as you molded me into the scholar I am today. Heather, I want you to know that I am extremely grateful that you decided to take me in as one of your graduate students when Lisa left A&M. Your generous support and fresh perspective made it possible for me to complete my dissertation and greatly improved the final product. Steve, thank you for always being so open and willing to support me in whatever endeavor I engaged in, including last-minute letter of recommendation requests! And Hart, thank you for convincing me of the value and necessity of learning multilevel modeling. Although your insistence that I use MLM in my dissertation initially frustrated me, I have since come around to appreciating it as a necessary kick-in-the-pants to get me to improve my research toolkit—an improvement that has since opened up many new possibilities for me.

Thanks also go to all of my friends and colleagues, and the department faculty and staff for making my time at Texas A&M University some of the best years of my life. To my fellow graduate student friends, you all are some of the best friends I have ever had and I sincerely hope that we will be able to stay in touch. And, to those of you who have not yet reached this stage of your graduate education, I wish you the very best of luck with your dissertation and your aspirations after graduate school. To the many undergraduate research assistants that have helped me implement my dissertation studies and analyze the resulting data, you have my sincere thanks and appreciation for your help. It has been a pleasure working with such talented students, and I am confident that you all will go on to achieve great things.

Finally, thanks go to my family, especially my mother, for all of your love and support. I know these last few years have not been easy for you, yet despite everything that has happened you continued to always be there for me whenever I needed you. Know that I am entirely serious when I say that you are the best parent anyone could ask for, and that I would not be where I am today were it not for your tireless efforts to give Cynthia, Kimberly, and me the best possible future.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Lisa Geraci of the Psychology Department at the University of Massachusetts Lowell, Drs. Heather Lench and Steve Smith of the Department of Psychological and Brain Sciences at Texas A&M University, and Dr. Hart Blanton of the Department of Communication at Texas A&M University.

Drs. Lisa Geraci and Heather Lench contributed tremendously to the framing and theoretical interpretations in this dissertation, and provided feedback at various stages of the writing of this dissertation. Dr. Hart Blanton guided me to an appropriate analytical strategy for Studies 2 and 3. Undergraduate students Sarah Way, Katie Gray, Lizbeth Okumura, Denise Ortiz, Ariana Kozitza, Ashley Riggins, and Raena Slate helped prepare materials that were used in the studies included in this dissertation; Katie Gray, Elise Yellitz, Li Wen Jan, Harrison Gibbs, Elizabeth Glass, and Cherri Yang coded participants' responses on the Alternate Uses Task used in Studies 2 and 3. All other work conducted for this dissertation was completed by Robert Tirso independently.

Funding Sources

Graduate study was supported by a fellowship and teaching positions from Texas A&M University.

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGEMENTS iv

CONTRIBUTORS AND FUNDING SOURCES vi

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES x

CHAPTER I INTRODUCTION 1

An Introduction to Metacognitive Monitoring 2
The Origins of Metacognition 2
The Benefits of Metacognitive Monitoring Accuracy 4
Sources of Metacognitive Bias 7
Self- and Other-judgments of Cognitive Ability 10
Differences in Information Between Selves and Others 11
Differences in Motivation Between Selves and Others 13
A Framework for Self- and Other-judgments 15
The Self-Other Knowledge Asymmetry Model 15
Metacognition and the SOKA Model 19
The Goal of the Current Studies 21
Research Question 1: Which Cognitive Abilities are Relatively High or Low in Evaluativeness and Observability? 21
Research Question 2: What Effect Does Evaluativeness Have on the Accuracy of Metacognitive Self- and Other-judgments? 22
Research Question 3: What Effect Does Observability Have on the Accuracy of Metacognitive Self- and Other-judgments? 24

CHAPTER II STUDIES 1A AND 1B: DETERMINING THE EVALUATIVENESS AND OBSERVABILITY OF COGNITIVE ABILITIES 26

Study 1A 26
Method 26
Results 29

Study 1B..31
Method..31
Results 34

CHAPTER III STUDY 2: TESTING THE SOKA MODEL..36

Study 2..36
Method..36
Results 47

CHAPTER IV STUDIES 3 – 4B: MANIPULATING EVALUATIVENESS AND OBSERVABILITY 54

Study 3..54
Method..55
Results 59
Study 4..66
Method..68
Results 71

CHAPTER V DISCUSSION 73

Major Findings and Implications 73
Limitations and Future Directions..79
Conclusion..83

REFERENCES..85

APPENDIX A EXAMPLE ITEM FROM STUDY 1A 99

APPENDIX B EXAMPLE ITEM FROM STUDY 1B 100

APPENDIX C HIGH EVALUATIVENESS ARTICLE FROM STUDY 3.................101

APPENDIX D LOW EVALUATIVENESS ARTICLE FROM STUDY 3 102

APPENDIX E TABLES..103

APPENDIX F FIGURES 113

LIST OF FIGURES

Page

Figure 1. The effect of evaluativeness on participants’ judgment resolution in Study 2. 113

Figure 2. The effects of observability on participants' judgment resolution in Study 2. 114

Figure 3. The effects of evaluativeness on informants' judgment resolution in Study 2. 115

Figure 4. The effects of observability on informants' judgment resolution in Study 2.. 116

Figure 5. The effects of judgment scale on informants’ judgment resolution for their low evaluativeness judgments in Study 3. 117

Figure 6. The effects of judgment scale on informants’ judgment resolution for their high evaluativeness judgments in Study 3. 118

LIST OF TABLES

Page

Table 1. Mean (SD) evaluativeness ratings, Krippendorff’s alpha, lower and upper bounds of the 95% CI around alpha, mean (SD) observability ratings, and mean (SD) funding allocations from Studies 1 and 2. Items are displayed from largest to smallest alpha. 103

Table 2. Mean (SD) performance and predictions from Study 2. 104

Table 3. Correlation matrix of participants' raw performance on the experimental tasks and participants' and informants' Likert scale judgments from Study 2. Bold indicates a correlation between a judgment and actual performance. ...105

Table 4. Correlation matrix of participants' percentile ranks on the experimental tasks and participants' and informants' percentile rank judgments from Study 2. Bold indicates a correlation between a judgment and actual performance. ...106

Table 5. Model 1.x results from Study 2. Standard errors are in parentheses. Asterisks indicate the significance level of effects and improvement in overall model fit compared to the preceding model. 107

Table 6. Mean (SD) predictions, performance, evaluativeness ratings, and manipulation check items from Study 3. 108

Table 7. Correlation matrix of participants' actual performance (fluency) on the AUT and participants' and informants' general ability judgments from Study 3. Bold indicates a correlation between a judgment and actual performance. ...109

Table 8. Correlation matrix of participants' actual percentile ranks on the AUT and participants' and informants' percentile rank judgments from Study 3. Bold indicates a correlation between a judgment and actual performance. 110

Table 9. Model 2.x results from Study 3. Standard errors are in parentheses. Asterisks indicate the significance level of effects and improvement in overall model fit compared to the preceding model. 111

Table 10. Krippendorff’s alpha and the lower and upper bounds of the 95% CI around alpha for all items from Studies 1B and 4. Items are displayed from largest to smallest alpha based on values originally calculated from Study 1B.........112

CHAPTER I

INTRODUCTION

Both the capacity for cognition and the capacity to monitor and reflect upon cognition are core components of the human experience. The ability to monitor and reflect upon one's own cognitive processes and abilities is referred to as metacognition (Dunlosky, Serra, & Baker, 2007; Flavell, 1979; Schraw & Dennison, 1994). Since its inception, the primary focus of research on metacognition has been metacognition as it pertains to the self—in other words, how do people monitor their own cognitive abilities and learning, and what factors affect the accuracy of their metacognitive self-judgments? More recently though, some attention (Helzer & Dunning, 2012; Koriat & Ackerman, 2010; Miller & Geraci, 2016; Tirso & Geraci, 2020; Tirso, Geraci, & Saenz, 2019) has shifted to metacognition as it pertains to others—for example, how does the accuracy of metacognitive judgments made about another individual (metacognitive other-judgments) compare to the accuracy of metacognitive self-judgments, and what factors affect the accuracy of metacognitive other-judgments? It is important to investigate metacognitive other-judgments in addition to self-judgments for a variety of reasons. For instance, metacognitive other-judgments play an invaluable role in education—instructors must be able to gauge their students' existing knowledge as well as the progress of ongoing learning to deliver high quality instruction. Knowledge of others' abilities (other-knowledge) also informs knowledge of one's own abilities (self-knowledge), as students no doubt compare themselves to their peers as a means of determining whether they themselves are performing relatively well or relatively poorly. Judgments about others' cognition no doubt play a significant role in a variety of important decisions too, such as when a hiring committee evaluates applicants for a position, or when a jury must decide whether an eyewitness's testimony is reliable or not. However, research on metacognitive other-judgments has produced mixed results: sometimes metacognitive other-judgments are more accurate than self-judgments (Helzer & Dunning, 2012), but sometimes the reverse is true (Koriat & Ackerman, 2010; Miller & Geraci, 2016; Tirso & Geraci, 2020; Tirso et al., 2019). Research from other areas of psychology has focused on the relationship between self- and other-judgments and attempted to explain the factors that affect them. In particular, the personality literature has found that, like in the metacognitive literature, sometimes other-judgments are more accurate than self-judgments, but sometimes the reverse is true. In the personality literature, these mixed results can be attributed to the influence of two factors on judgments: evaluativeness and observability (Beer & Vazire, 2017; Vazire, 2010; Vazire & Mehl, 2008). The goal of this dissertation was to determine the extent to which evaluativeness (how desirable or important a trait is to an individual—a motivational factor) and observability (how visible a trait is to outside observers—an informational factor) influence the relationship between metacognitive self- and other-judgments and whether their effects on judgment accuracy are consistent with those predicted by the Self-Other Knowledge Asymmetry (SOKA) model (Vazire, 2010).

An Introduction to Metacognitive Monitoring

The Origins of Metacognition

Before discussing the value of investigating the effects of observability and evaluativeness on metacognitive self- and other-judgments further, it is crucial to first discuss what metacognitive self- and other-judgments are and why one should care about them. As mentioned, both the capacity for cognition and the capacity to monitor and reflect upon cognition are core components of the human experience. People have been reflecting upon their own cognition (i.e., metacognition) since long before empirical studies of metacognition began in earnest. For example, one common mnemonic technique, the method of loci, can be traced back to roughly 2,500 years ago and provides an example of an early attempt at theorizing about memory and metacognition. As the tale goes, the ancient Greek poet Simonides was attending a banquet when suddenly the banquet hall's roof collapsed and killed many of the guests. Being one of the survivors, Simonides was asked to help identify the bodies. In the process, Simonides found that using mental imagery made it easier to remember who was in attendance that night. He subsequently hypothesized that to improve memory one should store to-be-remembered items as mental images within a mental scene of a location, as doing so meant that simply recalling the mental scene would also recall the items that needed to be remembered. Thus, the method of loci was born (Thomas, 2018). This is an example of the two basic processes that make up metacognition: monitoring and control. Simonides recognized that memorizing a list of items could be made easier by using mental images to represent those items within a mental representation of a location (an example of metacognitive monitoring). He then refined and shared this technique so that he and others could use it to improve their memory capacities in the future (an example of metacognitive control).

Although people have considered metacognition for millennia, John Flavell’s (1979) seminal paper on the topic marks a point at which empirical research on metacognition began in earnest. This is not to say that no research on metacognition existed prior to 1979—works such as Arbuckle and Cuddy (1969), Hart (1965), and Underwood (1966) certainly disprove that notion. Rather, the work by Flavell (1979) served as a catalyst that promoted the idea of metacognitive phenomena as an area of inquiry and spurred interest in metacognitive research. In his paper, Flavell noted that there are differences in children’s knowledge of cognition, or metacognition as he called it, with older children typically being better at monitoring their own memory and comprehension than younger children. This observation led to a profound question: to what extent can metacognition be developed? Clearly it was improving among children as they aged, so what might fully-developed metacognition look like? Flavell proposed a model of metacognition that consisted of four components: (1) metacognitive knowledge that consisted of knowledge of how cognitive processes might be influenced by various factors; (2) metacognitive experiences that consisted of the subjective experiences or “feedback” accompanying cognitive processes; (3) the goals set for a cognitive process; and (4) the strategies used to reach a set goal. Subsequent work has since reduced this model to two main components: metacognitive monitoring—the focus of the current paper—and metacognitive control.

The Benefits of Metacognitive Monitoring Accuracy

A common approach to studying metacognitive monitoring is to compare metacognitive monitoring judgments to some cognitive measure to determine a subject's metacognitive accuracy. Metacognitive monitoring judgments, or metacognitive judgments for short, come in a variety of forms such as the judgment of learning (JOL), judgment of remembering and knowing (JORK), feeling-of-knowing (FOK) judgment, ease-of-learning (EOL) judgment, and predictions of performance on cognitive tasks. What all metacognitive judgments share is that they require people to reflect on their cognitive processes—for instance, estimating the likelihood that one will remember a paired associate later, or monitoring the contents of memory to estimate how well one is likely to perform on an upcoming final exam. These judgments can then be compared to performance, such as whether or not the participant actually remembered a paired associate, or how close a student's grade prediction was to his or her actual grade on an exam.

Studies using this approach to examine metacognition have found that better metacognitive accuracy is desirable because it is associated with better performance across a variety of laboratory and classroom tests (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Hartwig & Dunlosky, 2014; Kruger & Dunning, 1999; Miller & Geraci, 2011a, 2011b; Thiede, 1999; Thiede, Anderson, & Therriault, 2003; Tirso & Geraci, 2020). Common sense would also suggest that better metacognitive accuracy leads to better performance because it allows individuals to recognize when performance is in need of improvement (but see also Koriat, Ma'ayan, & Nussinson, 2006). Consistent with this view, evidence suggests that metacognitive monitoring influences performance on cognitive tasks through the regulation of study time (i.e., metacognitive control and the monitoring-affects-control hypothesis; Dunlosky & Ariel, 2011; Metcalfe & Finn, 2008; Nelson, Dunlosky, Graf, & Narens, 1994; Thiede, 1999; Thiede et al., 2003). For instance, allowing participants to self-pace their studying in paired-associates learning tasks results in better performance than fixed-pace studying for the same amount of time, which suggests that the proper use of metacognitive monitoring leads to better learning and performance (Koriat et al., 2006). Other studies have reported differences in performance after manipulating metacognitive accuracy (Dunlosky & Rawson, 2012; Thiede et al., 2003). In the Thiede et al. study, participants rated their comprehension after reading several passages, took a comprehension test, then selected which passages to reread before taking a final comprehension test. Metacognitive accuracy was manipulated by having participants write keywords to summarize each passage during the initial reading phase (1) after a delay for maximum accuracy (see Nelson & Dunlosky, 1991), (2) immediately after reading each passage for reduced accuracy, or (3) not at all. Participants then provided comprehension ratings for each passage. Participants in the delayed keywords condition provided more accurate comprehension ratings than participants in the immediate and no keywords conditions after the initial reading and testing phase, indicating the metacognitive accuracy manipulation was successful. Participants in the delayed keywords condition were also more likely to select poorly understood texts for restudy and performed better on the final comprehension test than participants in the immediate and no keywords conditions despite participants in all conditions spending a similar amount of time restudying. Thus, the data demonstrate a causal chain between metacognitive accuracy, regulation of study, and performance (Thiede et al., 2003). In other words, there is evidence that better metacognitive monitoring can improve performance through better regulation of study time.

Sources of Metacognitive Bias

Given the benefits of accurate metacognition, it is unfortunate that a large body of work demonstrates that people often overestimate their cognitive abilities and knowledge. For example, people overestimate how well they will perform on logical reasoning tasks (Hartwig & Dunlosky, 2014, Study 4; Kruger & Dunning, 1999; Tirso & Geraci, 2020), their grammatical knowledge (Krueger & Mueller, 2002; Kruger & Dunning, 1999; Tirso & Geraci, 2020), how well they have learned paired associates (Metcalfe & Finn, 2008; Miller & Geraci, 2014, 2016; Thiede, 1999), their performance on practice GRE tests (Tirso & Geraci, 2021), and classroom tests spanning topics such as psychology (Al-Harthy, Was, & Hassan, 2015; Foster, Was, Dunlosky, & Isaacson, 2017; Hartwig & Dunlosky, 2014; Miller & Geraci, 2011a; Saenz et al., 2017; Serra & DeMarree, 2016; Tirso et al., 2021), and statistics (Hartwig & Dunlosky, 2017).

A number of factors may contribute to the pervasiveness of overconfidence in performance. In classroom settings, students’ study strategies may contribute to widespread overconfidence. Although students understand that studying leads to better memory and performance, they appear to have a poor understanding of good and bad studying habits. Many students report that they use rereading or repetition and self-testing to prepare for their exams (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007).

However, only about 11% of students recognize self-testing as a means of improving memory (Karpicke, Butler, & Roediger, 2009) despite its value as a learning tool (Roediger & Karpicke, 2006a; Roediger & Karpicke, 2006b). Instead, students appear to view testing as a form of assessment—a means of figuring out what material to reread, which is a relatively ineffective method of study (Callender & McDaniel, 2009)—rather than a strategy that improves memory in and of itself (Kornell & Bjork, 2007). Further complicating matters, although students recognize that spaced study leads to better performance than massed study or "cramming", most students choose massed study over spaced study (Wissman, Rawson, & Pyc, 2012). Unawareness and avoidance of good study habits could partially explain the prevalence of overconfidence in the classroom, as students could be basing their overconfident predictions on having spent a lot of time studying for their exams, not realizing that these efforts were not particularly effective.

Poor knowledge of what cues signal that something has been learned may also play a role in producing overconfidence. Evidence from laboratory studies using JOLs suggests that people often base JOLs and predictions of future performance off of misleading information or cues. It is well documented that fluency—how quickly an item is perceived, encoded, or retrieved from memory—is related to participants' predictions of future memory performance, even though such information is not diagnostic of future recall (see Bjork, Dunlosky, & Kornell, 2013 for a review). For example, participants believe they are more likely to remember words printed in large font than words printed in small font even though future recall is unaffected by font size (Kornell, Rhodes, Castel, & Tauber, 2011; McDonough & Gallo, 2011; Rhodes &

Castel, 2008). Participants also believe that they are more likely to remember the answers to questions after being primed with words present in the question, an effect that has been attributed to the primes facilitating perceptual processing and thus increasing fluency (Reder, 1987). There is some discussion about whether it is the subjective experience of fluency (see Kornell et al., 2011) or beliefs about the relationship between fluency and memory performance (e.g., that words in large font would be easier to remember than words in small font; see Mueller, Dunlosky, Tauber, & Rhodes, 2014) that causes metacognitive errors. But, regardless of the exact mechanism, fluency-related cues clearly exert substantial influence over metacognitive judgments and can often lead to overconfident judgments.

Another source of bias that contributes to overconfident metacognitive judgments is a desire to perform well, especially in high stakes situations such as in the classroom. In one study, students were asked to predict their performance before taking their exams as well as their desired grade, the lowest grade they would be happy with, and the extent to which they based their predictions on their study habits, their performance on a previous exam, and/or their attendance. Results showed that motivational factors—students' desired and lowest accpetable grades—were much stronger predictors of their grade predictions than academic factors traditionally assumed to influence predictions and performance, such as study habits, prior performance, and attendance (Saenz et al., 2017; see also Serr & DeMarree, 2016). This pattern persisted across multiple exams and suggests that, whether they realize it or not, students might be predicting the grades they want to receive rather than the grades they think they will receive. Laboratory research using JOLs has found a similar result: the level of performance participants want to achieve increased their JOLs but does not their performance, leading to overconfidence (Ikeda, Yue, Murayama, & Castel, 2016; Soderstrom & McCabe, 2011).

In short, despite the benefits of accurate metacognition, people often exhibit poor metacognitive accuracy. This can occur when people lack the information they need to make an accurate metacognitive judgment, such as when students mistakenly believe they are well-prepared for an exam after having spent many hours engaging in ineffective studying strategies, or when they rely on misleading cues such as how easily something comes to mind. Metacognitive errors also arise whenever people engage in some form of motivated reasoning, such as when students base their grade predictions on their desired grades rather than their past grades.

Self- and Other-judgments of Cognitive Ability

Whereas much work has examined the accuracy of self-judgments of cognitive ability, it is also informative to examine the accuracy of judgments of others' cognitive abilities (other-judgments) and how they compare to self-judgments, as knowledge of others' abilities informs knowledge of one's own abilities. To illustrate, consider the following example. A college student might know that she received scores of 157 and 160 on the verbal and quantitative portions of the GRE, and that these scores are above the minimum cutoff for the graduate program to which she is applying. However, she would have a more complete understanding of her abilities and competitiveness as an applicant if she also possessed good other-knowledge, such as knowing how well others performed on the GRE. In this case, she performed better than 76 percent of GRE testtakers (Educational Testing Services, 2017). Thus, this student would know not only that she performed above some threshold, but that she performed better than a majority of her peers.

Differences in Information Between Selves and Others

How do metacognitive other-judgments compare to metacognitive self-judgments? Obviously, others do not have the same intimate access that selves do regarding their own thoughts, cognitive processes, and knowledge. This deprives people of potentially useful information when assessing others' cognitive abilities, such as idiosyncratic mnemonic information or cues stemming from people's own experiences during study (Koriat, 1997; Koriat & Ackerman, 2010; Matvey, Dunlosky, & Guttentag, 2001; Miller & Geraci, 2016; Mueller et al., 2014; Tullis & Fraundorf, 2017; see also Kelley & Jacoby, 1996). It is logical to assume that, all else being equal, having access to less information may adversely influence other-judgment accuracy. There is some evidence to support this assumption. For instance, when studying word-pairs for a memory test, participants' JOLs are inversely related to study time—participants correctly judge that they are less likely to recall items studied for longer during testing, presumably because the longer study time indicates the item was hard to learn (Koriat et al., 2006; Thiede, 1999). However, when participants made JOLs for another person whom they watched study the same list of word pairs, their JOLs decreased with increasing study time instead of increasing, a pattern that typically results in poor accuracy on such a task unless there are differing levels of incentive to remember items (Koriat et al., 2006). This pattern was reversed—participants made higher JOLs for another student for items studied for longer—when participants had firsthand experience with the task prior to making JOLs for others (Koriat & Ackerman, 2010). In other words, participants made inaccurate metacognitive other-judgments when they lacked anything comparable to the mnemonic information individuals had while studying.

Although Koriat and Ackerman (2010) demonstrated the role mnemonic information can play in the accuracy of metacognitive other-judgments, due to the design and goals of their experiments it was unclear whether metacognitive other-judgments were inaccurate because they overestimated others' abilities or because they underestimated others' abilities. The direction of error (i.e., over- or underestimation) in metacognitive other-judgments is an important consideration. To illustrate, consider the earlier example of a college student trying to assess her abilities and competitiveness for graduate school compared to her peers. Similar to how students might underprepare for an exam if they are overconfident about their preparedness, if our hypothetical graduate school applicant underestimates her peers' qualifications relative to her own then she may not adequately prepare herself for graduate school, or she might apply to programs for which she is not competitive. On the other hand, if she overestimates her peers then she might overprepare for graduate school or apply to lower-tier programs for which she is overqualified for, which entails its own opportunity cost but is arguably a more desirable outcome than not getting accepted into a graduate program; thus, the direction of error matters.

Unfortunately, there are only a few studies that have investigated the direction of errors in metacognitive other-judgments. In one series of studies (Miller & Geraci, 2016), “judges” studied a list of Lithuanian-English paired associates, then read about a separate “learners” condition in which people studied the same list, completed several practice items, and were then tested over the full list of items. The judges were given information about learners’ performance on the practice items (i.e., which items they got right), and were asked to predict how well the learner would perform on the final test. Judges’ predictions of the learners’ performance were more overconfident and less accurate than learners’ own predictions. This finding was attributed to differences in mnemonic information, as judges did not have access to learners’ thoughts and experiences during the retrieval practice because they did not participate in retrieval practice themselves—they merely read about it (Miller & Geraci, 2016). In another series of studies, Tirso and Geraci (2020) asked participants to predict their own performance and either another classmate’s, a college friend’s, a stranger’s, or a close friend or family member’s performance on various cognitive tasks. Across each of these studies, performance predictions for others (other-predictions) were more optimistic and less accurate than self-predictions. This pattern occurred across a variety of cognitive tasks and could not be explained by the nature of the relationship between selves and others or by optimism in self-predictions dropping just prior to testing while optimism in other-predictions remained fixed (but see Shepperd, Ouellette, & Fernandez, 1996). In short, there is evidence that metacognitive other-judgments are less accurate and more inflated than self-judgments, and that this is a replicable and robust finding.

Differences in Motivation Between Selves and Others

Although there is evidence that metacognitive other-judgments are less accurate than self-judgments due to differences in the information available to judges, there are some occasions in which metacognitive other-judgments are more accurate than self-judgments. According to self-enhancement theory, people are motivated to protect and enhance their self-image because doing so leads to better self-esteem and well-being (Brown, 2012; Sedikides & Gregg, 2008; Sedikides & Strube, 1997), especially when the trait or ability in question is important to them (Ludeke, Weisberg, & Deyoung, 2013; Sedikides, Gaertner, & Toguchi, 2003). Consistent with a self-enhancement theory perspective, research has shown that metacognitive self-judgments are biased towards overconfidence by personal motivations, such as the desire for high performance (Saenz et al., 2017; Serra & DeMarree, 2016). However, metacognitive other-judgments do not appear to be so easily swayed by motivational factors. A study by Helzer and Dunning (2012) reported that students prioritize motivational information, such as their desired grade, more than diagnostic information, such as their grade on a previous exam, when predicting their own performance on an upcoming exam. However, their priorities changed when asked to predict another student's grade on the same exam—they believed the other student's performance on the previous exam was more important for making an accurate prediction than the other student's desired grade. This reduction in motivational biases among metacognitive other-judgments also led to greater accuracy—students were more accurate at predicting others' performance on the exam than their own. Thus, despite differences in the information available to selves and others, metacognitive other-judgments do enjoy some advantages over self-judgments, namely a reduction in motivational biases that can lead to greater accuracy, but this does not appear to happen often (c.f. Miller & Geraci, 2016; Tirso & Geraci, 2020).

A Framework for Self- and Other-judgments

To summarize, both informational and motivational factors influence the accuracy of metacognitive self- and other-judgments. But, why are metacognitive other-judgment sometimes more accurate than metacognitive self-judgments (e.g., Helzer & Dunning, 2012), and sometimes less accurate than self-judgments (e.g., Koriat & Ackerman, 2010; Miller & Geraci, 2016; Tirso & Geraci, 2020)? Despite the obvious similarities between metacognitive self- and other-judgments, there is currently no theoretical framework that attempts to predict and explain the accuracy of *both* metacognitive self- and other-judgments that could be used to understand why their accuracy varies. Indeed, it has recently been highlighted that one challenge the field of metacognition faces is the continued lack of a unified definition of metacognition. Instead, a plethora of various definitions of metacognition propagate the field, each consisting of different components and attempting to explain different metacognitive phenomena (Azevedo, 2020; see also Dunlosky & Rawson, 2019; Panadero, 2017; Schunk & Greene, 2018). The current paper does not attempt to provide such a unified definition, but it does attempt to push the field closer to one by testing a framework that could account for the accuracy of both metacognitive self- and other-judgments.

The Self-Other Knowledge Asymmetry Model

One potential explanation behind the mixed pattern of results seen when comparing metacognitive self- and other-judgment accuracy is that informational and motivational factors affect self- and other-judgments differently, and this leads to situations in which self-judgments are more accurate than other-judgments and to situations in which the reverse is true. This idea takes inspiration from Vazire's (2010) Self-Other Knowledge Asymmetry (SOKA) model. The SOKA model originated from the Realistic Accuracy Model (Funder, 1995) in personality assessment research. The Realistic Accuracy Model specified that accurate judgment of another's personality requires that (1) personality traits produce behaviors, (2) these behaviors are phenomena that can be observed by an outside judge, (3) the judge takes note of these behaviors, and (4) the information gained from observing behavior is properly used when making a judgment. The SOKA model might be interpreted as a simplification of this model, which also captures how motivations can influence these processes by collapsing these factors into two dimensions: a trait's observability (or visibility to outside observers) and a trait's evaluativeness (or how desirable it is to possess). The SOKA model also goes beyond work on the Realistic Accuracy Model by predicting not only when other-judgments of a trait will be accurate or inaccurate, but when they will be more, less, or as accurate as self-judgments (Vazire, 2010).

According to the SOKA model, selves have unparalleled access to information regarding their own traits, thoughts, and abilities, whereas outside observers do not. As a result, the observability of a trait primarily influences the accuracy of other-judgments of that trait. If a trait is highly observable, then both selves and others will have access to the information needed to make an accurate judgment of that trait. However, if a trait is low in observability, then others will lack the information needed to make an accurate judgment unless they know the target participant very well (Vazire, 2010). Thus, higher observability generally leads to increases in other-judgment accuracy but may be unnecessary if the self and other are well acquainted. With regard to a trait's evaluativeness—the second critical factor in determining other-judgment accuracy—motivational biases are said to lead people to inflate self-judgments, but not other-judgments, of highly evaluative traits because people wish to possess high levels of these traits; this is consistent with findings from the self-enhancement theory literature (Brown, 2012; Sedikides & Gregg, 2008; Sedikides & Strube, 1997; Taylor & Brown, 1988). Thus, according to the SOKA model, other-judgments will be more accurate than self-judgments of highly evaluative traits (Vazire, 2010).

Preliminary tests of the SOKA model have supported these predictions. Vazire (2010) compared the accuracy of self- and other-judgments for three traits that differed in their observability and evaluativeness: neuroticism (low observability, low evaluativeness), extraversion (high observability, low evaluativeness), and openness/intellect (low observability, high evaluativeness). Self- and other-judgments were equally accurate at predicting extraversion-related behaviors, self-judgments were more accurate than other-judgments at predicting neuroticism-related behaviors, and other-judgments from close others exhibited the strongest correlations with selves' performance on IQ and creativity tests—all outcomes predicted by the SOKA model. Vazire (2010) did not include a trait that was high in both evaluativeness and observability when testing her model, however we might infer how the results would have turned out based on another study examining perceptions of dating appeal. In their study, Preuss and Alicke (2009) found that participants provided more inflated judgments of their dating appeal than did yoked observers. Based on the biological and societal pressures to find a mate and the relationship between physical attractiveness and dating appeal, it is reasonable to assume that dating appeal is relatively high in evaluativeness and observability. If this is indeed the case, then this outcome is also consistent with the SOKA model—self-judgments of dating appeal were inflated due to high evaluativeness, and other-judgments were more accurate because of dating appeal's relatively high observability. It should be noted that determining the accuracy of self- and other-judgments in Preuss and Alicke's (2009) study is difficult due to the lack of an objective criterion from which to derive accuracy. Nevertheless, Preuss and Alicke's (2009) findings appear to be consistent with the SOKA model.

Vazire's (2010) findings and the SOKA model were also supported by a series of meta-analyses published in the same year (Connelly & Ones, 2010). Upon aggregating 1,510 interrater reliability coefficients from 114 different samples, Connelly and Ones (2010) found that, among the Big Five personality factors (openness, conscientiousness, extraversion, agreeableness, neuroticism/emotional stability), interrater reliability among assessments from others was highest for extraversion followed by conscientiousness. This was interpreted as evidence that extraversion and conscientiousness were in fact highly observable traits, as others' assessments of these traits tended to agree regarding what they "saw" in the target. In contrast, the interrater reliability of other-judgments was lower for agreeableness and especially openness and emotional stability, indicating that these traits were not as easily judged by others and thus are low in observability.

Connelly and Ones (2010) also found that interrater reliability was higher among family and friends than roommates, coworkers, and incidental acquaintances across all five personality factors, but that this difference was smallest for extraversion. In other words, traits considered to be highly observable were in fact highly observable as evidenced by greater interrater reliability—raters agreed on what they saw in the target—and the SOKA model's prediction that close others' assessments would be less affected by low trait observability was supported. Finally, Beer and Vazire (2017) have also replicated most of Vazire's (2010) original findings using naturalistic observation instead of laboratory assessments to index participants' personalities.

Not all of the literature is consistent with the SOKA model, however. In an earlier study on the impact of observability and evaluativeness on self-other and other-other agreement, John and Robins (1993) found that although observability and evaluativeness did influence some traits in ways consistent with the SOKA model's predictions, evaluativeness and observability could not explain all of the differences in interjudge agreement between traits. Stated more concretely, once the effects of evaluativeness and observability on interjudge agreement were partialed out, extraversion no longer had an effect on interjudge agreement but agreeableness did. This meant that although a model using evaluativeness and observability could adequately account for interjudge agreement in extraversion, the same was not true for agreeableness (John & Robins, 1993).

Metacognition and the SOKA Model

What relevance do self- and other-reported assessments of personality, evaluativeness, and observability have to metacognitive self- and other-judgments? Although assessing someone's personality is different from gauging the progress of someone's learning or their cognitive abilities, there are undeniable similarities between metacognitive judgments and personality judgments. After all, both kinds of judgments require people to know the target's capabilities and characteristics, and both judgments are influenced by available information and motivational biases (Beer & Vazire, 2017; Saenz et al., 2017; Serra & DeMarree, 2016; Thielmann, Zimmermann, Leising, & Hilbig, 2017; Tirso & Geraci, 2020; Vazire, 2010). Evaluativeness and observability may map on to the same underlying constructs as the motivational and informational factors observed in metacognition research, and it is apparent from the literature reviewed thus far that motivational and informational factors appear to be significant determinants of metacognitive self- and other-judgment accuracy. Furthermore, there is evidence that motivational and informational factors affect metacognitive self- and other-judgments differently and in ways consistent with the SOKA model. For example, Helzer and Dunning (2012) found that students readily made use of information about prior performance when predicting another student's grade on an upcoming exam but not when predicting their own grades. Additionally, Tirso and Geraci (2020) found that metacognitive other-judgments were less accurate than self-judgments when both perspectives lacked this information. These findings suggest that the information available to judges (i.e., observability) affects metacognitive other-judgments more than self-judgments. With regard to motivational factors, Helzer and Dunning (2012) also found that students' desired grades are strong predictors of their future exam grades and thus based their grade predictions on their desired grades. In contrast, knowing other students' desired grades did not exert nearly the same level of influence when students were asked to predict others' future exam grades, suggesting that metacognitive self-judgments are more heavily influenced by how desirable or important (i.e., evaluative) cognitive or academic ability is than other-judgments are, which would also be consistent with the SOKA model. Despite being designed with self- and other-judgments of personality in mind, the predictions from the SOKA model may generalize to metacognitive self- and other-judgments too, but additional research is required to make this determination.

The Goal of the Current Studies

Accordingly, the goal of the current studies was to determine to what extent the SOKA model's framework could be adapted for use with metacognitive self- and other-judgments. More specifically, the current studies set out to answer three related research questions. One, which cognitive abilities are relatively high or low in evaluativeness, and which are relatively high or low in observability? Two, what effect does evaluativeness have on the accuracy of metacognitive self- and other-judgments? And, three, what effect does observability have on the accuracy of metacognitive self- and other-judgments?

Research Question 1: Which Cognitive Abilities are Relatively High or Low in Evaluativeness and Observability?

In Studies 1A and 1B (Chapter II), I address the first research question: Which cognitive abilities are relatively high or low in evaluativeness and observability? Studies 1A and 1B were designed to determine the evaluativeness (Study 1A) and observability (Study 1B) of eight different cognitive abilities, the Big Five personality factors, and physical attractiveness for a total of 14 items. Participants were asked to rate how important each of these 14 items were to them personally regardless of their actual standing on these items in Study 1A—this served as a measure of each item's evaluativeness. In Study 1B, participants rated a series of individuals (targets) on these same 14 items, and interrater reliability for each item was used as measure of item observability consistent with prior work (Connelly & Ones, 2010). Data were collected for personality factors in addition to physical attractiveness in these studies to serve as points of comparison, as the observability and evaluativeness of these is already relatively well-established (Vazire, 2010).

Research Question 2: What Effect Does Evaluativeness Have on the Accuracy of Metacognitive Self- and Other-judgments?

Studies 2 (Chapter III) and 3 (Chapter IV) address the second research question, what effect does evaluativeness have on the accuracy of metacognitive self- and other-judgments? In Study 2, participants made a series of metacognitive judgments before completing a series of cognitive tasks designed to measure four cognitive abilities that were deemed by Studies 1A and 1B as varying substantially in their evaluativeness and observability. These abilities were creativity (low evaluativeness, high observability), visuospatial ability (low evaluativeness, low observability), working memory (high evaluativeness, high observability), and prospective memory (high evaluativeness, low observability). The tasks used to measure these abilities were an alternate uses task for creativity, a mental rotation task for visuospatial ability, an n-back for working memory, and an imaging task with a prospective memory component for prospective memory. Participants made a series of metacognitive judgments—rating their ability in each of these domains and predicting their relative performance on each task—prior to completing the experimental tasks. Participants also provided contact information for two people who knew them well and could serve as informants; these informants were later contacted and asked to make the same metacognitive judgments that participants did, but for their corresponding participants instead of for themselves. Multilevel modeling was used to examine how observability and evaluativeness affected participants' and informants' metacognitive judgment accuracy. If the SOKA model framework is sufficiently applicable to metacognitive self- and other-judgments, then evaluativeness should interact with participants' judgments by reducing judgment accuracy when evaluativeness is high and increasing judgment accuracy when evaluativeness is low; observability on the other hand should interact with informants' judgments by reducing judgment accuracy when observability is low and increasing judgment accuracy when observability is high.

Study 3 investigated the effects of evaluativeness on metacognitive self- and other-judgments by manipulating evaluativeness and examining the resulting effects on metacognitive self- and other-judgment accuracy. Participants read a series of fake articles designed to either increase creativity's evaluativeness by portraying it as a highly desirable quality, or reduce (or at least hold constant) creativity's evaluativeness by portraying it as a largely useless and unimportant quality. As in Study 2, participants made a series of metacognitive judgments and provided contact information for two informants before completing an alternate uses task designed to measure creativity. Informants were later contacted and underwent the same evaluativeness manipulation as their corresponding participants prior to making their metacognitive judgments. Once again, multilevel modeling was used to examine how participants' and informants' metacognitive judgments were affected by evaluativeness. If the SOKA model framework is sufficiently applicable to metacognitive self- and other-judgments, then participants' judgments should exhibit worse accuracy when made under conditions of high evaluativeness compared to conditions of low evaluativeness, whereas informants' judgment accuracy should be relatively unaffected by the evaluativeness manipulation.

Research Question 3: What Effect Does Observability Have on the Accuracy of Metacognitive Self- and Other-judgments?

The third and final research question investigated by the current studies was what effect does observability have on the accuracy of metacognitive self- and other-judgments? This question is the topic of Chapters III and V, and was addressed by Studies 2 and 4. How Study 2 investigates this question was already addressed in the previous section; Study 4 investigates this question by laying the groundwork for future research. Whereas Study 3 sought to manipulate creativity's evaluativeness to examine its downstream effects on metacognitive self- and other-judgment accuracy, Study 4 tests a new manipulation designed to increase creativity's observability so that future research may use this manipulation to examine observability's effects on metacognitive self- and other-judgments. This observability manipulation consisted of providing participants' informants an example of their corresponding participants engaging in a creative task in the hope that doing so would increase how observable a participant's creativity is to their informants.

CHAPTER II

STUDIES 1A AND 1B: DETERMINING THE EVALUATIVENESS AND OBSERVABILITY OF COGNITIVE ABILITIES

Study 1A

The purpose of Study 1A was to determine the evaluativeness of eight different cognitive abilities considered for inclusion in Studies 2 through 4. This was accomplished by having participants complete an online survey that instructed them to rate the personal importance of eight different cognitive abilities. Participants also rated the importance of the Big Five personality factors—extraversion, conscientiousness, emotional stability, openness, and agreeableness—and physical attractiveness. These noncognitive traits were included for comparison purposes, as the relative evaluativeness and observability of these traits has already been determined by prior works.

Method

Participants

Ninety-nine (82 females, 17 males) undergraduate students 18-30 years of age (*M* = 19.02, *SD* = 1.55) participated in Study 1A to fulfill a research requirement for course credit. Participants had completed 12.89 years of education on average (*SD* = 1.04). Sixty-four participants (64.6%) identified as being Caucasian/White, 25 (25.3%) identified as Hispanic, 8 (8.1%) identified as Asian or Pacific Islander, and 2 (2%) identified as other.

Materials and Procedure

Participants were given a link to a Qualtrics survey. After providing informed consent, participants saw a screen with the following instructions:

> *"On the next few pages you will be presented with a variety of traits and abilities. Please rate* ***how important*** *each one is to you personally.* ***Note that you are not rating how you see yourself on these traits.*** *Instead, you are rating how highly you value these traits and abilities, regardless of whether you excel at them or not. If you rate a trait as "very important" then it should matter a great deal to you whether you excel at that trait or not. In contrast, if you rate a trait as "not at all important" then it should not matter at all to you whether you excel at that trait or not." (Bold in original materials).*

The eight cognitive abilities participants rated included: retrospective memory, attentional control, working memory, logical reasoning, creativity, prospective memory, processing speed, and visuospatial ability. Participants also rated the importance of the Big Five personality factors—extraversion, conscientiousness, emotional stability, openness, and agreeableness—and physical attractiveness, resulting in 14 items in total. Since the evaluativeness of the Big Five personality factors is already known (John & Robins, 1993; Vazire, 2010), they were included in the survey to serve as a reference point against which the evaluativeness of cognitive abilities could be compared.

The 14 items participants rated were presented in a randomized order to each participant. Each item was also accompanied by a two-sentence description of the item that included a lay definition of the item followed by several examples or descriptive terms to control for differences in participants' knowledge of these items (Hayes & Dunning, 1997). For example, the item for retrospective memory read:

> *"RETROPSECTIVE MEMORY. Retrospective memory refers to your ability to remember things that have already happened. Retrospective memory is often reflected in your ability to remember where you parked your car, what your professor said about a concept or theory in a previous lecture, or that the Houston Astros won the 2017 World Series."*

Underneath each item's description was a scale labeled from 0 to 100 with the anchors "not at all important," "slightly important," "moderately important," "very important," and "extremely important," and the words "How important is [item name] to you?" to the left of the scale; see Appendix A for an example item. Participants responded by moving a slider with their mouse to the appropriate value. After rating the evaluativeness of each of the 14 items, participants were asked to make similar ratings about the observability of the same 14 items, and how much out of a $1,000 budget they would like their university to allocate to teaching students how to improve that trait or ability in a hypothetical series of workshops. These data are not discussed in this dissertation but can be found in Table 1. Item presentation order within the evaluativeness rating block, observability rating block, and budgeting block was randomized within each block, however the three aforementioned blocks were always presented in the same order to all participants. Upon completing their evaluativeness, observability, and budgeting decisions, participants completed a brief demographic questionnaire and were debriefed. A full transcript of the survey can be found online at https://osf.io/ny743/?view_only=b62690ae9a7b4e14857bd394dbe33cf8.

Results

Statistical significance for all analyses was set at $p < .05$. Descriptive statistics are reported in Table 1. Evaluativeness ratings for extraversion were compared to those of the rest of the Big Five personality factors to serve as a validity check and point of comparison. Prior work has identified extraversion as one of the least evaluative of the Big Five personality factors (Connelly & Ones, 2010; Funder & Colvin, 1988; John & Robins, 1993), and the results of Study 1A replicated this finding; extraversion was rated as less evaluative than openness ($t(97) = -5.34$, $p < .001$, $d_z = -.54$, 95% CI [-18.03, -8.25]), agreeableness ($t(96) = -6.28$, $p < .001$, $d_z = -.64$, 95% CI [-20.88, -10.85]), conscientiousness ($t(97) = -5.92$, $p < .001$, $d_z = -.60$, 95% CI [-21.70, -10.79]), and emotional stability ($t(97) = -7.30$, $p < .001$, $d_z = -.74$, 95% CI [-22.27, -12.75]). Thus, at least with regard to the Big Five personality factors, the results of Study 1A replicated prior work, and extraversion served as a reference point for what constitutes a low evaluativeness rating.

Cognitive abilities that were on the extreme ends of the evaluativeness spectrum were sought. Thus, the three highest rated abilities (logical reasoning, prospective memory, working memory) were compared to the three lowest (processing speed, creativity, visuospatial ability) in terms of evaluativeness. Logical reasoning was rated as significantly more evaluative than processing speed ($t(97) = 6.20$, $p < .001$, $d_z = .63$, 95% CI [8.23, 15.97]), creativity ($t(97) = 8.88$, $p < .001$, $d_z = .90$, 95% CI [13.75, 21.68]), and visuospatial ability ($t(97) = 10.91$, $p < .001$, $d_z = 1.10$, 95% CI [18.47, 26.69]). Similarly, prospective memory and working memory were also both rated as more evaluative than processing speed (prospective memory $t(98) = 5.04$, $p < .001$, $d_z = .51$, 95% CI [6.36, 14.63]; working memory $t(98) = 4.49$, $p < .001$, $d_z = .45$, 95% CI [5.52, 14.28]), creativity (prospective memory $t(98) = 8.16$, $p < .001$, $d_z = .82$, 95% CI [11.97, 19.66]; working memory $t(98) = 7.16$, $p < .001$, $d_z = .72$, 95% CI [11.01, 19.44]), and visuospatial ability (prospective memory $t(98) =8.39$, $p < .001$, $d_z = .84$, 95% CI [16.18, 26.20]; working memory $t(98) = 8.57$, $p < .001$, $d_z = .86$, 95% CI [15.83, 25.36]). Furthermore, evaluativeness ratings for extraversion were only marginally lower than evaluativeness ratings for processing speed ($t(97) = -1.90$, $p = .061$, $d_z = -.19$, 95% CI [-10.82, .25]), virtually identical to evaluativeness ratings for creativity ($t(97) = -.18$, $p = .857$, $d_z = .02$, 95% CI [-5.98, 4.98]), and marginally higher than evaluativeness ratings for visuospatial ability ($t(97) = 1.93$, $p = .057$, $d_z = .19$, 95% CI [-.17, 10.80]).

In summary, processing speed, creativity, and visuospatial ability were identified as a set of cognitive abilities low in evaluativeness relative to other cognitive abilities and according to the standards set by prior work (e.g., Vazire, 2010), whereas logical reasoning, prospective memory, and working memory were identified as a set of cognitive abilities high in evaluativeness. It should be noted that openness/intellect, which creativity is often considered a part of, has been considered high in evaluativeness by prior work (Beer & Vazire, 2017; John & Robins, 1993; Vazire, 2010), whereas the current study found that creativity is quite low in evaluativeness. These different patterns of results could be due to the fact that the current study assessed creativity specifically, whereas prior work assessed openness/intellect, and creativity was only one of several traits from which data were collected and used to create a single, composite rating.

Study 1B

Study 1B was conducted to determine the observability of cognitive abilities and how they compare to the Big Five personality factors and physical attractiveness. As in Study 1A, Study 1B was conducted online using undergraduate students and involved the same 14 items used in Study 1A. Instead of rating the evaluativeness of these 14 items however, participants in Study 1B rated target others on these items after watching videos depicting target others conversing with one another and/or engaging in various tasks. Interrater reliability was then computed for each of the 14 items and compared. Higher interrater reliability should be indicative of higher observability because it is a sign that judges are all "seeing" the same thing; on the other hand, lower interrater reliability should indicate lower observability because it is a sign that the item is more ambiguous and difficult for outside observers to judge (Funder & Colvin, 1988; Funder & Dobroth, 1987; John & Robins, 1993).

Method

Participants

Ninety (66 females, 23 male, 1 missing data) undergraduate students 18-24 years of age (*M* = 19.38, *SD* = 1.20) participated in Study 1B to fulfill a research requirement for course credit. Participants had completed 12.90 years of education on average (*SD* =

1.08). Fifty-two participants (57.8%) identified as being Caucasian/White, 17 (18.9%) identified as Hispanic, 17 (18.9%) identified as Asian or Pacific Islander, 2 (2.2%) identified as Black, and 1 (1.1%) identified as other.

Materials and Procedure

As in Study 1A, participants were given a link to a Qualtrics survey. After providing informed consent, participants saw a screen with the following instructions:

> *"On the next few pages you will be presented with a series of videos depicting people in various contexts. A series of questions will follow each video. These questions will ask you to rate the personality characteristics and cognitive abilities of each individual from these videos. Please pay close attention to each person in these videos. Because these videos were taken from the internet, some of them may include advertisements; feel free to disregard any advertisements you see as they are not a part of this study."*

The videos participants saw were publicly available videos from YouTube embedded into a Qualtrics survey and presented in a randomized order to participants. Each video depicted one to four people (targets) conversing with one another or the camera about a story they were sharing or a task they were engaged in (cooking, shopping for groceries, comparing snacks, etc.). Every target in every video received significant screen time and their own introduction. A full list of the videos used and links to view them can be found online at https://osf.io/ny743/?view_only=b62690ae9a7b4e14857bd394dbe33cf8.

Each video was presented on its own separate page within the survey to reduce potential distractions as participants watched, and participants could not advance to the next page until a hidden timer equal in length to the current video had expired. Immediately after viewing each video, participants rated the targets depicted in the video on the same 14 items (8 cognitive abilities, Big Five, physical attractiveness) from Study 1A. Each item was accompanied by the same two-sentence description used in Study 1A to control for differences and deficiencies in participants' knowledge of the items. For example, the item for extraversion read:

> *"EXTRAVERSION. Extraversion refers to your desire and affinity for social interactions. Extraverted people are best described as social, fun-loving, energetic, and talkative. How would you rate this individual on this domain?"*

Underneath the item description was a 1 to 7 scale with the anchors not at all extraverted, moderately extraverted, and extremely extraverted. Participants were only allowed to rate one target at a time and in a randomized order when a video depicted more than one target, and a still image of the target and the target's name appeared at the top of each block of ratings. The items within each block of ratings were presented in a randomized order. Participants rated 15 targets in total from 7 different videos ranging from 3 to 8 minutes in length. After watching all videos and rating all targets, participants completed a brief demographic questionnaire and were debriefed. A transcript of the survey can be found online at https://osf.io/ny743/?view_only=b62690ae9a7b4e14857bd394dbe33cf8 and an example item can be found in Appendix B.

Results

To quantify the observability of each cognitive ability, personality trait, and physical attractiveness, Krippendorff's alpha was computed for each of the 14 items using the "KALPHA" macro written for SPSS, which can be downloaded from http://afhayes.com/spss-sas-and-mplus-macros-and-code.html (Hayes & Krippendorff, 2007). It was reasoned that higher interrater reliability is indicative of higher observability because it is a sign that judges are all "seeing" the same thing; lower interrater reliability, on the other hand, is indicative of lower observability because it is a sign that the item is a more difficult for outside observers to judge (Funder & Colvin, 1988; John & Robins, 1993). Confidence intervals for each alpha were computed using 5,000 bootstrapped samples in order to compare alpha values and to categorize items as relatively high or low in observability. These results, integrated with the evaluativeness ratings from the preceding pilot study, are reported in Table 2. Although alpha values were low for all items, they mirrored results from prior work (Funder & Colvin, 1988; John & Robins, 1993; Vazire, 2010) and made intuitive sense: physical attractiveness was the most observable item by far, followed by extraversion, whereas internally-oriented items such as visuospatial ability, emotional stability, and openness were the least observable. Alpha values, confidence intervals, and evaluativeness ratings from Study 1A for all cognitive abilities, the Big Five personality factors, and physical attractiveness are reported in Table 1. These data and the evaluativeness data from Study

1A were used to select four cognitive abilities for additional study: working memory (high evaluativeness, high observability), prospective memory (high evaluativeness, low observability), creativity (low evaluativeness, high observability), and visuospatial ability (low evaluativeness, low observability).

CHAPTER III

STUDY 2: TESTING THE SOKA MODEL

Study 2

Study 2 served as an initial investigation into the effects of observability and evaluativeness on self- and other-judgments of various cognitive faculties. Participants completed self-judgments and measures of various cognitive abilities representing the three combinations of observability and evaluativeness investigated by Vazire (2010) and a fourth that was left uninvestigated: working memory, prospective memory, creativity, and visuospatial ability. Participants made a general ability judgment and a percentile rank judgment about each of the four cognitive abilities that were investigated, provided contact information for potential informants, and then completed a series of tasks designed to measure their working memory, prospective memory, creativity, and visuospatial ability. Informants were sent a recruitment email and asked to make the same judgments participants made, but with participants as the target of their judgments instead of the informants themselves (i.e., a standard informant-report procedure). Based on the SOKA model, it was expected that:

1. Participants' metacognitive judgments would be less accurate for highly evaluative tasks compared to less evaluative tasks.
2. Informants' metacognitive judgments would be more accurate for highly observable tasks compared to less observable tasks.

Method

Participants

Participants were undergraduate students recruited from the Sona subject pool system. One hundred ninety-seven participants were recruited in total, and each participant was asked to provide contact information for two informants. This sample size was selected because the true effect sizes for this work are unknown, and a priori power analyses indicate that this sample size would enable the detection of small to somewhat small sized effects in most analyses with 80% power or more (Faul, Erdfelder, Lang, & Buchner, 2007; Faul, Erdfelder, Lang, & Buchner, 2009; Schönbrodt & Perugini, 2013).

Because the current study's analyses required that data be available from both participants and informants, participants whose informants did not participate in the study were excluded from the sample. Additionally, one participant was excluded as an outlier because their informant provided judgments that were more than 3 standard deviations below the mean. This resulted in a final sample of 105 participants (75 female, 30 male) and 133 informants (94 females, 37 males, 2 missing). Of the undergraduate participants in the final sample, 67 freshmen were, 25 were sophomores, 7 were juniors, and were 6 seniors. Additionally, 70 identified as Caucasian/white, 26 identified as Hispanic, 6 identified as Asian or Pacific Islander, 1 identified as black, 1 identified as American Indian or Alaskan Native, and 1 identified as "other". Of the informants in the final sample, 85 identified as Caucasian/white, 39 identified as Hispanic, 5 identified as Asian or Pacific Islander, and 2 identified as black. Informants were primarily close friends and family members—55 were parents, 43 were friends, 18 were siblings, 14 were romantic partners, 2 were other relatives, and 1 was a co-worker or supervisor. This final sample of 105 participants and 133 informants was used in all analyses.

Materials

All materials can be found online at https://osf.io/ny743/?view_only=b62690ae9a7b4e14857bd394dbe33cf8.

Mental Rotation Task

A mental rotation task (MR task) was constructed using Lab.js (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2019) to assess visuospatial ability. It was hosted on Open Lab (https://open-lab.online) and was identical to the mental rotation task used by Peronnet and Farah (1989), but with 160 trials instead of 200. In this task, participants were informed that they will be shown a series of letters, and that for each letter they must decide as quickly as possible whether it is a normal letter by pressing the N key on their keyboard or the mirror image of a letter by pressing the M key. Participants then viewed six example trials that included brief explanations about how they would respond to that specific trial during the experiment; participants could only advance through these examples by pressing the correct response (N for normal letters or M for mirror images). These examples included two upright letters (0°) and 4 rotated letters (two at 90° and two at 180°). Three of these examples were normal letters and three were mirrored. After completing all six examples, participants completed 16 practice trials that were randomly selected from the pool of all possible experimental trials, meaning that some trials were presented twice in the task. After completing the practice trials, participants completed the 160 experimental trials in a random order.

The letters F, G, J, and K were used as the stimuli. Each letter was presented 8 times rotated to 0°, 4 rotated to 45°, 4 rotated to 90°, 4 rotated to 135°, 8 rotated to 180°, 4 rotated to 225°, 4 rotated to 270°, and 4 rotated to 315°. This resulted in 8 trials each requiring 0°, 45°, 90°, 135°, and 180° of rotation, 40 trials in total for each letter, and 160 trials in total across all four letters. Half of all trials for each letter at each orientation were normal and the other half were mirrored. Each letter presentation was preceded by a fixation cross which remained on screen for 500ms, followed by a 2000ms interstimulus interval. During the practice trials, letters remained on screen until participants made a response (N or M), but during the experimental trials, letters remained on screen for a maximum of 2000ms or until participants made a response—whichever occurred first. Participants were scored based on their median correct reaction time (RT). The mental rotation task can be found online at https://rtirso-diss-mentalrotation.netlify.com for demonstration purposes.

Alternate Uses Task

The alternate uses task (AUT; Christensen, Guilford, Merrifield, & Wilson, 1960) was used to assess creativity. It was administered via a Qualtrics survey. Participants were instructed that they will be given a creativity task in which they must produce as many different uses as they can think of for a given object that differ from the object's stated normal use. The AUT consisted of a single item: a brick, with the stated normal use of building walls. Participants had four minutes to come up with as many alternate uses for this item as they could, and entered their responses into a text box. Participants' responses on the AUT were scored for fluency—the number of viable alternate uses they provided for a brick—by a team of six undergraduate research assistants. Krippendorff's alpha indicated a high degree of interrater reliability ($\alpha = .88$) among coders; thus, coders' fluency scores were averaged together and used to create a single measure of fluency on the AUT.

Prospective Memory Task

To assess prospective memory, a prospective memory task (PM task) was constructed using Lab.js (Henninger et al., 2019) and hosted on Open Lab (https://open-lab.online) for participants to complete using a computer. It was based off of the mental imagery task used by Scullin, Einstein, and McDaniel (2009), with some modifications. In this task, participants were instructed that they would be shown a series of words, and that their job was to rate how difficult they find it is to create mental images representing each word displayed by pressing the 1 (easy), 3 (difficult), or 2 (somewhere in between) keys. Participants were also informed that they had 2.5 seconds to rate each word before it disappeared, and were encouraged to work as quickly and accurately as possible. Ten practice trials followed these instructions. After the first ten practice trials, participants were informed that they would also have a secondary task to perform during the experiment: pressing "Q" instead of making an imagery rating when the word displayed was an animal. Participants were reminded on this same screen that their primary job was the image rating task, and that they should work as quickly and as accurately as possible. Another block of 10 practice trials containing two critical trials began after participants acknowledged the instructions. After the second practice block, participants were reminded of their primary and secondary tasks before beginning a series of 160 experimental trials.

Of the experimental trials, trials 35, 75, 115, and 155 were always critical trials. Both practice trial blocks and the blocks of filler trials separating the critical trials were presented in a random order to participants. Each trial consisted of a fixation cross in the center of the screen for 500ms, followed by the stimulus word, which stayed on screen for 2500ms or until participants made an appropriate response (1, 2, 3, or Q), followed by a 200ms interstimulus interval. The response scale for the image-rating portion of the task, but not the prospective memory portion, appeared at the bottom of the screen for every trial. All filler and critical words used in the PM task were matched on frequency of occurrence based on the norms developed by Brysbaert and New (2009). The critical words used in the second practice block were *horse* and *fish*, and the critical words used in the experimental trials were *cat* (35), *dog* (75), *bird* (115), and *pig* (155). The number of experimental trials and critical trials was double that used by Scullin et al. (2009) in order to provide a wider range of possible performance. Given the number of critical words, animals as a category was used as the prospective memory cue rather than specific words such as *corn* and *dancer* (Scullin et al., 2009) to avoid prospective memory failures due to participants failing to sufficiently encode all four critical words before beginning the task. Participants were scored based on the proportion of critical trials they correctly remembered to press the “Q” key on. The PM task can be found online at https://rtirso-diss-pmtask.netlify.com for demonstration purposes.

N-back Task

To assess working memory, an n-back task was also constructed in Lab.js (Henninger et al., 2019) by using the n-back template made by Felix Ludwig and hosted on Open Lab (https://open-lab.online). An instructions screen told participants that they would be shown sequences of letters, and that their job was to judge whether or not each letter shown matches a set of conditions unique to that sequence. These "conditions" were the specific instructions for the different blocks (0-back, 1-back, etc.) and were explained to participants at the start of each block. Participants pressed the "j" key on their keyboards to indicate that the letter shown on screen matches the sequence's conditions, and the "k" key if it does not; this information remained at the bottom of participants' screens throughout the task. There was a practice and an experimental block for 0-back, 1-back, 2-back, and 3-back conditions, resulting in 8 total blocks. Practice blocks consisted of 15 trials, 3 of which were match trials. Experimental blocks consisted of 50 trials, 16 of which were match trials. A 500ms interstimulus interval preceded each letter presentation, and each letter was presented for 2500ms or until participants made a response. The letters B, b, T, t, V, v, G, and g were used as stimuli, and participants were instructed to treat uppercase and lowercase letters as the same letter. Participants were scored based on the number of critical trials they responded correctly to. The n back task is available online at https://rtirso-diss-nback.netlify.com for demonstration purposes.

Procedure

Study 2 and all subsequent studies took place entirely online. Participants were given a link to a Qualtrics survey that hosted the questionnaire component of the study. After providing consent, participants were told that the study they are participating in is investigating how accurately people can assess their own prospective memory, working memory, visuospatial ability, and creativity. Then, participants were told that they would be completing a series of cognitive tasks and would complete a series of eight judgments in pairs of two (two judgments for each ability being tested). The same definitions for prospective memory, working memory, visuospatial ability, and creativity that were provided to participants in Studies 1A and 1B appeared alongside their corresponding judgment pairs. These judgment pairs were presented in a random order for each participant. The first prediction participants made in each pair was a general ability judgment adapted from Vazire (2010) that utilized a 1 to 15 Likert scale with the anchors "1 – Extremely Bad" and "15 – Extremely Good." The exact wording of this item differed slightly for each domain, but the creativity version serves as a representative example:

> *"Given your knowledge of your own abilities, how would you describe your creativity compared to the average university student? For example, if you think you are extremely creative compared to most university students, you might select 14 or 15 on the slider below. Conversely, if you think you are extremely uncreative compared to most university students, you might select 1 or 2 on the slider below."*

After making their general ability judgment, participants were asked to make a percentile rank judgment:

> *"Imagine that these tasks will be given to a sample of 100 university students (including you). Given your knowledge of your own abilities, how well do you think you will perform on this task compared to everyone else? Specifically, of the 99 other students in this sample, how many do you think you will outperform on this task? Use the slider below to make your response."*

After completing all eight self-judgments, participants completed a brief informant questionnaire. This questionnaire asked them for the names and email addresses of two people who know them well and could serve as informants. Participants were told that informants would be emailed a brief explanation of the study and a link to an online survey that would ask informants to make the same judgments that they just made on the previous page, and that informants' responses would be confidential. The informant questionnaire also inquired about the number of years participants had known each of their informants for, the nature of their relationship with each of their informants, and to estimate how frequently they talk or interact with their informants.

The AUT, prospective memory, n-back, and mental rotation tasks followed after the informant questionnaire in a random order for each participant. The AUT was administered as part of the Qualtrics survey. However, for the prospective memory, n-back, and mental rotation tasks, participants were given an ID number in Qualtrics, instructed to use a link to access the Open Lab page that each task is hosted on, and enter their ID number when prompted. It was explained to participants that they would receive a completion code upon completing each task, and that they must enter that completion code into the Qualtrics survey in order to continue. Once participants had received and entered all of their completion codes into the Qualtrics survey, they answered some basic demographic questions and were debriefed.

Informants were sent a recruitment email sometime after their referring participant had completed the study. This email mentioned informants' referring participant by name, provided a brief summary of the study their participant participated in, and invited informants to answer a few questions about their participant and predict how well their participant had performed in the lab. It was also explained to informants that their responses would be kept confidential, that their participation in the study was completely voluntary, and that neither they nor their referring participant would be penalized in any way should they choose not to participate. The recruitment email ended with a link to a Qualtrics survey containing the informant response form, the referring participant's subject number (for entry into the informant response form), and the author's contact information. Informants simply needed to click or copy and paste the included link into a web browser to access and complete the informant response form. The form itself began with the informed consent process, then asked informants to provide their first and last name and enter their referring participant's ID number. These steps served to link informants' responses to participants' data for analysis. Afterwards, informants completed the same judgment procedure as participants, but with their participants serving as the target rather than the informants themselves. For example, the informants' Likert scale general ability judgment for creativity read:

> *"Given your knowledge of your participant, how would you describe their creativity compared to the average university student? For example, if you think they are extremely creative compared to most university students, you might select 14 or 15 on the slider below. Conversely, if you think they are extremely uncreative compared to most university students, you might select 1 or 2 on the slider below."*

Just like for participants, each general ability judgment was also followed by a percentile rank judgment for the same cognitive ability for informants. The percentile rank judgment for creativity read:

> *"Imagine that the tasks your participant completed were given to a sample of 100 university students (including your participant). Given your knowledge of your participant's abilities, how well do you think they performed on the creativity task compared to everyone else? Specifically, of the 99 other students in this sample, how many do you think your participant outperformed on this task? Use the slider below to make your response."*

Additionally, the first judgment pair informants' referring participants completed (creativity, visuospatial ability, working memory, or prospective memory) was presented to informants first. The informant response form concluded with a demographic questionnaire and the debriefing. Reminder emails were sent to informants one week, two weeks, and three weeks after all participants' data had been collected to boost response rates.

Results

Judgment accuracy was defined as the strength of the association between judgments (either participants' or informants') and actual ability as measured by the AUT, n-back, prospective memory, and mental rotation tasks. The primary goal of Study 2 was to examine how participants' and informants' judgments interacted with the observability and evaluativeness of the domains (creativity, working memory, prospective memory, and visuospatial ability) being judged. Given that the focus of the current study was on judgment accuracy and not the direction of errors (i.e., over- and underconfidence effects), and due to the nature of the predictions participants made, the current studies relied primarily upon the linear association between predicted performance and actual performance—also known as judgment resolution—as the measure of judgment accuracy. Additionally, because the hypotheses being tested required comparisons between participants' performance, participants' judgments, and informants' judgments, only participants who had at least one informant report available were used for the following analyses. As previously mentioned, this resulted in a sample size of 105 participants and 133 informant reports. When data from more than one informant was available for a given participant, informant judgments were averaged together to create a single set of aggregated informant judgments for that participant. To anticipate, whether or not informant data were based on a single informant report or two informant reports did not moderate informant judgment accuracy (b = .03, SE = .07, t(438.12) = .39, p = .696). Additionally, the order in which participants and informants completed their judgments and the experimental tasks bore no relationship to judgment accuracy (all ps > .361). Note that differences in degrees of freedom across analyses were due to missing data.

Preliminary Analyses

Overconfidence

It has been well-documented that when people predict their performance on cognitive tasks they tend to overestimate it, and this overestimation tends to be greater the lower one's actual performance is (the Dunning-Kruger effect; Kruger & Dunning, 1999). The predictions used in the current study—a 1-15 Likert scale rating of ability and a percentile rank prediction—were not entirely conducive to investigating overconfidence effects because this was not the primary goal of the study. Nevertheless, participants' predicted percentile rankings can be compared to their actual percentile rankings to shed some light on this question. The results of these comparisons indicated that, on average, participants did not over- or underestimate their percentile rankings on the prospective memory task (t(96) = 1.77, p = .080, d_z = .18), the AUT (t(99) = -1.09, p = .277, d_z = -.11), the n-back (t(92) = -.41, p = .680, d_z = -.04), or the mental rotation task (t(66) = -1.13, p = .264, d_z = -.14). Looking at the data by performance quartile shows a different story, however. Participants in the lowest performance quartile on each task overestimated their percentile ranks on the prospective memory task (t(23) = 13.02, p < .001, d_z = 2.66), n-back (t(24) = 6.88, p < .001, d_z = 1.38), mental rotation task (t(16)

= 4.59, $p < .001$, $d_z = 1.11$), and AUT ($t(25) = 7.03$, $p < .001$, $d_z = 1.38$). In contrast, participants in the highest performing quartile on each task underestimated their percentile ranks on the prospective memory task ($t(12) = -4.77$, $p < .001$, $d_z = -1.32$), n-back ($t(21) = -7.57$, $p < .001$, $d_z = -1.61$), mental rotation task ($t(16) = -5.35$, $p < .001$, $d_z = -1.30$), and AUT ($t(24) = -10.34$, $p < .001$, $d_z = -2.07$). In other words, there was clear evidence that the current study replicated the often-observed Dunning-Kruger effect (sometimes called the *unskilled and unaware effect*; Kruger & Dunning, 1999). Full descriptive statistics can be found in Table 2.

Comparing Participants' and Informants' Judgments

In addition to underconfidence among high performers and overconfidence among low performers, prior work has also found that metacognitive other-judgments are often more optimistic than metacognitive self-judgments (e.g., Tirso & Geraci, 2020); paired samples t-tests comparing participants' predictions to informants' predictions revealed that the current study replicated this finding. Every single prediction informants made—their general ability and percentile rank judgments for all four experimental tasks—was significantly higher than their corresponding participants' predictions (all ps < .001, all d_zs > .56).

Two correlation matrices were calculated to provide an initial glimpse into the accuracy of participants' and informants' general ability and percentile rank judgments—one matrix for Likert scale judgments, and one matrix for percentile rank judgments. Neither participants' nor informants' predictions were correlated with actual performance on any task. These correlation matrices are reported in their entirety in Table 3 and in Table 4.

Main Analyses

Given that multiple judgments (general ability and percentile rank) and measures of performance (raw score and percentile rank) were nested within each of the four tasks ($\rho = .81$), and these tasks were in turn nested within participants ($\rho = .13$), a 3-level multilevel modeling approach was used. Participants' performance on the experimental tasks was the dependent variable. Participants' and informants' judgments, along with the scale (general ability or percentile rank) that they were made on were level 1 variables; task observability and evaluativeness were level 2 variables. There were no predictors entered at level 3, but level 3 was retained to accommodate the nested nature of the data. Participants' raw scores and percentile ranks for each experimental task, along with participants' and informants' corresponding judgments, were standardized; this permitted the author to determine whether relatively high judgments actually predicted relatively high performance or not, and the extent to which observability and evaluativeness moderated the relationship between judgments and performance. All multilevel models were run in R version 3.6.3 (R Core Team, 2020) using RStudio version 1.2.5042 (RStudio Team, 2020), using the packages lme4 (Bates, Maechler, Bolker, & Walker, 2015), lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017), merTools (Knowles & Frederick, 2019) jtools (Long, 2020), reghelper (Hughes, 2020), jmv (Selker, Love, & Dropmann, 2020), and interactions (Long, 2019). All models also used full maximum likelihood estimation and $p < .05$ as the threshold for statistical significance. All *p* values for the following analyses were calculated using Satterthwaite's formula for degrees of freedom.

As outlined earlier, the current study defined judgment accuracy as the linear association between judgments and actual performance. Additionally, participants' performance, participants' judgments, and informants' judgments were all converted to z-scores to facilitate comparisons across the various ranges of possible performance for each task. Thus, a zero represents mean performance for any given task or the mean judgment for any given judgment. The judgment scale (general ability or percentile rank) was entered using dummy coding. Task evaluativeness and observability were both dummy-coded to indicate whether a given judgment was for a task that was high or low in evaluativeness and high or low in observability. An iterative process was used to construct the model, starting with the empty, null model (Model 1.0) to serve as a baseline against which to compare later models and to calculate the intraclass correlation coefficients for the four tasks (level 2; $\rho = .81$) and participants (level 3; $\rho = .13$). The ICC indicates the percentage of variance accounted for by the clustering variables themselves; the high ICC values observed in the current study clearly indicate the need for a multilevel modeling approach. Additional predictors were then added in subsequent models, with each new model's overall fit compared to that of the previous one.

The final model, Model 1.2, revealed that participants' judgments ($b = .10$, SE = .04, $t(481.76) = 2.19$, $p = .029$) significantly predicted their actual performance but informants' judgments ($b = .05$, SE = .04, $t(468.02) = 1.22$, $p = .222$) did not. Judgment scale (general ability/Likert scale or percentile rank) did not moderate participants' (*b* = -

.01, SE = .02, *t*(354.96) = -.63, *p* = .528) or informants' judgment accuracy (*b* = .00, SE = .02, *t*(356.53) = .07, *p* = .942). Similarly, observability did not significantly moderate the accuracy of participants' judgments (*b* = -.04, SE = .05, *t*(498.45) = -.87, *p* = .385) or informants' judgments (*b* = -.06, SE = .04, *t*(491.27) = -1.40, *p* = .162). However, evaluativeness did significantly moderate the accuracy of participants' judgments (*b* = -.10, SE = .05, *t*(494.80) = -2.27, *p* = .024), but not informants' judgments (*b* = -.02, SE = .04, *t*(491.64) = -.38, *p* = .708). Finally, at no point in the model-building process did adding any predictors result in a statistically significant increase in model fit compared to the null model. Full model results are reported in Table 5. The interactions between task evaluativeness, task observability, participants' judgments, and informants' judgments are depicted in Figures 1 through 4.

In summary, participants demonstrated greater metacognitive accuracy than informants did. Participants' and informants' metacognitive accuracy was not affected by the type of scale they made their judgments on as indicated by the correlation matrices calculated earlier and the results from Model 1.2. Observability did not have any effect on participants' judgment accuracy, which is consistent with the literature on the SOKA model, but observability also did not affect informants' judgment accuracy, which is inconsistent with the SOKA model. Evaluativeness on the other hand did influence participants' judgment accuracy in a manner consistent with the SOKA model, with participants' judgments on highly evaluative tasks exhibiting less accuracy than their judgments on less evaluative tasks. Lastly, evaluativeness had no discernable effect on informants' judgment accuracy, which is consistent with the SOKA model. Overall, results would best be described as mixed. Evaluativeness affected judgment accuracy just as predicted by the SOKA model, but observability did not. It is possible that observability did not moderate informants’ judgment accuracy because the SOKA framework simply cannot be fully adapted to metacognitive self- and other-judgments. However, it is also possible that observability did not influence informant judgment accuracy due to a range restriction problem with the magnitude of the differences in observability between the four experimental tasks. In other words, it is possible that creativity, working memory, prospective memory, and visuospatial ability simply do not differ enough in their observability for observability to serve as a useful predictor of judgment accuracy. This potential limitation was partly addressed in Study 4. At present however, evaluativeness seems to be a useful predictor of judgment accuracy, highlighting the role that motivational biases play in reducing metacognitive accuracy.

CHAPTER IV

STUDIES 3 – 4: MANIPULATING EVALUATIVENESS AND OBSERVABILITY

Study 3

Evaluativeness is said to bias self-reported judgments more so than informant-reported judgments because of the motivational and ego-related biases present in self-judgments of highly evaluative traits (Vazire, 2010). Indeed, the results from Study 2 showed that evaluativeness significantly affected participants' judgment accuracy. However, to the best of my knowledge, no published study using the SOKA model has manipulated evaluativeness and examined the resulting effect on self- and informant-reported metacognitive judgments. Therefore, the purpose of Study 3 was to provide a stronger test of the SOKA model by manipulating evaluativeness and examining its impact on self- and informant-reported judgments. Participants made two sets of general ability and percentile rank judgments prior to completing the experimental task. Before each set of judgments, participants read a brief, fictional article extolling the benefits and importance (or lack thereof) of creativity, resulting in a within-subjects design. After both reading-judgment cycles, participants in all three conditions completed the same informant questionnaire and AUT that were used in Study 2.

It should be noted that whereas Study 2 investigated all four potential combinations of evaluativeness and observability, Study 3 investigated only one: creativity (low evaluativeness, high observability). This design choice was due to two reasons. One, investigating the same four domains from Study 2 in Study 3 would have required quadrupling the total number of participants enrolled in Study 3 (from approximately 200 to 800), or sacrificing a considerable amount of statistical power; both of these options would have greatly complicated the reporting and interpretation of the results. Two, creativity's evaluativeness from Study 1A made it the best choice for Study 3 because there was still considerable room separating it from the floor and ceiling of evaluativeness ratings, and yet it was not rated quite as low as visuospatial ability was. In other words, there was ample room for creativity's evaluativeness to increase or decrease, and there was already evidence that higher and lower evaluativeness scores were indeed possible to achieve (see Table 1).

Method

Participants

Participants were undergraduate students recruited from the Sona subject pool system. Two hundred and three participants were recruited in total, and each participant was asked to provide contact information for two informants just as in Study 2. This sample size was selected because the true effect sizes for this study is unknown, and a priori power analyses indicated that this sample size would enable the detection of small to somewhat small sized effects in most analyses with 80% power or more (Faul et al., 2007; Faul et al., 2009; Schönbrodt & Perugini, 2013).

As in Study 2, the primary analyses required that each participant had data from at least one informant. Thus, only participants with data from one or more informants available were included. Additionally, 3 participants were dropped from the dataset because their informants provided judgments that were more than 3 standard deviations above or below the mean. These steps resulted in a final sample of 96 (74 female, 22 male) participants and 124 (84 female, 40 male) informants. Of the undergraduate participants, 55 were freshmen, 26 were sophomores, 13 were juniors, and 2 were seniors. Additionally, 60 of them identified as Caucasian/white, 21 as Hispanic, 12 as Asian or Pacific Islander, 1 as black, 1 as American Indian or Alaskan Native, and 1 as “other”. Of the informants in the sample, 81 identified as Caucasian/white, 29 as Hispanic, 12 as Asian or Pacific Islander, and 2 as black. Like in Study 2, informants were primarily friends and family members—57 were parents, 31 were friends, 25 were siblings, and 11 were romantic partners. All analyses were based on this final sample.

Materials

All materials can be found online at https://osf.io/ny743/?view_only=b62690ae9a7b4e14857bd394dbe33cf8.

Article Manipulation

Two fake articles were created to manipulate the evaluativeness of creativity. The high evaluativeness (HE) article was titled “Creativity: The New Key to Success,” and described creativity as being essential for success in today’s modern economy. To create the low evaluativeness (LE) article, the title of the HE article was changed to “Creativity: Much Ado About Nothing,” and all positive statements in the HE article were changed to neutral statements that portrayed creativity as being irrelevant. For example, “Because of these internal motivations and higher levels of productivity, employers strongly favor more creative applicants,” was changed to “When asked to choose between highly creative individuals and non-creative individuals, employers did not favor either type of applicant." The articles were 110 and 136 words in length, respectively, and were loosely based on Forbes and Domm (2004) and Smith (2001), with some creative liberties taken to increase the chances of the manipulation having its desired effect. The articles themselves can be found in Appendices C and D

Alternate Uses task

The alternate uses task (AUT) used in Study 3 was identical to the AUT used in Study 2.

Procedure

Participants were given a link to a Qualtrics survey that contained all of the study materials. The survey began with the informed consent process, followed by the article manipulation. For the manipulation, participants were told that the study was designed to measure creativity, and that they were to read an excerpt from a recent article on creativity published by *Academia Daily* (a fictional organization). Then, participants were presented with Article 1. For half of the participants, Article 1 was the HE article; for the other half of participants, Article 1 was the LE article. Then, participants were given the same definition of creativity used in Studies 1A, 1B, and 2, and were asked to make general ability and percentile rank judgments just as in Study 2. Participants were also asked to rate how important creativity is to them using a slightly modified version of creativity's Likert-scale evaluativeness item from Study 1A; this item was intended to serve as a manipulation check.

After Article 1 and its corresponding judgments, participants were given a cover story. This story claimed that a secondary goal of the study was to investigate the effects of misinformation and fake news on decision-making, and that this goal was accomplished by presenting participants with a fake news article (Article 1) that was created by taking an article from the Association for Psychological Science, changing a few words, and inventing an academic-sounding publication ("Academia Daily") to make the fabricated article seem more credible. Participants were then asked to rate how believable they found Article 1 (1 – not very believable to 5 – very believable) and told that they were to read the original, unaltered version of Article 1 to correct any misconceptions it introduced, and then provide an update to their previous judgments. They were told specifically, "Your new predictions can increase, decrease, or stay the same—do not worry about your previous predictions, simply answer the questions to the best of your ability." Then, participants were shown Article 2 (either the LE or HE article—whichever one was not used for Article 1) and asked to make a second set of general ability, percentile, and evaluativeness judgments. Finally, participants rated how believable they found Article 2 compared to Article 1 (1 – not very believable to 5 – very believable). This manipulation yielded two sets of general ability and percentile rank judgments: one set immediately after the HE article (HE judgments), and one immediately after the LE article (LE judgments). After the article manipulation, participants completed the same informant questionnaire and AUT that were used in Study 2, and then were debriefed. To anticipate, no significant order effects from the within-subjects nature of this manipulation were observed.

Informants were sent a recruitment email similar to the one used in Study 2. After providing consent and entering their information into the survey, informants completed the same article manipulation that their referring participants did, including all of the same judgments (but with their referring participants as the targets instead of informants). Additionally, the HE and LE articles were presented to informants in the same order that they were presented to their referring participants. Afterwards, informants completed a brief demographic questionnaire and were debriefed.

Results

As in Study 2, judgment accuracy was defined as the strength of the association between judgments (either participants' or informants' judgments) and actual creativity as measured by the AUT. The primary goal of Study 3 was to examine how evaluativeness interacted with participants' and informants' judgments. Doing so required comparisons between participants' performance, participants' judgments, and informants' judgments, meaning that only participants who had at least one informant report available were used for the following analyses. As previously mentioned, this resulted in a sample size of 96 participants and 124 informant reports. When data from more than one informant were available for a given participant, informant judgments were averaged together to create a single set of aggregated informant judgments for that participant. To anticipate, whether or not informant data were based on a single informant report or two informant reports did not moderate the accuracy of informants' low evaluativeness (b = -.09, SE = .13, t(102.01) = -.74, p = .460) or high evaluativeness (b = .17, SE = .13, t(103.18) = 1.34, p = .183) judgments. Note that differences in degrees of freedom across analyses were due to missing data. Descriptive statistics are reported in Table 6.

Preliminary Analyses

Manipulation Check

Participants' low evaluativeness judgments (i.e., their judgments made immediately after reading the low evaluativeness article) were compared to their high evaluativeness judgments to determine if the manipulation produced a difference in participants' judgments. These comparisons revealed participants' low evaluativeness Likert-scale general ability judgments did not differ from their high evaluativeness general ability judgments ($t(95) = 1.39$, $p = .168$, $d_z = .14$). Participants' low evaluativeness percentile rank judgments did not differ from their high evaluativeness percentile rank judgments either ($t(95) = -.04$, $p = .969$, $d_z = .00$). However, participants rated creativity as being more evaluative after reading the high evaluativeness article ($M = 71.14$, $SD = 19.15$) compared to after reading the low evaluativeness article ($M = 68.23$, $SD = 20.34$, $t(95) = 2.93$, $p = .004$, $d_z = .30$). Thus, the evaluativeness manipulation succeeded in producing a statistically significant difference in evaluativeness, but this difference does not appear to be practically meaningful—it amounted to a 3-point difference on a 100-point scale, and it did not affect participants' metacognitive judgments.

Informants also went through the same evaluativeness manipulation as their participants did, and their data essentially mirrored their participants' data. More specifically, the evaluativeness manipulation did not produce any changes in informants' general ability judgments ($t(95) = -.91$, $p = .365$, $d_z = -.09$) or their percentile rank judgments ($t(95) = .51$, $p = .609$, $d_z = .05$) despite informants rating creativity as more evaluative after reading the high evaluativeness article (*M* = 80.12, *SD* = 13.58) compared to after reading the low evaluativeness article (*M* = 76.08, *SD* = 16.12, *t*(95) = 4.39, *p* < .001, d_z = .45). Once again, the evaluativeness manipulation seems to have produced a statistically significant difference in evaluativeness, but not a meaningful one—in this case, the difference amounted to a 4-point difference on a 100-point scale, but no difference in informants' actual judgments.

Finally, the within-subjects nature of the manipulation did not result in any order effects. Whichever article participants and informants read first bore no relationship to the accuracy of their high evaluativeness judgments (participants: *b* = -.02, *SE* = .13, *t*(102.13) = -.20, *p* = .844; informants: *b* = -.13, *SE* = .12, *t*(105.48) = -1.07, *p* = .286) or their low evaluativeness judgments (participants: *b* = .03, *SE* = .12, *t*(100.27) = .24, *p* = .810; informants: *b* = .13, *SE* = .12, *t*(106.65) = 1.08, *p* = .282).

Overconfidence

Participants' predicted percentile rankings were compared to their actual percentile rankings. Given that the evaluativeness manipulation had no discernable effect on judgments, the results of these comparisons were nearly identical for participants' low and high evaluativeness judgments. These results indicated that, on average, participants' low evaluativeness judgments (*t*(95) = 1.88, *p* = .063 d_z = .19) and high evaluativeness judgments (*t*(95) = 1.89, *p* = .062, d_z = .19 were not significantly higher than their actual percentile rankings on the AUT. However, looking at the data by performance quartile revealed the presence of a Dunning-Kruger effect. For participants in the lowest performance quartile, both low evaluativeness (*t*(23) = 13.83, *p* < .001, d_z =

2.82) and high evaluativeness ($t(23) = 14.73$, $p < .001$, $d_z = 3.01$) judgments exceeded participants' actual percentile ranks on the AUT; for participants in the highest performance quartile, both low evaluativeness ($t(23) = -5.93$, $p < .001$, $d_z = -1.21$) and high evaluativeness ($t(23) = -5.48$, $p < .001$, $d_z = -1.12$) judgments were lower than participants' actual percentile ranks. Full descriptive statistics can be found in Table 6.

Comparing Participants' and Informants' Judgments

Previous research has shown that informant-reported metacognitive judgments are often more overconfident than self-reported metacognitive judgments (Tirso & Geraci, 2020). Paired samples t-tests comparing participants' predictions to informants' predictions revealed that the current study replicated this finding. Every prediction that informants made—general ability and percentile rank judgments under conditions of both low evaluativeness and high evaluativeness—was significantly higher than their corresponding participants' predictions (all *p*s < .001, all d_zs > .67). Interestingly, informants also rated creativity as being more evaluative than participants after both the low evaluativeness ($t(95) = -3.18$, $p = .002$, $d_z = -.32$) and the high evaluativeness ($t(95) = -3.81$, $p < .001$, $d_z = -.39$) articles.

Two correlation matrices, one for general ability judgments and another for percentile rank judgments, were created to provide an initial glimpse into the accuracy of participants' and informants' judgments. Neither participants' nor informants' predictions were correlated with actual performance on any task, however the correlations between participants' low evaluativeness ($r(94) = .17$, $p = .094$) and high evaluativeness ($r(94) = .19$, $p = .068$) general ability judgments and AUT fluency came notably close to reaching significance, especially when compared to all other judgment-performance correlations. Full correlation matrices can be found in Table 7 and in Table 8.

Main Analyses

Similar to Study 2, multiple judgments (general ability and percentile rank) and measures of performance (raw score and percentile rank) were nested within each participant ($\rho = .95$). Given this nested nature of the data and the high intraclass correlation coefficient (possibly due to how similar each participant's and informant's judgments were given the manipulation's lack of an effect), a 2-level multilevel modeling approach was used. Participants' performance on the AUT was the dependent variable. Participants' and informants' low evaluativeness and high evaluativeness judgments, along with the scale (general ability or percentile rank) that they were made on were level 1 variables. The order in which participants and informants completed the evaluativeness manipulation was entered at level 2, but as previously mentioned there were no significant order effects so this variable was removed. Also like in Study 2, participants' performance on the AUT—fluency and percentile rank—along with participants' and informants' judgments were standardized to permit determinations about whether relatively high judgments actually predicted relatively high performance or not, and whether high evaluativeness judgments were less accurate predictors of performance than low evaluativeness judgments. Thus, zero represented mean performance on the AUT and the mean for any given judgment. The judgment scale (general ability or percentile rank) was entered using dummy coding. Participants' and informants' high and low evaluativeness judgments were entered as separate predictors along with their interactions with judgment scale. All multilevel models were run in R version 3.6.3 (R Core Team, 2020) using RStudio version 1.2.5042 (RStudio Team, 2020), using the packages lme4 (Bates, Maechler, Bolker, & Walker, 2015), lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017), merTools (Knowles & Frederick, 2019) jtools (Long, 2020), reghelper (Hughes, 2020), jmv (Selker, Love, & Dropmann, 2020), and interactions (Long, 2019). All models also used full maximum likelihood estimation and $p < .05$ as the threshold for statistical significance. All *p* values for the following analyses were calculated using Satterthwaite's formula for degrees of freedom.

Judgment accuracy was defined as the linear association between judgments and actual performance. An iterative process was used to construct the model, starting with the empty, null model (Model 2.0) to serve as a baseline against which to compare later models and to calculate the intraclass correlation coefficient (ρ = .95). The ICC indicates the percentage of variance accounted for by the clustering variables themselves—in this case, the individual participants each cluster of judgments pertained to. Additional predictors were then added in subsequent models, with each new model's overall fit compared to that of the previous one. The final model, Model 2.1b, revealed that neither participants' low evaluativeness judgments (*b* = -.05, *SE* = .06, *t*(99.42) = -.78, *p* = .436) nor their high evaluativeness judgments (*b* = .05, *SE* = .07, *t*(99.80) = .72, *p* = .471) significantly predicted their actual performance on the AUT. Informants' judgments did not fare any better—neither their low evaluativeness judgments (*b* = -.02, *SE* = .07, *t*(102.00) = -.23, *p* = .818) nor their high evaluativeness judgments (*b* = .06, *SE* = .06, *t*(100.21) = .95, *p* = .344) actually predicted their participants' performance on the AUT. Judgment scale (general ability or percentile rank) did not moderate the accuracy of participants' low evaluativeness judgments (*b* = .07, *SE* = .10, *t*(99.55) = .75, *p* = .453) or their high evaluativeness judgments (*b* = -.11, *SE* = .10, *t*(99.26) = -1.17, *p* = .246). However, judgment scale did moderate the accuracy of informants' low evaluativeness judgments (*b* = -.17, *SE* = .08, *t*(98.04) = -2.05, *p* = .043) and their high evaluativeness judgments (*b* = .18, *SE* = .08, *t*(97.70) = 2.19, *p* = .031). These interactions are depicted in Figure 5 and in Figure 6. Full model results can be found in Table 9.

Overall, Study 3's results could not be used to speak directly to the research questions of interest. Study 3 was designed to further test the SOKA model's applications in metacognition by manipulating evaluativeness and examining its downstream effects on participants' and informants' metacognitive accuracy, but the evaluativeness manipulation did not produce meaningful changes in creativity's evaluativeness. Instead, participants' and informants' high evaluativeness and low evaluativeness judgments were nearly identical to one another. Unsurprisingly, evaluativeness did not predict metacognitive accuracy for participants or informants in the current study, as there was essentially no difference between these two types of judgments. Unfortunately, these results are neither consistent nor inconsistent with the SOKA model—instead, they merely indicate that evaluativeness may be difficult to manipulate. It is possible that people's preexisting beliefs about the value of cognitive abilities such as creativity require more than just a quick article to change.

The interactions between informants' judgments and scale type were surprising. In decomposing these interactions (see Figure 5 and Figure 6), informants' general ability judgments bore no relationship to participants' actual performance on the AUT. However, informants' low evaluativeness percentile rank judgments were negatively associated with participants' actual percentile ranks on the AUT, yet informants' high evaluativeness percentile rank judgments were positively associated with participants' actual percentile ranks. At present, there is no compelling theoretical reason why evaluativeness and judgment scale would only affect informants' judgments and not participants' judgments, let alone why it would only affect informants' percentile rank judgments specifically. Furthermore, it is unclear why informants' low evaluativeness percentile rank judgments would be negatively associated with participants' actual percentile ranks. These findings are discussed further in the Chapter V of this dissertation.

Study 4

As discussed earlier, one potential limitation to Study 2 was that the various domains—creativity, prospective memory, working memory, and visuospatial ability—did not differ enough in their observability for observability to serve as a useful predictor of metacognitive self- and other-judgment accuracy. Study 4 sought to address this limitation by testing a manipulation that was designed to increase creativity's observability so that future research might implement this manipulation to test the causal relationship between observability and judgment accuracy.

Participants in both the control and the high observability conditions of this manipulation were shown the same series of videos and completed the same rating procedure used in Study 1B. However, participants assigned to the high observability condition were also told that the people in those videos completed a “Just Suppose” (JS) task designed to measure their creativity by having them describe what they thought would happen if people could walk on air or fly. During the rating task after each video, high observability participants were shown these individuals’ “responses” (which were actually supplied by undergraduate research assistants). The rationale behind this manipulation was that presenting judges—in this case, participants—with an example of how the individuals they are judging (targets) performed in a creative task will make targets’ creative ability more public, thus making it easier to observe. Interrater reliability served as the measure of observability and was compared across conditions to determine if the observability manipulation had its desired effect.

As in Study 3, creativity was the only one of the four cognitive abilities from Study 2 that was assessed in Study 4. There were four main reasons for this design choice. First, the task used for the manipulation needed to tap into the same underlying construct measured by one of the four tasks used in Study 2; in this case, both the Just Suppose task and the AUT are divergent thinking tasks. Second, the task used for the manipulation needed to be open-ended in nature such that, if given information about participants’ performance, it would be difficult for informants to simply anchor to a specific number representing prior performance when making their judgments. The Just Suppose task works perfectly here, as there is no set number of items that can be answered correctly or incorrectly—performance cannot be measured and reported on a 0-100% scale that subsequent judgments could be anchored to. Third, information about performance on the task used for the manipulation needed to be meaningful to lay observers. Laypeople likely have some familiarity with creativity and what constitutes a creative as opposed to an uncreative response on a divergent thinking task. For instance, most people would likely agree that for the AUT, it would not be as creative to list "a doorstop," as one of the alternate uses for a brick as it would be to list "a ramp for my pet hamster." Fourth, the manipulation needed to have a reasonable chance to make private, internal processes more public. Again, the Just Suppose task met this requirement because responses on the Just Suppose task are not too far removed from a think aloud protocol in that they capture participants' internal thoughts to some degree (as opposed to a series of single, one-word or one-item responses), which was hoped would further boost observability. In short, although Study 1B indicated that creativity is relatively high in observability (but, importantly, not as high in observability as working memory or extraversion), creativity was used in Study 4 because the divergent thinking tasks used to measure creativity satisfy all of the criteria that an effective observability manipulation would need to satisfy.

Method

Participants

A total of 206 (129 female, 64 male) undergraduate students were recruited to participate in Study 4. Of these participants, 109 identified as Caucasian/white, 43 as Hispanic, 34 as Asian or Pacific Islander, 4 as black, and 3 as "other". Additionally, 149 of the participants were freshmen, 31 were sophomores, 8 were juniors, and 5 were seniors. These participants were randomly assigned to either the control condition (*n* = 95) or the high observability condition (*n* = 101).

Materials

All materials can be found online at https://osf.io/ny743/?view_only=b62690ae9a7b4e14857bd394dbe33cf8.

Just Suppose Task

A Just Suppose (JS) task from the Abbreviated Torrance Test for Adults (Goff, 2002), along with several sets of responses, were used for the observability manipulation. The JS task item read "Just suppose you could walk on air or fly without being in an airplane or similar vehicle. What problems might this create? List as many as you can." Several sets of responses to this item were produced by undergraduate research assistants, and both the item and the responses were provided to participants under the guise that the responses came from the individuals featured in the videos participants watched.

Videos and Ratings

The videos and subsequent rating task used in Study 1B to measure observability were also used in Study 4 to provide targets for participants to judge.

Procedure

Like the previous studies, Study 4 was conducted entirely online. Participants were given a link to a Qualtrics survey that contained all of the study materials. The survey began with the informed consent process. After providing informed consent, participants in the control condition completed the same procedure used in Study 1B. Participants in the high observability condition completed a variation of the procedure used in Study 1B which began with the following instructions:

> *"On the next few pages you will be presented with a series of videos depicting people in various contexts. We contacted the people in these videos and asked them to complete a brief creativity assessment. After watching each video, you will be shown each person's responses to this creativity assessment and asked to make several judgments about each person. Please pay close attention to each person in these videos. Because these videos were taken from the internet, some of them may include advertisements; feel free to disregard any advertisements you see as they are not a part of this study."*

The videos participants saw were the same as the ones used in Study 1B, and were presented in the same format: each was embedded from YouTube into a Qualtrics survey on its own, separate page and presented in a randomized order to participants. Participants were unable to advance to the next page until a time equal to the length of the current video they were watching had passed. After each video, participants saw a still image of an individual from the video along with that individual's name, the JS task, and a set of responses to the JS task ostensibly provided by the target from the video. Participants were instructed to review the target's responses on the JS task and to proceed to the next screen to make their ratings. When participants advanced to the next screen, the JS task and responses were replaced by the same 14 items used in Study 1B. This JS task and rating process repeated for each person in a given video before participants moved on to the next, randomly selected video. When a video contained more than one target individual, targets were reviewed and rated in a random order. Once participants had watched, reviewed, and rated every target from all of the videos they completed a brief demographics questionnaire and were debriefed.

Results

As in Study 1B, Krippendorff's alpha was used as the measure of observability for each item, and it was computed using the "KALPHA" macro for SPSS (Hayes & Krippendorff, 2007). Confidence intervals for each Krippendorff's alpha were computed using 5,000 bootstrapped samples in order to facilitate the comparison of observed observability across conditions. These results, along with the original observability values obtained in Study 1B, are presented in Table 10. Once again, alpha values were low for all items but, in the case of the control condition, their rank order tended to mirror findings from prior work (Funder & Colvin, 1988; John & Robins, 1993; Vazire, 2010) and from Study 1B. Of primary interest however was how creativity's observability differed between the control and high observability conditions. To this end, the results suggest that the observability manipulation was quite successful in increasing creativity's observability—the alpha for creativity went from being among the bottom half of all 14 items in Study 1B ($\alpha = .09$) and the control condition ($\alpha = .06$) to being second only to physical attractiveness in the high observability condition ($\alpha = .21$). In short, the observability manipulation proved capable of increasing the observability of creativity substantially, and can be used in future research to examine the causal relationship between observability and judgment accuracy. In fact, I am currently in the process of conducting said research, however it is not included in this dissertation because it is not yet complete.

CHAPTER V

DISCUSSION

The goal of the current studies was to determine to what extent the SOKA model's framework could be adapted for use with metacognitive self- and other-judgments. In doing so, the current studies focused on three related research questions. One, which cognitive abilities are relatively high or low in evaluativeness and observability? Two, what effect does evaluativeness have on the accuracy of metacognitive self- and other-judgments? And, three, what effect does observability have on the accuracy of metacognitive self- and other-judgments? I conducted a series of five studies to answer these questions. In Study 1A, participants completed a survey in which they reported how important a variety of 8 cognitive abilities, the Big Five personality factors, and physical attractiveness were to them, and in Study 1B, participants completed a similar survey in which they rated other people on these same 14 items. Study 2 was designed to test the extent to which evaluativeness and observability predicted differences in metacognitive self- and other-judgment accuracy across four cognitive abilities that differed in evaluativeness and observability. Study 3 attempted to manipulate evaluativeness to examine its downstream effects on metacognitive self- and other-judgment accuracy, and Study 4 tested the effectiveness of a manipulation designed to increase how observable creativity is for future research.

Major Findings and Implications

The results from Study 1A indicated that creativity, visuospatial ability, and processing speed were relatively low in evaluativeness, whereas working memory, prospective memory, and logical reasoning were all relatively high in evaluativeness. These findings were used to inform the design of Studies 2 and 3, but they are also interesting on their own. No published work has investigated the evaluativeness of cognitive abilities. Although some prior work has documented the role that motivational bias, such as a desire for high performance, can play in inflating metacognitive self-judgments such as grade predictions and JOLs (e.g., Helzer & Dunning, 2012; Ikeda et al., 2016; Saenz et al., 2017; Serra & DeMarree, 2016; Soderstrom & McCabe, 2011), no study has systematically examined how various cognitive tasks might differ in their tendency to elicit motivational bias from participants. Thus, in addition to informing the design of Studies 2 and 3, Study 1A also helped address this gap in the literature. The results from Study 1A suggest that cognitive abilities vary considerably in their evaluativeness. Accordingly, we now have evidence to suggest that there may be comparatively less motivational bias present in judgments of creativity, visuospatial ability, and processing speed, and comparatively more motivational bias present in judgments of working memory, prospective memory, and logical reasoning. Research in both classroom settings (e.g., Saenz et al., 2017) and laboratory settings (Saenz et al., 2019) has demonstrated the importance of reducing the influence of motivational biases to improve metacognitive accuracy and academic performance. Thus, the results from Study 1A suggest that motivation debiasing interventions may be particularly useful whenever accurate monitoring of one’s working memory, prospective memory, or logical reasoning ability is needed. Conversely, motivation debiasing interventions may be relatively ineffective when applied to low evaluativeness tasks, such as creative tasks or tasks involving visuospatial ability; in these cases, feedback-based interventions may be more appropriate if one's goal is to improve metacognitive accuracy.

To the best of my knowledge, no published study has investigated differences in observability between cognitive abilities either. In this regard, Study 1B contributes to the literature by helping to fill this gap in the literature. The results from Study 1B indicated that creativity, working memory, and attentional control are relatively high in observability and thus should be comparatively easy for others to judge. In contrast, visuospatial ability, processing speed, and prospective memory are relatively low in observability, and should be comparatively difficult for others to judge accurately. In general though, cognitive abilities may be on the low end in terms of observability, at least when compared to extraversion ($\alpha = .14$) and physical attractiveness ($\alpha = .25$). Thus, although there might be differences in how easy or difficult specific cognitive abilities are for others to observe and judge accurately, cognitive abilities in general may be relatively difficult for others to judge. This interpretation of the results is consistent with what has been seen in prior work comparing metacognitive self- and other-judgments. More specifically, when others lack any obvious clues about a target individual's cognitive abilities, such as prior performance on a similar task, metacognitive other-judgments exhibit considerable inaccuracies and often err on the side of overconfidence or overestimation (cf. Tirso & Geraci, 2020; Miller & Geraci, 2016; see also Koriat & Ackerman, 2010). In fact, there is some evidence that metacognitive other-judgments are affected by motivational biases much like self-judgments whenever others lack adequate information to base their judgments on (Tirso & Geraci, 2020). This particular finding is not altogether consistent with the SOKA model, which posits that motivational biases (i.e., evaluativeness) should only affect self-judgments. In the current dissertation, metacognitive other-judgments were consistently more optimistic than metacognitive self-judgments in both Studies 2 and 3, and, with the exception of informants' high evaluativeness percentile rank judgments in Study 3, informants' metacognitive judgments were never positively associated with participants' actual cognitive abilities. It should be noted however that Study 2 found no evidence that metacognitive other-judgment accuracy was affected by evaluativeness. Nevertheless, these findings lend credence to the idea that, without some sort of feedback or additional information, others make for poor judges of our cognitive abilities in part because cognitive abilities are difficult to observe.

Based on the results from Studies 1A and 1B, Study 2 was designed to compare the accuracy of metacognitive self- and other-judgments across four different cognitive abilities that were identified as differing in evaluativeness and observability: creativity (low evaluativeness, high observability), visuospatial ability (low evaluativeness, low observability), working memory (high evaluativeness, high observability), and prospective memory (high evaluativeness, low observability). The results from Study 2 indicated that evaluativeness moderated metacognitive self-judgment accuracy in a manner consistent with the SOKA model. That is, metacognitive self-judgments for highly evaluative cognitive abilities—prospective memory and working memory in this case—exhibited no relationship with actual performance on tasks designed to measure prospective and working memory. In contrast, metacognitive self-judgments for cognitive abilities that were low in evaluativeness—creativity and visuospatial ability—were positively associated with actual ability, meaning there was some level of accuracy present even if some participants under- or overestimated their actual ability. However, the results regarding observability were not consistent with the SOKA model. Observability did not significantly moderate metacognitive self-judgment accuracy, which was to be expected, but observability did not moderate metacognitive other-judgment accuracy either, which was inconsistent with the SOKA model. It is possible that observability simply does not affect metacognitive other-judgment accuracy. However, it is also possible that, as alluded to earlier, the cognitive abilities included in the current studies were all relatively low in observability and thus did not vary enough in observability for it to emerge as a significant moderator of judgment accuracy. This potential limitation was partially addressed by Study 4, which tested a manipulation that proved capable of substantially increasing creativity's observability. However, additional research is still needed to implement this intervention and determine if observability influenced metacognitive other-judgment accuracy. Overall, Study 2 demonstrated that at least some aspects of the SOKA model—specifically, evaluativeness—can serve as useful predictors of metacognitive accuracy, but more research is needed to determine whether observability predicts metacognitive accuracy.

Despite its limited success, Study 3 still provides some valuable information. Most notably, the results from Study 3 suggest that evaluativeness may be difficult to change within a short period of time, at least for creativity. The evaluativeness manipulation used in Study 3 did succeed in producing a statistically significant difference in creativity's evaluativeness, but this difference was quite small—it amounted to a 3-point difference on a 100-point scale. For comparison, the difference in evaluativeness between creativity and the two highly evaluative tasks in Study 2 amounted to approximately 15 points on the same 100-point scale—five times the size of the difference produced by the evaluativeness manipulation used in Study 3. In light of this result, perhaps it is not surprising that this manipulation did not affect participants' or informants' judgments. Unfortunately, this prevented any conclusions from being drawn about the potential causal relationship between evaluativeness and metacognitive accuracy, but it does inform us that evaluativeness might be resistant to change. Given that evaluativeness is negatively associated with metacognitive self-judgment accuracy as shown by Study 2, this finding suggests that attempts to improve metacognitive accuracy by reducing task evaluativeness may prove unsuccessful. Instead, it may be more fruitful to target the motivational biases that evaluativeness is likely to elicit rather than targeting evaluativeness itself. Stated differently, it may be more effective to simply warn people about the pitfalls of motivated reasoning when making metacognitive judgments about a highly evaluative task rather than trying to prevent any and all motivated reasoning from occurring. Indeed, several studies (e.g., Saenz et al., 2017, 2019) have already demonstrated how effective specifically targeting motivational biases can be in improving metacognitive accuracy, without attempting to reduce task evaluativeness.

Finally, Study 4 demonstrated that it is in fact possible to manipulate observability. Despite the novelty of its manipulation, Study 4 demonstrated that providing someone with an example of a target individual engaging in a creative task is sufficient to increase the observability of creativity for the individual in question. In fact, the results from this observability manipulation were rather impressive—it increased creativity's observability (α = .21) to the point where creativity was nearly as observable (as determined by interrater reliability) as physical attractiveness (α = .22), which was the most observable of all items measured. For comparison, the next most observable item was extraversion (α = .15), which has been identified as one of if not the most observable of the Big Five personality factors (Vazire, 2010). This finding holds considerable promise for future research seeking to investigate the potential causal relationship between observability and metacognitive other-judgment accuracy, as it provides a means of effectively manipulating observability to determine if other-judgment accuracy increases as a result.

Limitations and Future Directions

This dissertation's ambitious and exploratory nature resulted in several potential limitations that should be kept in mind at the very least and, ideally, explored in future research. The current studies relied heavily on the data obtained in Studies 1A and 1B, yet there were some limitations to these studies. Study 1A relied on self-report measures—participants rated how important various cognitive abilities and other items were to them personally. Naturally, this brings with it concerns applicable to any self-report measure, such as socially desirable responding. In particular, physical attractiveness was rated very low in evaluativeness, which appears to contrast with the importance American culture places on physical attractiveness. It is possible that this result might be indicative of socially desirable responding—participants may not have wanted to appear vain by rating physical attractiveness as important to them. This begs the question, to what extent were evaluativeness ratings for other items, including the cognitive abilities used in Studies 2-4, accurate reflections of actual evaluativeness? It should be noted that, whatever biases might have been present in evaluativeness ratings in Study 1A, evaluativeness ratings made intuitive sense—items that a layperson might associate with success or utility in daily life, such as logical reasoning, conscientiousness, and the various types of memory, were rated as the most evaluative. Furthermore, Study 2 demonstrated that evaluativeness as measured by Study 1A predicted metacognitive self-judgment accuracy in a manner consistent with the SOKA model, suggesting that the evaluativeness ratings from Study 1A were at least accurate enough to prove useful. Nevertheless, additional research might further investigate this question, perhaps by using behavioral measures of evaluativeness instead of or in addition to self-report measures.

Whereas Study 1A relied on self-report measures, Study 1B relied on a behavioral measure—participants rated target individuals depicted in short YouTube videos, and interrater reliability was used as a measure of observability. One limitation to this approach however was that interrater reliability was generally low for all items—alpha values ranged from .25 on the high end to .04 on the low end. It is not clear why interrater reliability was so low. It is possible that the limited exposure participants had to the individuals they were judging in Study 1B made it possible for participants' judgments to reach some level of agreement, but difficult for them to reach high levels of agreement. Confidence intervals were created around each estimate of interrater reliability and used to identify which cognitive abilities differed in their observability, but given that interrater reliability was abysmally low for all cognitive abilities (from .11 to .04), whether these differences in observability actually meant anything is another matter. Although working memory ($\alpha = .10$) and visuospatial ability ($\alpha = .04$) might have differed statistically in their observability, this difference might not have been large enough for it to affect metacognitive other-judgment accuracy. As alluded to earlier, this may explain why observability did not moderate metacognitive other-judgment accuracy in Study 2. However, additional research is needed to determine whether these small differences in observability reflect limitations in the way observability was measured in the current study, or the fact that cognitive abilities in general are all similarly low in observability.

Future research might also test whether increasing the observability of a given cognitive ability produces an increase in other-judgment accuracy, perhaps by implementing a manipulation similar to the one used in Study 4. Caution is advised however because although the manipulation used in Study 4 succeeded in increasing observability, it may be limited in that it was designed specifically for creativity. Part of the rationale behind the design of Study 4's manipulation was that providing informants with an example of a target individual engaging in a task designed to measure the cognitive ability of interest (i.e., creativity) would increase that cognitive ability's observability. It was assumed that for this to work, responses on this task needed to be easily interpretable by laypeople and not result in anchoring of informants' subsequent judgments. Creativity has several advantages here in that most laypeople likely have at least some familiarity with creativity—they should be able to distinguish creative ideas from uncreative ideas with at least some level of accuracy. In contrast, laypeople would likely have a harder time understanding what constitutes an impressive performance on, say, a mental rotation task unless they are already familiar with tasks involving reaction times. Thus, care should be taken when trying to adapt Study 4's observability manipulation for use with other cognitive abilities such as visuospatial ability.

Study 4 was also limited in that, while it tested the effectiveness of an observability manipulation, it did not actually implement this intervention and examine its effects on metacognitive self- and other-judgment accuracy. This prevented any strong conclusions about the role observability plays in metacognitive accuracy from being made in this dissertation. However, additional research intended to address this limitation is already in progress. Currently, I am conducting a study that implements the manipulation tested in Study 4. In this study, participants are randomly assigned to either a high observability or low observability condition. Participants in the high observability condition complete the JST and the AUT, and their responses on the JST are presented to their informants in the same way that JST responses were presented to participants in Study 4. Participants in the low observability condition do not complete the JST, just the AUT. Metacognitive self- and other-judgment accuracy will be compared across conditions to determine if the increase in observability brough on by the manipulation tested in Study 4 will have any effect on metacognitive accuracy. Currently, 200 participants have been recruited for this study and I am in the process of recruiting the informants that participants provided contact information for.

Finally, Study 3 is limited in the conclusions that can be drawn from it regarding the relationship between evaluativeness and metacognitive self-judgment accuracy. The goal of Study 3 was to create a sizable difference in evaluativeness and determine if this had any effect on metacognitive accuracy. However, the manipulation used in Study 3 did not produce a sizable difference in evaluativeness, nor any difference in the metacognitive judgments the participants and informants provided. Therefore, I could not draw any strong conclusions about the role that evaluativeness plays in metacognitive self-judgment accuracy from Study 3 alone, because essentially there was no variability in evaluativeness across judgments. Nevertheless, findings from Study 2 did suggest that evaluativeness plays a significant role in metacognitive self-judgment accuracy. Still, future research seeking to manipulate evaluativeness should use a more potent manipulation to ensure meaningful differences in evaluativeness are produced.

Conclusion

In this dissertation I set out to determine the extent to which the SOKA model can be used to predict and explain differences in metacognitive self- and other-judgment accuracy. I accomplished this goal by first determining which cognitive abilities are relatively high or low in their observability and evaluativeness in Studies 1A and 1B. In Study 2, I collected metacognitive self- and other-judgments for four different cognitive abilities that differed in their observability and evaluativeness according to the results from Studies 1A and 1B—creativity, visuospatial ability, working memory, and prospective memory. Evaluativeness predicted metacognitive self-judgment accuracy, with metacognitive judgments for cognitive abilities low in evaluativeness (creativity and visuospatial ability) exhibiting greater accuracy than judgments for cognitive abilities high in evaluativeness (working memory and prospective memory)—a finding consistent with the SOKA model. However, observability did not affect metacognitive other-judgment accuracy at all, which was inconsistent with the SOKA model. I also attempted to manipulate evaluativeness to examine its effects on metacognitive self-judgment accuracy in Study 3, but the manipulation used in Study 3 was largely unsuccessful. Study 4 tested a manipulation that successfully increased observability for creativity, and additional research in which this manipulation is implemented to determine if it affects metacognitive other-judgment accuracy is already underway. Overall, the results from this dissertation suggests that while portions of the SOKA model—specifically the relationship between evaluativeness and self-judgment accuracy—generalize to metacognitive judgments, other aspects of the model—specifically, the relationship between observability and other-judgment accuracy—do not at present seem to generalize to metacognitive judgments. Additional research is needed (and underway) to determine if this is due to the low observability of cognitive abilities in general.

REFERENCES

Al-Harthy, I. S., Was, C. A., & Hassan, A. S. (2015). Poor performers are poor predictors of performance and they know it: Can they improve their prediction accuracy. *Journal of Global Research in Education and Social Science, 4*, 93-100.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*, 126-131. doi:10.1037/h0027455

Azevedo, R. (2020). Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition and Learning, 15*, 91-98. doi:10.1007/s11409-020-09231-x

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48. doi:10.18637/jss.v067.i01

Beer, A., & Vazire, S. (2017). Evaluating the predictive validity of personality trait judgments using a naturalistic behavioral criterion: A preliminary test of the self-other knowledge asymmetry model. *Journal of Research in Personality, 70*, 107-121. doi:10.1016/j.jrp.2017.06.004

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444. doi:10.1146/annurev-psych-113011-143823

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Weinert, & R. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Erlbaum.

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin, 38,* 209-219. doi:10.1177/0146167211432763

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990. doi:10.3758/BRM.41.4.977

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology, 34,* 30-41. doi:10.1016/j.cedpsych.2008.07.001

Christensen, P., Guilford, J. P., Merrifield, P. R., & Wilson, R. C. (1960). *Alternate Uses*. Beverly Hills, CA: Sheridan Psychological Service.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136,* 1092-1122. doi:10.1037/a0021212

Dunlosky, J., & Ariel, R. (2011). The influence of agenda-based and habitual processes on item selection during study. *Journal of Experimental Psychology: Learning Memory, and Cognition, 37,* 899-912. doi:10.1037/a0023064

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271-280. doi:10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Serra, M. J., & Baker, J. M. (2007). Metamemory. In F. T. Durso, R. S. Nickerson, S. Dumais, S. Lewandowsky, & T. J. Perfect (Eds.), *Handbook of Applied Cognition* (pp. 137-160). New Jersey: John Wiley & Sons Inc.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83-87. doi:10.1111/1467-8721.01235

Educational Testing Services. (2017). GRE® Verbal and quantitative reasoning concordance tables.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98-121. doi:10.1016/j.obhdp.2007.05.002

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19-34. doi:10.1016/j.jml.2007.03.006

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906-911. doi:10.1037/0003-066X.34.10.906

Forbes, J. B., & Domm, D. R. (2004). Creativity and productivity: Resolving the conflict. *SAM Advanced Management Journal, 69*, 4-27.

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning, 12*, 1-19. doi:10.10037/s11409-016-9158-6

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 52*, 149-158. doi:10.1037/0022-3514.55.1.149

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52*, 409-418. doi:10.1037/0022-3514.52.2.409

Goff, K. (2002). *Abbreviated Torrance test for adults: Manual*. Bensenville, IL: Scholastic Testing Service

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126-134. doi:10.3758/s13423-011-0181-y

Hartwig, M. K., & Dunlosky, J. (2014). The contributions of judgment scale to the unskilled-and-unaware phenomenon: How evaluating others can exaggerate over-(and under-) confidence. *Memory & Cognition*, *42*, 164-173. doi:10.3758/s13421-013-0351-4

Hartwig, M. K., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology*, *49*, 80-90. doi:10.1016/j.cedpsych.2016.12.002

Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self-peer agreement in personality judgment. *Journal of Personality and Social Psychology, 72,* 664-677. doi:10.1037/0022-3514.72.3.664

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89.

Helzer, E. G., & Dunning, D. (2012). Why and when peer prediction is superior to self-prediction: The weight given to future aspiration versus past achievement. *Journal of Personality and Social Psychology*, *103*, 38-53. doi:10.1037/a0028124

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2019). Lab.js: A free, open, online study builder. Retrieved from https://lab.js.org

Hughes, J. (2021). reghelper: Helper functions for regression analysis. R package version 1.0.2. https://CRAN.R-project.org/package=reghelper

Ikeda, K., Yue, C. L., Murayama, K., & Castel, A. D. (2016). Achievement goals affect metacognitive judgments. *Motivation Science, 2*, 199-219. doi:10.1037/mot0000047

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*, 521-551. doi:10.1111/j.1467-6497.1993.tb00781.x

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*, 471-479. doi:10.1080/09658210802647009

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

Knowles, J. E., & Frederick, C. (2020). merTools: Tools for analyzing mixed effect regression models. R package version 0.5.2. https://CRAN.R-project.org/package=merTools

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

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and others during self-paced study. *Consciousness and Cognition, 19*, 251-264. doi:10.1016/j.concog.2009.12.010

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

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219-224.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*, 787-794. doi:10.1177/0956797611407929

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180-188. doi:10.1037/0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77,* 1121-1134. doi:10.1037/0022-3514.77.6.1121

Kruger, J., & Dunning, D. (2002). Unskilled and unaware--but why? A reply to Krueger and Mueller. *Journal of Personality and Social Psychology, 82*, 189-192. doi:10.1037/0022-3514.82.2.189

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1-26. doi:10.18637/jss.v082.i13

Long, J. A. (2019). interactions: Comprehensive, user-friendly toolkit for probing interactions. R package version 1.1.0. https://cran.r-project.org/package=interactions

Long, J. A. (2020). jtools: Analysis and presentation of social scientific data. R package version 2.1.0. https://cran.r-project.org/package=jtools

Ludeke, S. G., Weisberg, Y. J., & Deyoung, C. G. (2013). Idiographically desirable responding: Individual differences in perceived trait desirability predict overclaiming. *European Journal of Personality, 27*, 580-592. doi:10.1002/per.1914

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29*, 222-233. doi:10.3758/BF03194916

McDonough, I. M., & Gallo, D. A. (2011). Illusory expectations can affect retrieval-monitoring accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 391-404. doi:10.1037/a0025548

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

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

Miller, T. M., & Geraci, L. (2016). The influence of retrieval practice on metacognition: The contribution of analytic and non-analytic processes. *Consciousness and Cognition, 42*, 41-50. doi:10.1016/j.concog.2016.03.010

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1-12.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgements of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*, 267-271. doi:10.1111/j.1467-9280.1991.tb00147.x

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8,* 1-28.

Peronnet, F., & Farah, M. J. (1989). Mental rotation: An event-related potential study with a validated mental rotation task. *Brain and Cognition, 9*, 279-288. doi:10.1016/0278-2626(89)90037-7

Preuss, G. S., & Alicke, M. D. (2009). Everybody loves me: Self-evaluations and metaperceptions of dating popularity. *Personality and Social Psychology Bulletin, 35*, 937-950. doi:10.1177/0146167209335298

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*, 90-138. doi:1.1016/0010-0285(87)90005-3

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615-625. doi:10.1037/a0013684

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210. doi:10.1111/j.1745-6916.2006.00012.x

RStudio Team (2021). RStudio: Integrated development environment for R. RStudio, PBC, Boston, MA. http://www.rstudio.com

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology, 33*, 918-929. doi:10.1002/acp.3556

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

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

Schunk, D. & Greene, J. (2018). *Handbook of self-regulation of learning and performance* (2nd ed.). New York: Routledge.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609-612. doi:10.1016/j.jrp.2013.05.009

Scullin, M. K., Einstein, G. O., & McDaniel, M. A. (2009). Evidence for spontaneous retrieval of suspended but not finished prospective memories. *Memory & Cognition, 4*, 425-433. doi:10.3758/MC.37.4.425

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3*, 102-116. doi:10.1111/j.1745-6916.2008.00068.x

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology, 29*, 209-269. doi:10.1016/S0065-2601(08)60018-0

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology, 84*, 60-79. doi:10.1037/0022-3514.84.1.60

Selker, R., Love, J., & Dropmann, D. (2020). jmv: The 'jamovi' analyses. R package version 1.2.23. https://CRAN.R-prroject.org/package=jmv

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

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Personality and Social Psychology, 70*, 844-855. doi:10.1037/0022-3514.70.4.844

Smith, E. A. (2001). The role of tacit and explicit knowledge in the workplace. *Journal of Knowledge Management, 5*, 311-321.

Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1236-1242. doi:10.1037/a0023548

Taylor, S. E., & Brown, J. D. (1988). Illusions and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193-210. doi:10.1037/0033-2909.103.2.193

Thiede, K. W. (1999). the importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*, 662-667. doi:10.3758/BF03212976

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

Thomas, N. J. (2018). Mental imagery: Ancient imagery mnemonics. In N. Z. Edward, *The Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/spr2018/entries/mental-imagery/ancient-imagery-mnemonics.html

Tirso, R., & Geraci, L. (2020). Taking another perspective on overconfidence in cognitive ability: A comparison of self and other metacognitive judgments. *Journal of Memory and Language, 114*, 1-14. doi:10.1016/j.jml.2020.104132

Tirso, R., & Geraci, L. (2021). Unskilled and aware: Metaperception among low and high performers. Manuscript in preparation.

Tirso, R., Geraci, L., & Saenz, G. D. (2019). Examining underconfidence among high-performing students: A test of the false consensus hypothesis. *Journal of Applied Research in Memory and Cognition*, *8*, 154-165. doi:10.1016/j.jarmac.2019.04.003

Tirso, R., Saenz, G. D., & Geraci, L. (2021). Overconfidence and metacognitive accuracy: A meta-analysis. Manuscript in preparation.

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language, 95*, 124-137. doi:10.1016/j.jml.2017.03.003

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology, 71*, 673-679. doi:10.1037/h0023107

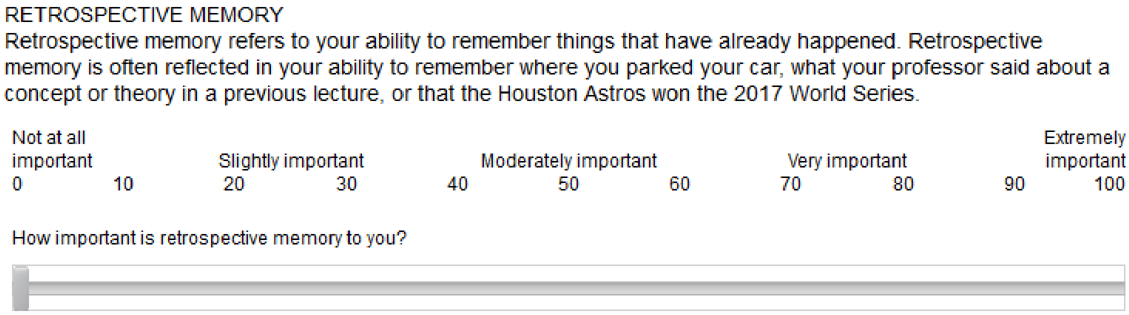
Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98,* 281-300. doi:10.1037/a0017908

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology, 95,* 1202-1216. doi:10.1037/a0013314

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20,* 568-579. doi:10.1080/09658211.2012.687052

APPENDIX A

EXAMPLE ITEM FROM STUDY 1A



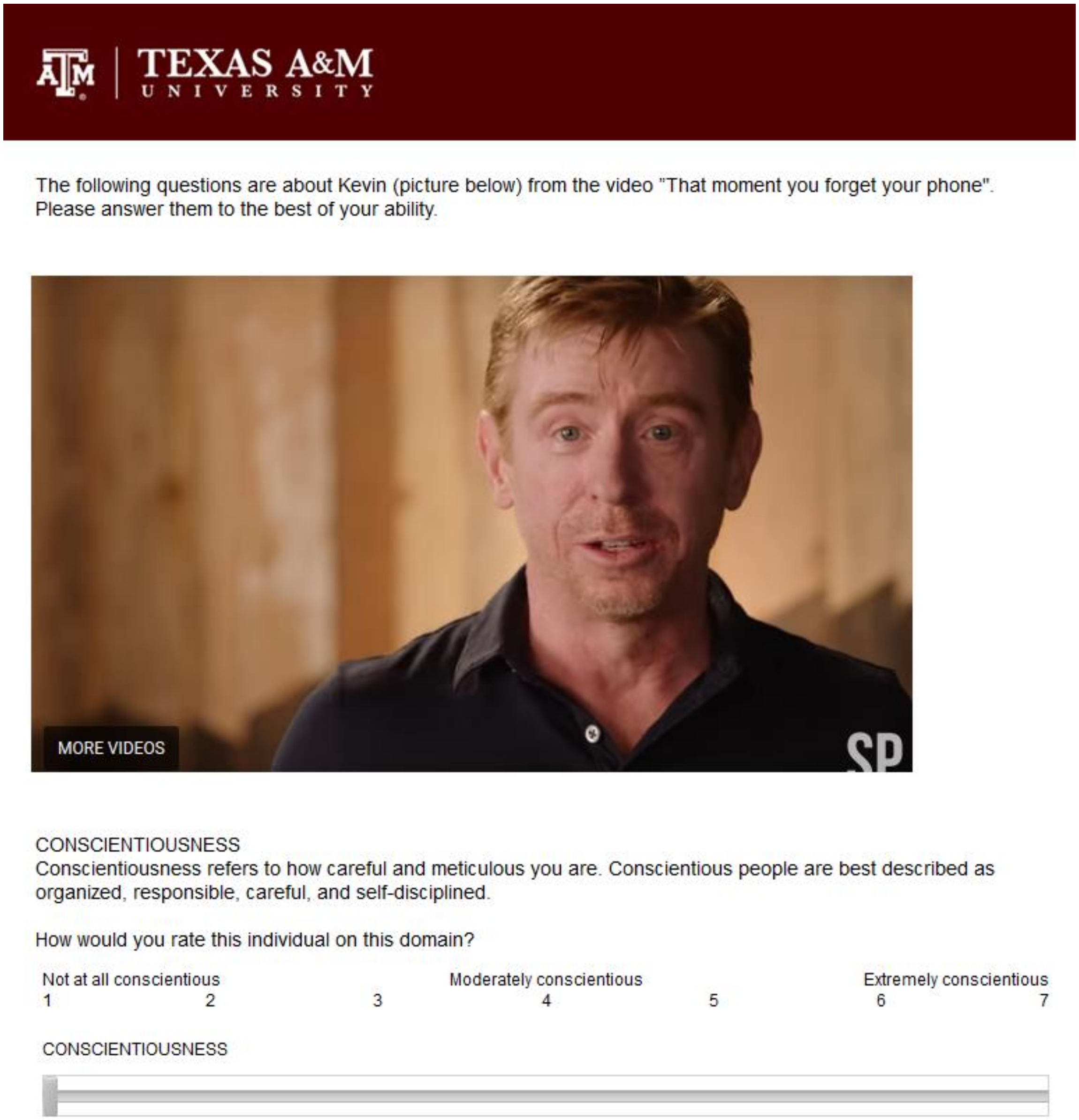
RETROSPECTIVE MEMORY

Retrospective memory refers to your ability to remember things that have already happened. Retrospective memory is often reflected in your ability to remember where you parked your car, what your professor said about a concept or theory in a previous lecture, or that the Houston Astros won the 2017 World Series.

How important is retrospective memory to you?

APPENDIX B

EXAMPLE ITEM FROM STUDY 1B



The following questions are about Kevin (picture below) from the video "That moment you forget your phone". Please answer them to the best of your ability.

CONSCIENTIOUSNESS
Conscientiousness refers to how careful and meticulous you are. Conscientious people are best described as organized, responsible, careful, and self-disciplined.

How would you rate this individual on this domain?

Not at all conscientious			Moderately conscientious			Extremely conscientious
1	2	3	4	5	6	7

CONSCIENTIOUSNESS

APPENDIX C

HIGH EVALUATIVENESS ARTICLE FROM STUDY 3

Creativity: The New Key to Success

In the modern economy, creativity is an essential quality for successful businesses Creativity provides employees with internal motivation by rewarding "thinking outside the box" in an ordinarily mundane environment. Dr. Teresa Amabile, a researcher of freedom and intrinsic motivation, claims that exploring new ideas and creating new products, services, and processes helps to foster productive internal, task-related rewards. The differences between creative and non-creative individuals are very apparent in the average workplace. Across the board, creative employees have higher levels of curiosity and drive that set them apart from the rest of their colleagues. Because of these internal motivations and higher levels of productivity, employers strongly favor more creative applicants.

APPENDIX D

LOW EVALUATIVENESS ARTICLE FROM STUDY 3

Creativity: Much Ado About Nothing

In the modern economy, creativity is not clearly advantageous or disadvantageous. Creative individuals are no better or worse than their less creative peers. According to Dr. Theresa Amabile, a researcher of freedom and intrinsic motivation, creativity is good for "internal motivation," but can distract employees with "thinking outside the box, which reduces overall productivity." Since creative work is intrinsically motivated, creative individuals can easily produce smaller amounts of quality work, but will often be reluctant to move on to the hard work of evaluating their ideas and putting them into practice. Because of this, creative workers produce greater quality work, but often end up producing less total work than other individuals. As a result, when asked to choose between highly creative individuals and non-creative individuals, employers did not favor one type of applicant over the other.

APPENDIX E

TABLES

Table 1. Mean (SD) evaluativeness ratings, Krippendorff's alpha, lower and upper bounds of the 95% CI around alpha, mean (SD) observability ratings, and mean (SD) funding allocations from Studies 1 and 2. Items are displayed from largest to smallest alpha.

Trait/Cognitive Ability	Evaluativeness	α	α LB95%CI	α UB95%CI	Observability	Funds Allocated
Physical Attractiveness	61.22 (22.53)	0.2547	0.2469	0.2629	82.25 (21.82)	$55.10 (89.95)
Extraversion	61.82 (24.21)	0.1384	0.1300	0.1471	80.96 (17.78)	$53.49 (47.05)
Retrospective Memory	71.20 (19.32)	0.1120	0.0788	0.1308	53.34 (23.95)	$53.37 (36.66)
Conscientiousness	78.18 (17.51)	0.1071	0.0985	0.116	67.71 (19.30)	$68.99 (46.34)
Attentional Control	74.18 (18.77)	0.1043	0.0955	0.1133	66.89 (21.31)	$72.24 (49.60)
Working Memory	76.63 (17.55)	0.1029	0.0857	0.1378	59.24 (23.16)	$70.60 (46.72)
Logical Reasoning	78.85 (14.94)	0.0992	0.0872	0.1414	60.62 (21.81)	$93.20 (71.72)
Creativity	61.40 (20.49)	0.0853	0.0758	0.0941	65.58 (21.89)	$77.07 (106.44)
Prospective Memory	77.22 (17.41)	0.0768	0.0450	0.0998	56.92 (25.47)	$62.04 (46.22)
Processing Speed	66.73 (20.68)	0.0766	0.0769	0.1303	65.11 (19.56)	$72.02 (57.22)
Visuospatial ability	56.03 (22.28)	0.0603	0.0312	0.0846	45.06 (25.14)	$43.15 (34.47)
Emotional Stability	79.33 (18.67)	0.0506	0.0411	0.0598	61.26 (24.86)	$127.33 (126.89)
Openness	74.35 (22.08)	0.0504	0.0278	0.0815	67.18 (21.05)	$83.86 (77.80)
Agreeableness	77.33 (17.50)	0.0432	0.0341	0.0523	74.95 (18.22)	$67.53 (51.00)

Table 2. Mean (SD) performance and predictions from Study 2.

	Participants	Informants (aggregated)
Prospective Memory		
General Ability Judgment	8.41 (2.98)	10.62 (2.61)
Percentile Rank Judgment	48.87 (21.26)	67.93 (19.36)
Actual Performance (Items Recalled)	1.80 (1.37)	-
Actual Percentile Rank	41.91 (29.78)	-
Creativity		
General Ability Judgment	8.60 (2.75)	10.97 (2.53)
Percentile Rank Judgment	48.96 (20.91)	69.87 (18.61)
Actual Performance (Fluency)	6.45 (4.19)	-
Actual Percentile Rank	52.74 (29.52)	-
Working Memory		
General Ability Judgment	8.60 (2.76)	11.20 (2.42)
Percentile Rank Judgment	50.44 (22.09)	70.06 (18.54)
Actual Performance (Trials Correct)	165.48 (16.39)	-
Actual Percentile Rank	52.68 (29.12)	-
Visuospatial Ability		
General Ability Judgment	8.43 (3.40)	10.93 (2.98)
Percentile Rank Judgment	47.13 (24.70)	70.25 (20.34)
Actual Performance (Median Correct RT)	929.32 (158.91)	-
Actual Percentile Rank	52.35 (28.71)	-

Table 3. Correlation matrix of participants' raw performance on the experimental tasks and participants' and informants' Likert scale judgments from Study 2. Bold indicates a correlation between a judgment and actual performance.

	1	2	3	4	5	6	7	8	9	10	11	12
1. PM Task Proportion Correct	-											
2. N-back Proportion Correct	0.09	-										
3. MR Task Median Correct RT	0.15	-0.02	-									
4. AUT Fluency	0.17	0.22*	0.04	-								
5. Participants' PM Judgment	**0.12**	-0.01	0.04	-0.08	-							
6. Participants' WM Judgment	0.04	**0.01**	0.00	-0.01	0.41***	-						
7. Participants' VS Judgment	0.09	0.04	**0.13**	-0.34**	0.18	0.37**	-					
8. Participants' Creativity Judgment	-0.05	0.03	-0.09	**0.15**	0.19~	0.18~	-0.1	-				
9. Informants' PM Judgment	**0.15**	-0.17	-0.09	-0.15	0.17	0.02	-0.08	-0.08	-			
10. Informants' WM Judgment	0.23*	**0.08**	-0.01	-0.08	-0.01	0.02	0.24*	0.17~	0.21*	-		
11. Informants' VS Judgment	-0.08	0.05	**-0.11**	-0.02	-0.09	0.12	0.13	0.34***	0.20*	0.51***	-	
12. Informants' Creativity Judgment	0.15	0.13	0.03	**-0.06**	0.05	0.16	0.09	0.00	0.44***	0.35***	0.30**	-

~*p* < .10, **p* < .05, ***p* < .01, ****p* < .001

Table 4. Correlation matrix of participants' percentile ranks on the experimental tasks and participants' and informants' percentile rank judgments from Study 2. Bold indicates a correlation between a judgment and actual performance.

	1	2	3	4	5	6	7	8	9	10	11	12
1. PM Task Percentile Rank	-											
2. N-back Percentile Rank	0.13	-										
3. MR Task Percentile Rank	0.09	0.14	-									
4. AUT Percentile Rank	0.16	0.32**	0.01	-								
5. Participants' PM Judgment	**0.03**	0.06	0.06	0.02	-							
6. Participants' WM Judgment	0.10	**-0.03**	0.09	0.04	0.52***	-						
7. Participants' VS Judgment	0.26*	0.19	**0.16**	-0.20~	0.30**	0.55***	-					
8. Participants' Creativity Judgment	0.04	0.03	-0.03	**0.17**	0.34***	0.39***	0.16	-				
9. Informants' PM Judgment	**0.16**	-0.06	-0.07	-0.09	-0.03	-0.12	-0.07	-0.06	-			
10. Informants' WM Judgment	0.18~	**0.05**	0.09	-0.09	0.04	0.10	0.25*	0.00	0.43***	-		
11. Informants' VS Judgment	0.05	0.13	**0.08**	-0.02	-0.03	-0.08	0.19	0.17~	0.42***	0.55***	-	
12. Informants' Creativity Judgment	0.16	0.08	-0.02	**-0.07**	-0.08	-0.08	0.12	-0.16~	0.61***	0.59***	0.38***	-

~p < .10, *p < .05, **p < .01, ***p < .001

Table 5. Model 1.x results from Study 2. Standard errors are in parentheses. Asterisks indicate the significance level of effects and improvement in overall model fit compared to the preceding model.

Model 1.x Results				
Fixed Effects	Model 1.0	Model 1.1a	Model 1.1b	Model 1.2
Intercept	-0.01 (0.06)	-0.01 (0.06)	-0.01 (0.06)	-0.01 (0.10)
Participants' Judgments	-	0.02 (0.02)	0.02 (0.02)	0.10 (0.04)*
Informants' Judgments	-	0.01 (0.02)	0.01 (0.02)	0.05 (0.04)
Judgment Scale	-	-	0.00 (0.02)	0.00 (0.02)
Participants' Judgments * Scale (Likert vs. percentile)	-	-	-0.01 (0.02)	-0.01 (0.02)
Informants' Judgments * Scale (Likert vs. percentile)	-	-	0.00 (0.02)	0.00 (0.02)
Task Evaluativeness (low vs. high)	-	-	-	0.00 (0.10)
Participants' Judgments * Evaluativeness (low vs. high)	-	-	-	-0.10 (0.05)*
Informants' Judgments * Evaluativeness (low vs. high)	-	-	-	-0.02 (0.04)
Task Observability (low vs. high)	-	-	-	0.00 (0.10)
Participants' Judgments * Observability (low vs. high)	-	-	-	-0.04 (0.05)
Informants' Judgments * Observability (low vs. high)	-	-	-	-0.06 (0.04)

~*p* < .10, **p* < .05, ***p* < .01, ****p* < .001

Table 6. Mean (SD) predictions, performance, evaluativeness ratings, and manipulation check items from Study 3.

	Participants	Informants (aggregated)
Following HE Article		
General Ability Judgment	8.92 (2.50)	10.93 (2.13)
Percentile Rank Judgment	57.76 (19.33)	76.56 (15.43)
Evaluativeness Rating	71.14 (19.15)	80.12 (13.58)
Following LE Article		
General Ability Judgment	8.75 (2.56)	11.03 (2.13)
Percentile Rank Judgment	57.79 (19.66)	76.21 (15.06)
Evaluativeness Rating	68.23 (20.37)	76.08 (16.12)
Performance		
AUT Fluency	6.89 (3.72)	-
Percentile Rank	51.10 (28.45)	-
Manipulation Checks		
1st Article Believability	3.68 (0.94)	3.64 (0.84)
2nd Article Believability	3.94 (0.95)	3.86 (0.80)

Table 7. Correlation matrix of participants' actual performance (fluency) on the AUT and participants' and informants' general ability judgments from Study 3. Bold indicates a correlation between a judgment and actual performance.

	1	2	3	4	5
1. Participants' LE Judgment	-				
2. Participants' HE Judgment	0.89***	-			
3. Informants' LE Judgment	0.31**	0.31**	-		
4. Informants' HE Judgment	0.21*	0.18~	0.87***	-	
5. AUT Fluency	**0.17~**	**0.19~**	**-0.08**	**-0.06**	**-**

~*p* < .10, **p* < .05, ***p* < .01, ****p* < .001

Table 8. Correlation matrix of participants' actual percentile ranks on the AUT and participants' and informants' percentile rank judgments from Study 3. Bold indicates a correlation between a judgment and actual performance.

	1	2	3	4	5
1. Participants' LE Judgment	-				
2. Participants' HE Judgment	0.92***	-			
3. Informants' LE Judgment	0.03	0.06	-		
4. Informants' HE Judgment	0.00	0.01	0.91***	-	
5. AUT Percentile Rank	**-0.02**	**-0.01**	**-0.07**	**-0.12**	-

~*p* < .10, **p* < .05, ***p* < .01, ****p* < .001

Table 9. Model 2.x results from Study 3. Standard errors are in parentheses. Asterisks indicate the significance level of effects and improvement in overall model fit compared to the preceding model.

Model 2.x Results			
Fixed Effects	Model 2.0	Model 2.1a	Model 2.1b
Intercept	0.00 (0.10)	0.00 (0.10)	0.00 (0.10)
Participants' LE Judgments	-	-0.02 (0.07)	-0.05 (0.06)
Participants' HE Judgments	-	0.04 (0.07)	0.05 (0.07)
Judgment Scale (Likert vs. percentile)	-	0.00 (0.03)	0.00 (0.03)
Ps' LE Judgments * Scale (Likert vs. percentile)	-	0.07 (0.10)	0.07 (0.10)
Ps' HE Judgments * Scale (Likert vs. percentile)	-	-0.13 (0.10)	-0.11 (0.10)
Informants' LE Judgments	-	-	-0.02 (0.07)
Informants' HE Judgments	-	-	0.06 (0.06)
Is' LE Judgments * Scale (Likert vs. percentile)	-	-	-0.17 (0.08)*
Is' HE Judgments * Scale (Likert vs. percentile)	-	-	0.18 (0.08)*

~*p* < .10, **p* < .05, ***p* < .01, ****p* < .001

Table 10. Krippendorff's alpha and the lower and upper bounds of the 95% CI around alpha for all items from Studies 1B and 4. Items are displayed from largest to smallest alpha based on values originally calculated from Study 1B.

	Original (Study 1B)			Control (Study 4)			High Observability (Study 4)		
Trait/Cognitive Ability	kalpha	lower	upper	kalpha	lower	upper	kalpha	lower	upper
Physical attractiveness	0.2547	0.2469	0.2629	0.1631	0.1555	0.1726	0.2174	0.2094	0.2246
Extraversion	0.1384	0.1300	0.1471	0.0802	0.0717	0.0893	0.1485	0.1406	0.1559
Retrospective memory	0.1120	0.0788	0.1308	0.0822	0.0737	0.0910	0.1224	0.1146	0.1302
Conscientiousness	0.1071	0.0985	0.1160	0.0886	0.0796	0.0974	0.1249	0.1169	0.1325
Attentional control	0.1043	0.0955	0.1133	0.0638	0.0549	0.0728	0.1231	0.1153	0.1311
Working memory	0.1029	0.0857	0.1378	0.0740	0.0650	0.0829	0.1076	0.1000	0.1156
Logical reasoning	0.0992	0.0872	0.1414	0.0674	0.0586	0.0762	0.0607	0.0525	0.0690
Creativity	**0.0853**	0.0758	0.0941	**0.0591**	0.0501	0.0680	**0.2056**	0.1982	0.2125
Prospective memory	0.0768	0.0450	0.0998	0.0907	0.0823	0.0994	0.0910	0.0832	0.0988
Processing speed	0.0766	0.0769	0.1303	0.0490	0.0398	0.0584	0.1158	0.1081	0.1237
Visuospatial ability	0.0603	0.0312	0.0846	0.0333	0.0239	0.0422	0.0608	0.0527	0.0688
Emotional stability	0.0506	0.0411	0.0598	0.0349	0.0259	0.0438	0.0551	0.0468	0.0632
Openness	0.0504	0.0278	0.0815	0.0301	0.0209	0.0387	0.0975	0.0891	0.1052
Agreeablenesss	0.0432	0.0341	0.0523	0.0217	0.0122	0.0308	0.0680	0.0599	0.0762

APPENDIX F

FIGURES

Figure 1. The effect of evaluativeness on participants' judgment resolution in Study 2.

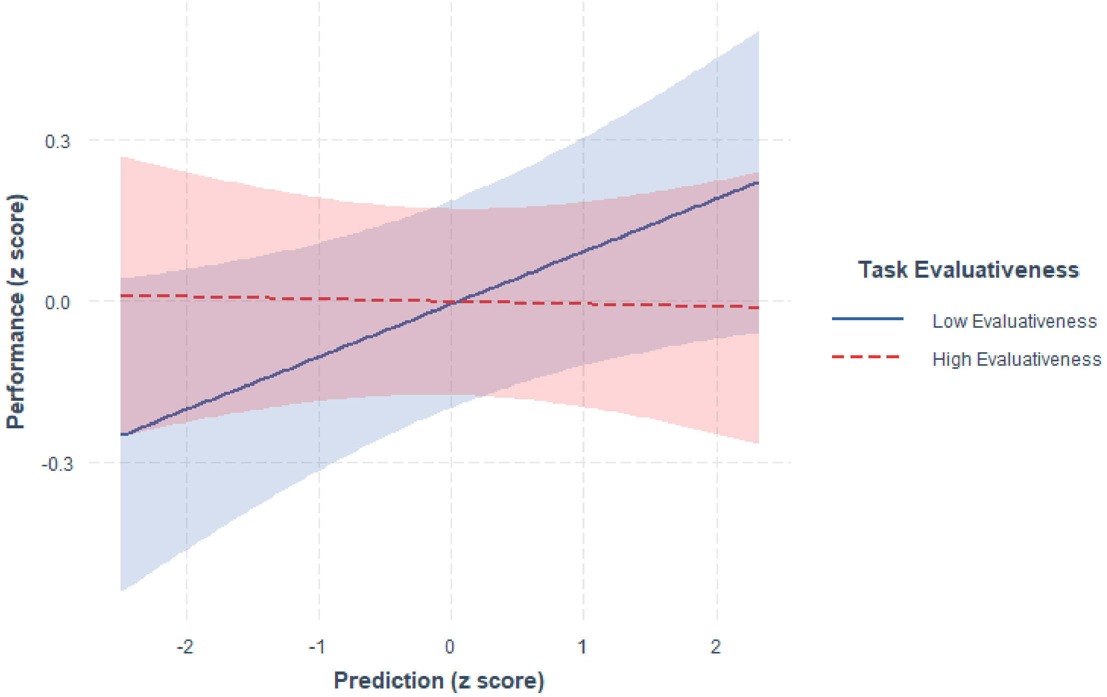


Figure 2. The effects of observability on participants' judgment resolution in Study 2.

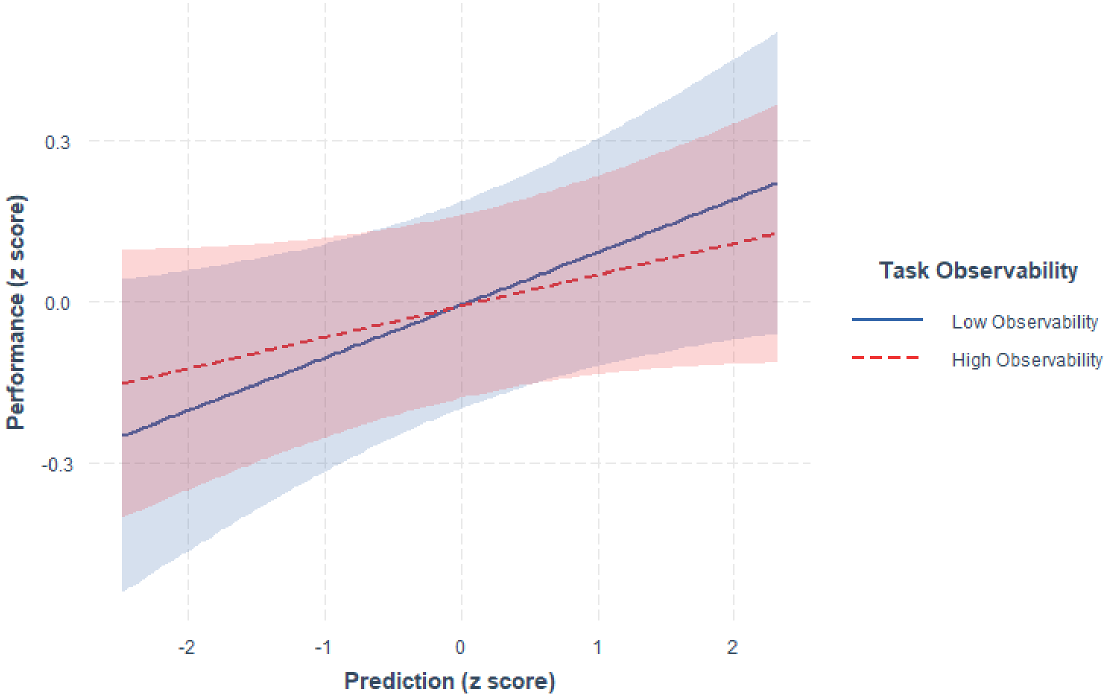


Figure 3. The effects of evaluativeness on informants' judgment resolution in Study 2.

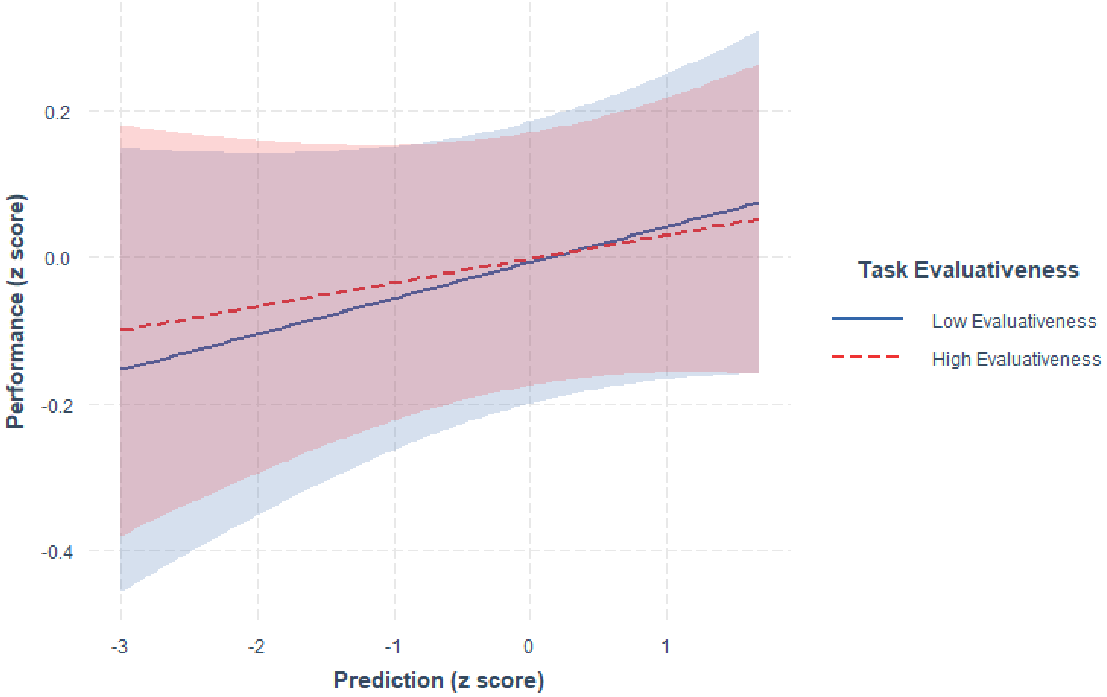


Figure 4. The effects of observability on informants' judgment resolution in Study 2.

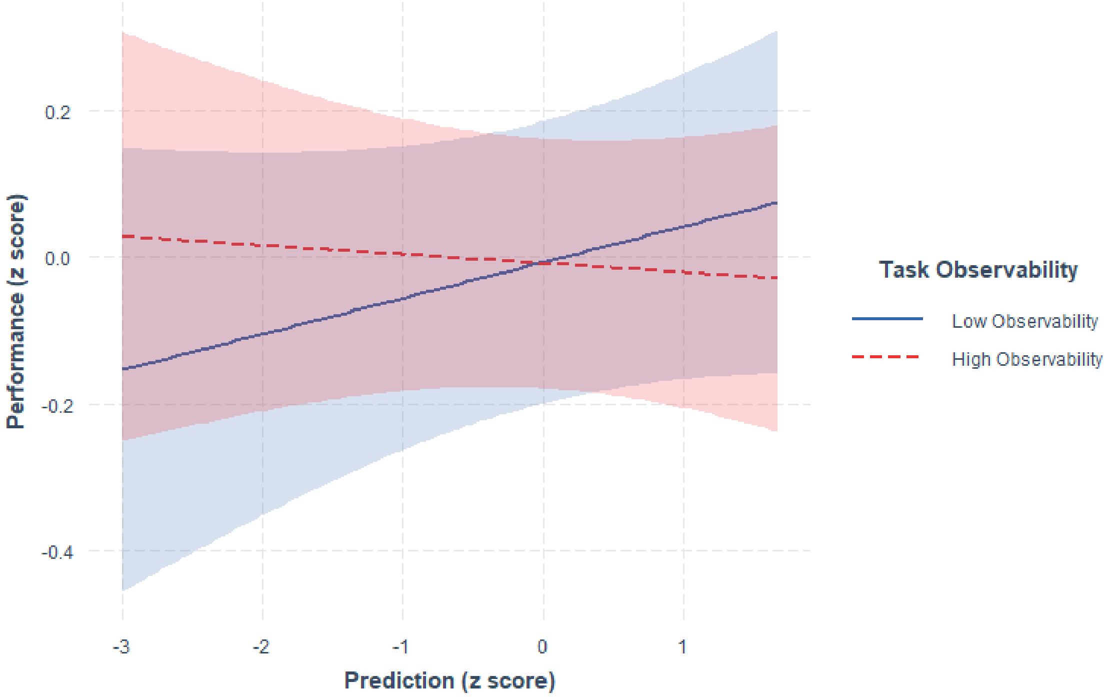


Figure 5. The effects of judgment scale on informants' judgment resolution for their low evaluativeness judgments in Study 3.

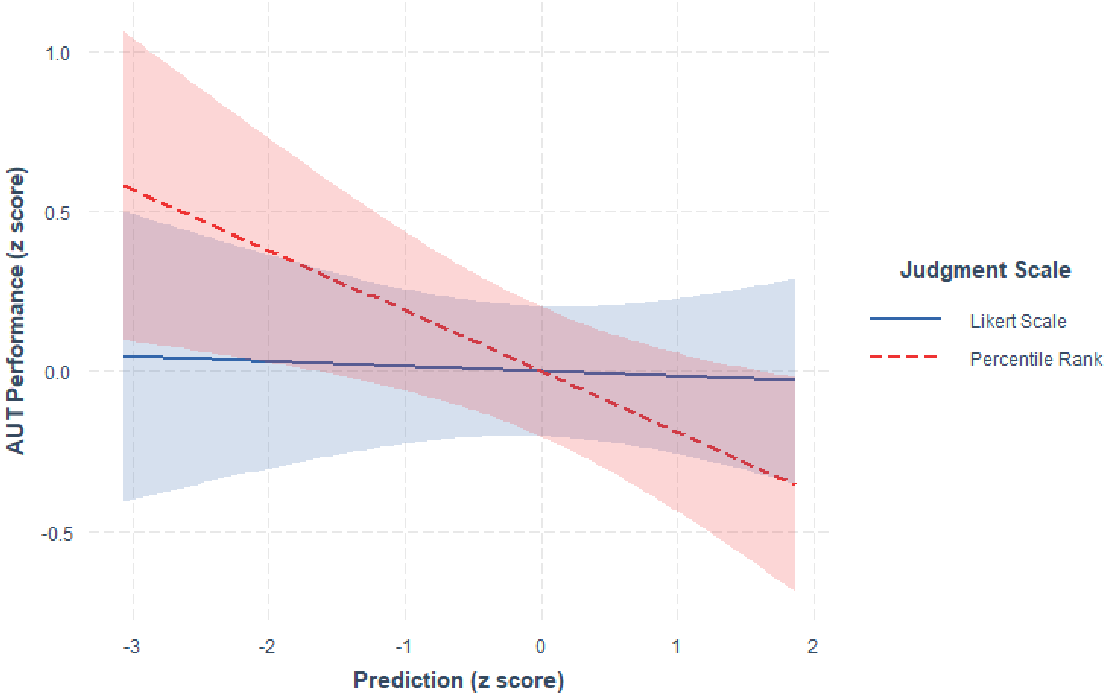
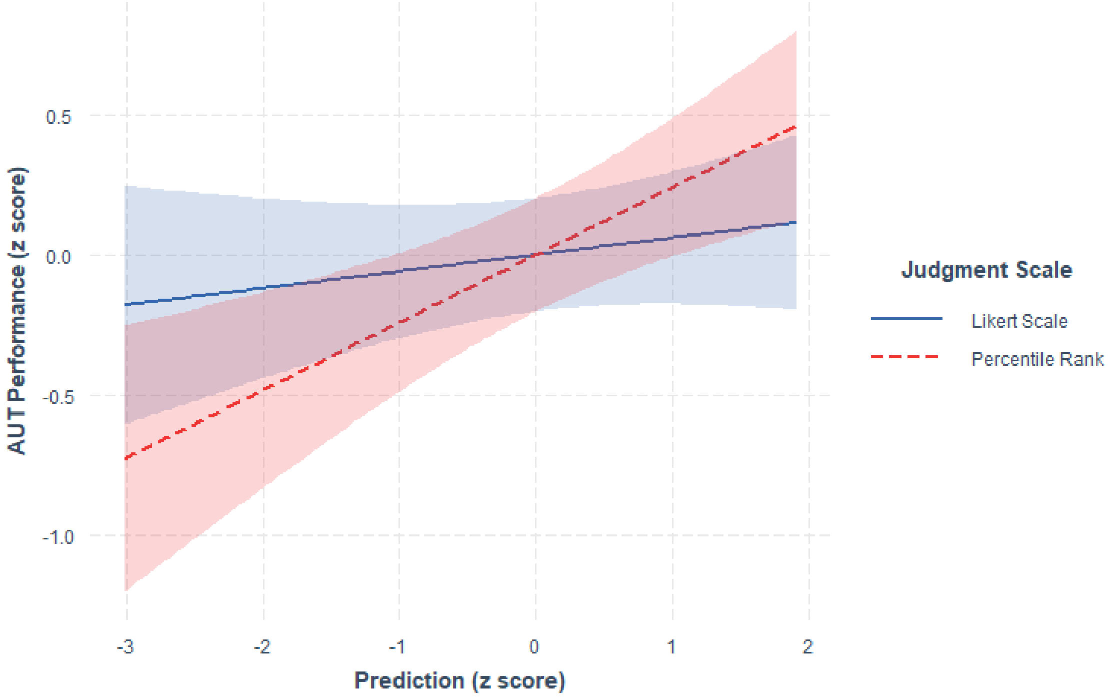


Figure 6. The effects of judgment scale on informants' judgment resolution for their high evaluativeness judgments in Study 3.


AUT Performance (z score)
Prediction (z score)
Judgment Scale
Likert Scale
Percentile Rank