SHIFTING FOCUS: A NEW THEORY EXPLAINING HARMFUL OVERCONFIDENCE IN STUDENTS

A Dissertation

by

GABRIEL DIEGO SAENZ

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Steven Smith
Committee Members,	Lisa Geraci
	Jessica Bernard
	Jeffrey Liew
Head of Department,	Heather Lench

August 2019

Major Subject: Psychology

ABSTRACT

Students are often overconfident (or otherwise metacognitively inaccurate) about how they will perform on exams, a condition that can have negative consequences for students as they may stop studying prematurely and perform poorly on tests. Interventions designed to improve student metacognition have had mixed results, and poor transference, potentially because researchers do not completely understand *why* students have poor metacognition. Preliminary data show one reason why interventions might not transfer to later tests, students become less confident after taking tests, but regain their confidence over just 10 minutes. Because no information was introduced during those 10 minutes, participants must have changed the way in which they were thinking about their past and/or future test performance, a "Shifting Focus" Effect. I propose that students' confidence follows a similar Shifting Focus pattern between class exams, potentially because motivations erode rational metacognitive judgements over time. The current dissertation was designed to accomplish two goals: to replicate the Shifting Focus effect in classroom and laboratory conditions, and to investigate its causes as well as test ameliorative interventions. Experiments 1-3 replicated preliminary findings suggesting a "Shifting Focus" Effect across a variety of conditions, and although my prior research indicates that students are *motivated* to think positively about their future, evidence connecting motivations to rising confidence was inconclusive. Experiments 4-5 tested interventions designed to prevent the Shifting Focus Effect. Results indicate that rising overconfidence in students may be prevented through minimalistic interventions in normal classroom settings, and may even improve test grades.

DEDICATION

For the people I love and the people who love me – you are always my motivation. I’ve learned so much in the last five years, and I might get a PhD too. Gracias and Thank you for everything, and your patience doubly so. Finally, this is for students, past and future, whom I hope to help at least a little bit.

ACKNOWLEDGMENTS

I have too many people to thank. First, thank you so much to my advisors and Drs. Lisa Geraci and Steven Smith. You saw value in me as a student, and you have been invaluable to me as mentors. Of course, I would also like to thank committee members Drs. Jessica Bernard and Jeffrey Liew, your thoughtful comments have and will enrich my research.

A most heartfelt thank you to the many research assistants that made this possible: Patricia Najera, Kyna Westbury, Casey Norton, Kristen Rinderknecht, and many, many more.

To the many friends and colleagues I met and worked with along the way, thank you: Robert Tirso, Anton Leontyev, Omar Garcia, and everyone of the PDOGS.

I would also like to thank the faculty and staff at the Psychology Department of Texas A&M, who have and do help me and others in innumerable ways.

Finally, my unwavering gratitude goes out to my parents, Laura Samperio and Gabriel David Saenz, for the security, help, invaluable education, and wondrous opportunities they have afforded me. And to Suzie Swain, who kept me sane.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Steven Smith, Lisa Geraci, and Jessica Bernard of the Department of Psychology and Professor Jeffrey Liew of the Department of Education and Human Development.

The data analyzed for Experiments 3 and 5 was obtained with help from Drs. Bernard and Geraci respectively. Drs. Steven Smith and Lisa Geraci contributed greatly to the framing and theoretical interpretations in this dissertation. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship and teaching positions from Texas A&M University. This work was also made possible in part by the Graduate Student Fellowship Application Support Program, from the Psychology Department at Texas A&M University.

TABLE OF CONTENTS

Page

CHAPTER I

INTRODUCTION

Imagine the following anecdote: A student visits their professor, expounding surprise at having received a low test grade. Upon review, the student had received similarly low grades on multiple previous tests, but reports honestly believing they would have performed better this time around. Without evidence of having studied more effectively, it seems the student has no reason to have gained confidence after past failures, so why do students seem to exhibit overconfidence so often? It seems reasonable that students *want* to perform well on tests, so it may be that when estimating their knowledge of course material, and future test performance, their desire could lead them to be more confident than they should be. Even after experiencing failure, students could be *shifting the focus* of their self-evaluations from past experiences to desired grades, leading to overconfidence.

The ability to monitor one's own learning (referred to as *metacognition)* is essential for students, because they must be able to accurately determine their preparedness for exams (Pintrich, Wolters, & Bexter, 2000; Winne, 2011). Ample evidence shows that good metacognition and good performance go hand in hand. For example, studies have shown that students with higher GPAs or SAT scores are also better metacognitive-monitors than students with lower scores (Everson & Tobias, 1998; Kelemen, Winningham, & Weaver, 2007; Shepperd, 1993) and that higher performing students are also better predictors of their own future performance (Bol, Hacker, O'Shea, & Allen, 2005; Kruger, & Dunning, 1999; Saenz, Geraci, Miller, & Tirso, 2017) than are lower-performing students.

Unfortunately, poor metacognition is rampant in self-evaluations as people generally make overconfident self-evaluations (Bol, et al., 2005; Dunlosky & Rawson, 2012; Foster, Was, Dunlosky, & Isaacson, 2017; Hacker, Bol, & Bahbahani, 2008; Hartwig & Dunlosky, 2014; Kelemen et al., 2007; Kruger & Dunning, 1999; Miller & Geraci, 2011a, 2011b, 2014; Nietfeld, Cao, & Osborne, 2005, 2006; Rawson, O'Neil, & Dunlosky, 2011; Saenz et al., 2017; Szpunar, Jing, & Schacter, 2014). Students are often particularly poor monitors of their classroom knowledge, and generally exhibit poor calibration (the difference between performance predictions and grades). For example, when asked, "What grade do you think you will get on this exam?" students are frequently five, ten, even twenty-five grade points overconfident in their predictions (Saenz et al., 2017). Furthermore, research suggests that overconfident students may not study enough because higher memory confidence is related to lower study choice (Dunlosky & Hertzog, 1998; Metcalfe, 2002; Metcalfe & Finn, 2008b).

Because better metacognition is associated with better academic performance, a number of investigators have designed interventions for improving metacognition in students. Regrettably, interventions to improve metacognition in the classroom are often ineffective, with many interventions failing to improve metacognitive accuracy and/or test performance (Hacker, Bol, & Keener, 2008; Nietfeld et al., 2005; Foster et al., 2017). Whereas there are some successful interventions to improve metacognitive accuracy (Miller & Geraci, 2011b; Nietfeld et al., 2006), these have not focused on transfer of metacognitive improvement. Furthermore, calibration does not naturally improve in students over time. Some speculate that students may "calibrate" their self-estimations over the course of a university semester by gathering feedback about their previous test performance and the types of tests a teacher writes (See also Pierce & Smith, 2001). However, calibration is resistant to incremental improvement, even with extensive feedback and metacognitive training (Nietfeld et al., 2005). Figure 1 demonstrates calibration over a university course semester during a study in which students took 13 tests, made predictions about their performance, and received feedback again and again. Students were unable to improve their calibration over the course of the semester despite the extensive prediction and feedback schedule (Foster et al., 2017).

Why do so many interventions fail to produce improvements in calibration? Intuitively, calibration should improve with feedback. To calibrate a watch, one checks the time from a reliable source and makes an appropriate adjustment to the watch's time, if there is a discrepancy. If such an adjustment cannot be made, it is not uncommon for people to use consistently inaccurate clocks by making appropriate mental adjustments (e.g., my watch is usually 5 minutes fast, so 2:35 is actually 2:30). In a similar way, it seems as though students should be able to improve the accuracy of their predictions over time by making educated guesses about their future performance, and adjusting their estimates based on feedback (i.e., actual test grades). However, as outlined above (Figure 1), student's calibration does not always improve with feedback, even when students are explicitly instructed to use the feedback to improve their predictions. This outcome is particularly strange as (anecdotally) students often report feeling more sure about their test grades and predictions as a university semester progresses, because they learn more about how an instructor teaches, and how tests are designed (See also Putwain & Sander, 2016). So if students do not improve their calibration between tests, and across university semesters, what *does* happen to their metacognitive thoughts during this time?

Metacognition Between Tests

The accuracy of self-evaluations is affected by the time in which they are made (Glenberg & Epstein, 1985; Hacker, Bol, Horgan, Rakow, 2000; Maki & Serra, 1992). Sometimes called the “postdiction superiority effect” (Pierce & Smith, 2001), postdictions (performance predictions made after taking a test, but before receiving the grade) are generally less confident, and more accurate than predictions. This finding makes intuitive sense, as people take tests, they may learn about their own level of knowledge, as well as the test difficulty and content (Devolder, Brigham, & Pressley, 1990; Lin, Moore, Zabrucky, 2001; Maki, 1998; Maki, Jonas, & Kallod, 1994; Pierce & Smith, 2001). Anecdotally, many students report feeling less confidence in their performance after taking tests (sometimes from, “I think I’ll do well” to, “I think I failed”) relative to before the test.

If calibration improves after testing, why does that metacognitive learning not transfer to later tests in a semester? Preliminary evidence demonstrates that grade predictions can become more confident even after just 10 minutes of time after completing a test and postdiction. These data stem from a study that compared the efficacy of five different metacognition-intervention designs (Saenz, Geraci, & Tirso, 2019). Participants completed two tests, made predictions and postdictions about each, and completed one 10-minute intervention after the first test's postdiction, and before the second test's prediction. Importantly, participants were clearly informed that the two tests were taken from the same question bank, and were of similar difficulty. In the control condition, participants completed maze puzzles instead of an intervention. Predictions decreased significantly from before to after the first exam, and stayed stable afterwards, this outcome differs from the mostly-invariant confidence students produce in actual classes; students learned about their performance, and they expressed that metacognitive knowledge before and after the second test. Alternatively, one of the intervention conditions appeared detrimental to predictive accuracy (See Figure 2; Saenz et al., 2019). Between the two prediction-test-postdiction cycles, participants made grade predictions once a minute, for ten minutes. This intervention was designed to make people think about predictions for an extended amount of time so that might improve their accuracy. Instead, grade predictions became steadily more confident, and less accurate, during the ten-minute intervention, an outcome that might mirror what happens to student's metacognitive confidence in an actual classroom! Confidence zig-zagged, increasing between tests, and decreasing after tests, without any changes to available metacognitive information, suggesting that participants *shifted the focus* of their metacognitive judgements away from feedback to some other thought process.

Why did confidence fluctuate in the repeated predictions condition, but stay flatly improved in the control condition? Both conditions involved improved calibration after test 1, suggesting that students improved the accuracy of their predictions, most likely through a form of self-feedback. Whereas this improvement did not happen twice in the flat control condition, it did re-appear in the repeated predictions condition after the second exam (Figure 2). Something about making repeated grade predictions must have increased metacognitive confidence. No new information was introduced, and participants didn’t have the chance to study (as students might between real tests), so it seems likely that the repeated predictions intervention changed the way participants thought about their past and future performance, a shifting focus effect. It may be that thinking about their future test performance made students “forget” their self-feedback from the earlier test, but this explanation seems unlikely as participants lowered their confidence from the last repeated prediction, to the second test’s prediction, potentially refocusing on the information they had learned earlier.

It may be that participants in the repeated predictions condition somehow rationalized their past performance and, in contemplating their future, allowed themselves to become more confident in the same way a person might regain their confidence before re-attempting a failed task. If such a motivational explanation is valid, the preliminary data might evidence that people have some metacognitive control over their motivated optimism. Although predictions became more optimistic during the time between tests (hereafter referred to as the *reflection period*), they began to decrease during the final reflection period prediction, and decreased to a similar extent between the prediction and postdiction for the second exam (Figure 2). These decreases in confidence may indicate that people are metacognitively wary of the optimism they are putting out, or at least that they face some form of anxiety as the final test approaches, lowering their confidence. There is some evidence for students lowering their grade predictions as time of feedback approaches, and that lowered predictions were associated to affect and anxiety (Taylor & Shepperd, 1998). This motivational explanation might also hint at the *way* in which confidence increased, it may be that *time* was not the most predictive factor of increasing optimism in metacognition as the control condition involved the same 10 minute period between tests, but did not result in degraded calibration. Instead, it is possible that as participants thought about their predictions, their motivations lead them to become more confident in their future performance. These speculations might be tested by varying the length of time, amount of predictions, and lab vs. classroom setting of the reflection period. Speculation aside, if this outcome can be replicated in classroom conditions, it may help explain why students don’t become better calibrated over time, why some interventions work, and others don’t.

Why Don’t Metacognitive Interventions Always Work?

Interventions may be targeting the wrong mental processes as there yet has been no comprehensive theory describing how students in the classroom become overconfident. To devise an effective classroom metacognition intervention, researchers need to understand the underlying processes that cause inaccurate metacognition, and then develop and test interventions that target these processes.

What might be the processes underlying students' poor metacognition? Some accounts (i.e., the Dunning-Kruger effect or the Unskilled and unaware effect, Kruger & Dunning, 1999) suggest that lower performing students may simply have a harder metacognitive task than others, because having less knowledge about a subject makes it harder to guess how much is missing[1]. Unfortunately this explanation leads to a circular argument: students don't know enough material so they are overconfident about their knowledge – overconfident students think they know enough material so they don't study, making them perform poorly. Another popular explanation might come from social psychology, where the better-than-average effect has been studied extensively. This effect states that most people identify themselves to be better than average when comparing their abilities to others (Alike & Govorun, 2005; Taylor & Brown, 1988; Sedikides & Gregg, 2003). Most explanations of this effect have focused on motivational accounts, suggesting that people focus on their own successes, and other people's failures because doing so feels good (Klein & Kunda, 1993, 1994; Klein & Weinstein, 1997; Kunda, 1987; Middleton, Harris, & Surman, 1996; Regan, Snyder, & Kassin, 1995; Taylor & Brown, 1988; Taylor, Wayment, & Collins, 1993; Weinstein & Klein, 1996).

[1] Note that this varies slightly from the more common understanding of the Dunning-Kruger effect, which seems to be commonly miss-conceived. Many people describe a Dunning-Kruger effect to occur in the entirety of a person, i.e. a person is generally incompetent and therefore generally overconfident. Instead, the Dunning-Kruger actually describes how a person can be ignorant in particular areas, and therefore overconfident in some things, but more accurate and competent in others.
See http://gabrielsaenz.com/unskilled-and-unaware-about-the-unskilled-and-unaware-effect for a list of popular-media misconception examples

Clearly, students may be motivated to be optimistic about their future test performance, and avoid contemplating imminent test failure. Beyond evidence from social psychology, mounting evidence suggests that students' grade predictions are strongly influenced by *motivational factors*, in a number of ways. For example, lower performing, overconfident students externalize explanations of their poor grades, ascribing these undesirable outcomes to issues outside of their control, such as unreasonably difficult questions, or poor instructors, while higher performing students attribute their success to themselves (Hacker & Bol, 2004; Bol et al., 2005; Hacker et al., 2008a, 2008b). Other researchers have found more direct connections between motivational information and grade predictions, indicating that students use information such as the *ideal* grade they would like to get on a test to influence their grade predictions above and beyond educational information (such as prior test performance or study habits; Saenz et al., 2017; Serra & DeMarree, 2016; See also the "better than myself" effect: Alicke, Vredenburg, Hiatt, & Govorun, 2001). My own research shows that interventions swaying students away from using motivational information to make grade predictions may improve metacognition (Saenz et al., 2017).

So how exactly does motivation information sway students towards overconfidence? I propose that students' motivations erode rational metacognitive judgements over time as students make optimistic self-judgements to avoid thinking about possible negative outcomes. During the time between tests, students may become less metacognitively accurate because they *shift the focus* of their predictions from realistic academic information to optimistic motivated information.

People are known to avoid thinking about unwanted information (Golman, Hagmann, & Loewenstein, 2017), and are motivated to favor optimistic thoughts (Lench & Bench, 2012). I propose that students rely on these same thought processes and that their confidence grows without feedback to replace more rational assessments of their knowledge. This bias may be particularly detrimental for low-performing students, who stand to benefit the most from more accurate assessments of knowledge, which could lead them to increase their studying behavior.

This explanation can tie together some of the apparently disparate findings in educational metacognitive research. The Postdiction Superiority Effect (Pierce & Smith, 2001) states that performance postdictions are more accurate than predictions, so students should learn and become more calibrated as they take university tests, but there is evidence that this does not happen (i.e., Foster et al, 2017). How could students take so many tests and not become more calibrated over time? I propose that students regain confidence between tests in the same way that participants regained confidence in my preliminary data, through some sort of *Shifting Focus* Phenomena, where students change the basis of information they use to make grade predictions. Under this Shifting Focus hypothesis, students *do* become more calibrated after taking tests, but become less calibrated by the time of the next test, possibly because they change the way they make their grade predictions. During the time between tests, students may shift away from making metacognitive judgements based on what they learned by taking a test, towards making metacognitive judgements based on the level of performance they wish to achieve, leading to no net-calibration change, as found in many longitudinal or intervention studies (i.e., Foster et al., 2017).

This explanation can incorporate a number of other findings in metacognitive literature. For example, interventions that correct overconfidence by improving domain knowledge (such as Dunning-Kruger, 1999, Study 4, and Nietfeld et al., 2006) could help improve calibration by aligning desired grades with actual grades, whereas interventions aimed at motivational bias (i.e., Saenz et al., 2017) may be preventing a shift in focus altogether. Another finding indicates that lower performing students, or in other words, more metacognitively inaccurate students, are also less sure about their metacognitive judgements (Miller & Geraci, 2011a; See also studies showing people can “try harder” to make better metacognitive judgements: Buratti & Allwood, 2012, 2013). Why would students purposefully make overconfident metacognitive judgements if they are unsure about them? Would it not be more reasonable to simply make less confident judgements, and be surer about them? One explanation for this apparent discrepancy would be to say that these students are, to some extent, aware that they are being overconfident, but allowing themselves to make overconfident metacognitive judgements, possibly because they are *motivated* to be confident, *but remember their past test performance*. In such a case, a Shifting Focus hypothesis, paired with a motivational explanation seems to fit.

Based on my preliminary data and the findings hereby outlined, I propose that people become less accurate in their self-evaluations (such as test performance predictions) because they discount information by *shifting the focus* of their self-evaluations from useful information (such as academic feedback) towards less factual information such as their desired grades. The preliminary data showed that, yes, people learn about a test shortly after taking it and make more accurate performance prediction, but those data also show that they become less and less accurate over time, resulting in a significantly less accurate performance prediction for the next exam. The proposed motivational explanation for the observed Shifting Focus effects explains this outcome by suggesting that people are more accurate in their self-evaluations immediately after tests because they base those self-evaluations more on factual information, such as how well they feel they performed on the test, any feedback information they received (such as their actual test grade), and on other factual information (such as how well they prepared for the test). Focusing on this information seems adaptive as it may allow one to put an outcome into perspective and accurately consider the results of one's actions. However, as time elapses after a test, and as another test approaches, it also seems adaptive to allow oneself to believe one will perform well on a future test, regardless of prior outcomes or actions. Indeed, there is ample evidence that believing in oneself and being motivated can be key factors in one's success (Cambria & Guthrie, 2010; Feather 1968, 1969; Lench & Bench, 2012; Maslow, 1943), and related theories and findings have been reported in other academic fields (economics: Bénabou & Tirole, 2002; Business: Ucbasaran, Westhead, Wright, 206). Therefore, why wouldn't a person naturally tend towards being confident in their future performance? Unfortunately, this confidence may also lead a person to study less than they need to, thinking that they know more than they actually do. Regaining confidence can be an adaptive motivational habit, but arriving at overconfidence may be a maladaptive self-regulatory mistake. In other words, the ability to regain confidence, that might normally allow people to try again after failure, may also be causing certain students to become overconfident, and under-study for tests, because it leads them to shift the focus of their metacognitive judgements towards desired outcomes instead of factual academic information.

The Current Studies

To investigate the Shifting Focus Hypothesis, observe whether a motivational explanation fits, and to better understand how these ideas fit metacognitive literature, the current experiments were designed to gather data on the following research questions:

1) **Why are test-performance predictions inaccurate and often overconfident? Can the pattern of prediction accuracy fluctuation found in the preliminary data (the Shifting Focus effect) be replicated?**
    a) **Are these results robust to the particular prediction type (absolute vs percentile)?**
    b) **What factors affect these fluctuations?**
    c) **Are the answers to these questions consistent across different laboratory and applied settings?**

d) **Are changes in prediction accuracy more strongly related to the amount of *time* between predictions, or the amount of *thought* put into those predictions?**

2) **Can interventions to improve self-evaluations across multiple tests be developed based on what we know about the causes of inaccuracy in predictions?**

a) **Do such interventions interact with the underlying factors in expected ways? (e.g., does an intervention based on reducing motivational bias in predictions actually do so, in addition to improving prediction accuracy?)**

b) **Can such interventions be implemented with minimal interference in the classroom?**

The remainder of this section describes how the current experiments addressed these research questions, and how these questions relate to the larger body of related literature. **Why are test-performance predictions inaccurate and often overconfident?** It may be that performance predictions become less accurate over time, as suggested by the aforementioned preliminary data. **Can we replicate the pattern of prediction accuracy fluctuation found in the preliminary data?** Experiments 1-3 were designed to replicate and extend my preliminary data in both laboratory and classroom paradigms. **Are these results robust to the particular prediction type (absolute vs percentile)?** Experiments 1-3 gathered self-evaluation data in two ways: absolute and percentile predictions. Whereas *absolute predictions* simply ask people to estimate how well they think they will do on an exam (e.g., "I will answer 80% of these questions correctly.”), *percentile predictions* ask people to estimate their performance in terms of a percentile rank compared to their peers (e.g., “I will perform better than 80% of people on this task.”). Percentile predictions have sometimes been used interchangeably with absolute predictions (Kruger & Dunning, 1999; Dunning, Johnson, Ehrlinger, & Kruger, 2003) and serve as an alternate outcome variable in any case where an absolute prediction is used. Results from percentile predictions are generally similar to those of absolute predictions, but there is evidence that percentile predictions amplify changes or differences in metacognition (Hartwig & Dunlosky, 2014; Tirso, Geraci, & Saenz, 2019).

What factors affect these fluctuations? Beyond self-evaluations, data have also been gathered on the role of motivational (e.g., desired grades) and educational factors (e.g., prior performance) in grade predictions. Saenz and colleagues (2017) demonstrated that overconfident grade predictions are associated with strong motivational bias. These data support the notion that changes in calibration may be associated to changes in the information students use to make their grade predictions. For example, postdictions may be more accurate than predictions because students may base their self-evaluations on the information they just gained from taking an exam. However, as time passes, students may be motivated to shift the focus of their self-evaluations away from academic information like their prior performance or previously observed exam difficulty. Instead, as time passes, people may be influenced by their own motivations to become more optimistic, and therefore make predictions based more on motivationally oriented information, such as their desired grades.

Additional data was gathered to investigate the possible role of anchoring in student performance predictions (anchoring factors) as students may be adjusting their grade predictions around certain pieces of arbitrary information (See Ferrel & McGoey, 1980; Burson, Larrick, & Klayman, 2006; Klayman, Soll, Gonzalez-Vallejo, Barlas, 1999; See also Footnote 2 for a description of how analytical artefacts can indicate overconfidence). For example, could students simply be predicting they will receive "average" grades? If both low and high performers simply estimate their performance at the 80% mark, low performers may seem less calibrated than high performers without there being a real difference in metacognition between the groups. Although such predictions would not be unreasonable, they might reflect a course-grained metacognitive strategy that does not differentiate between lower and higher performing students. In such a case, a ceiling effect would account for high performers more accurate metacognition.

Are the answers to these questions consistent across different laboratory and applied settings? Experiment 2 tested the ecological validity of the preliminary data by extending the reflection period to a week of time. In university courses, weeks of time may pass between two tests. Therefore, if the mechanism found in the 10-minute laboratory paradigm is the same as the one experienced by students in actual classrooms, the preliminary data should be replicable over a period greater than 10 minutes. Experiment 3.1 asked students in an actual university course to make grade predictions before taking two university exams, as well as to make predictions about their performance once per day, during each of the class periods (about 5) between the two exams. Experiment 3.2 followed the same procedures, but over the course of a month of time, and improved data collection and participant compliance by including an incentive to complete study procedures. These data were expected to replicate the Shifting Focus effect and motivational/educational interactions of Experiment 1.

Are changes in prediction accuracy more strongly related to the amount of time between predictions, or the amount of thought put into those predictions? Experiments 2-3 may also illuminate whether time and/or number of predictions are more associated to changes in metacognitive confidence. Because Experiments 1-3 involved different reflection period lengths, and quantities of reflection period predictions, variations in the size of the Shifting Focus effect may indicate a greater impact of time or quantity of predictions. For example, a larger confidence change in Experiment 1 (10 minutes, 10 predictions) compared to Experiment 3.2 (1 month, 4 predictions) would suggest that the quantity of predictions is more important than the amount of time elapsed since a test for increased confidence.

Can interventions to improve self-evaluations across multiple tests be developed based on what we know about the causes of inaccuracy in predictions? Experiments 4 and 5 tested the efficacy of two interventions aimed at preventing and eliminating rising overconfidence in the classroom. Both of these interventions were designed to be as noninvasive and as low-cost as possible, so as to be widely applicable in actual classes. Experiment 4 again asked students to make grade predictions before and after taking two in-class exams. During the weeks between these class exams, students were asked to make online grade predictions at home at a rate approximating the number of class periods (3 times per week, for four weeks). In addition to making simple predictions, students were asked to respond to indicate their initial test grades and predictions, this was intended to act as a form of self-feedback. As students recall their originally-overconfident judgements, they might hesitate to increase their predictions over time, leading to a permanence in the metacognitive learning associated with test taking. Put simply, Experiment 4 tried to prevent a Shifting Focus effect by reminding students of their past overconfidence while they made new grade predictions.

Do such interventions interact with the underlying factors in expected ways? In addition to gathering self-evaluation data during these intervention studies, anchoring, educational, and motivational factor data were also gathered to observe how the basis of performance predictions was affected by an intervention. Experiments 1-3 were expected to show motivational factors becoming more associated to predictions relative to educational factors as the reflection period progressed. Therefore, the relationship between motivational factors, educational factors, and performance predictions was not expected to change during the reflection period of Experiment 4 (and Experiment 5), because the intervention was designed to prevent that very shift.

Can such interventions be implemented with minimal interference in the classroom? Experiment 5 took this intervention a step further and attempted to prevent rising confidence in students in the least invasive way possible. Experiment 5 again asked students to make grade predictions during two in-class exams. However, instead of performing repeated predictions between the two exam dates, students' first grade predictions were posted alongside their first test grades, so as to remind participants of their prior overconfidence every time they looked at their grades online. As a further benefit of this intervention, data was gathered on the amount of times individual students viewed their grades, allowing for correlation between participation and metacognitive change. Students who viewed their grades more often, may be better inoculated against a Shifting Focus effect.

CHAPTER II

EXPERIMENTS 1-3: REPLICATING AND EXTENDING

Experiment 1

This experiment tested the hypothesis that grade predictions can increase in confidence significantly over just 10 minutes of time using a within-subjects paradigm. Participants took two tests, and made predictions and postdictions about each. Participants also completed 10 grade predictions during a reflection period between the two tests so they may reflect on their prediction accuracy. Data were also gathered on the impact of certain factors can explain this change in predictions (e.g., anchoring, educational, and motivational factors). As part of this exploratory replication, demographic information was also gathered to investigate individual differences in the Shifting Focus effect.

Methods.

Participants. One-hundred sixty four (164) undergraduate students were recruited through an introductory psychology research pool. Preliminary data showed a significant, though small effect size increase in prediction confidence with about 40 participants (Figure 2). So to investigate performance quartiles (along with a number of other moderating: age, gender, race, performance level, and first-generation student classification), approximately 40 people per performance quartile cell were recruited.

Participants were mainly younger-college aged (*M* = 18.62, *SD* = .99) with 86.6% of participants being either 18 or 19 years of age. Participants were mostly female (N = 115 or 70.1%), and mostly self-identified as either Caucasian (N = 98, 59.8%) or Hispanic (N = 40, 24.4%), with a number of smaller groups (N = 6 African American, N = 9 Asian Indian, and N = 11 Asian) that will hereafter be batched and referred to as an "other group" due to sample size restrictions. Of the 164 participants, only 28 (17.1%) identified as first generation students.

Materials. All participants completed two logical reasoning tests taken from the Official LSAT Preptest of June 2007, and from the LSAT logical/analytical reasoning self-assessment modules from testprepreview.com (See Appendix A for sample questions). These tests were designed to be very difficult while avoiding floor effects, and were the same as those used in the pilot study. Average performance in the pilot study was low (*M* = 37.11, *SD* = 13.46) by design to give participants ample room to improve their calibration between exams and predictions. Each test consisted of 20-multiple choice questions, to be answered in 20 minutes. Test order was counterbalanced across participants.

Participants responded to a self-evaluation survey before and after each of two exams (Appendix B). Each survey began with a self-evaluation question that formed the main dependent variables for the current study. Hereafter, these data points are referred to as Prediction 1/2, and Postdiction 1/2 for the self-evaluations made before and after each of the two exams (test 1/2). The wording of these judgements was based on a number of previously published works on this topic (Foster et al., 2017; Hacker et al., 2008; Saenz et al., 2017): "What grade do you think you will receive on this test?" Each of these judgements was to be the first item on a survey of questions investigating metacognition at each prediction time (see full survey under Appendix B). Additionally, the fourth question in the survey asked participants to estimate their performance in terms of percentile ranking compared to their peers (a percentile prediction and alternate dependent variable). The same survey was used in Experiments 1 and 2 (Experiments 3-5 used a very similar questionnaire described in Experiment 3).

Each of the explanatory factors (anchoring, educational, and motivational) were investigated using multiple items (See Appendix B). Anchoring was investigated with questions 2 and 3: “What is your goal to earn on this test?”, and, “What do you think will be the average grade on this test?” Questions 5-11 gauged the extent to which students grade predictions were based on academic (5-8: educational) or desire (9-11: motivational) –based information. These items were adapted from a recent study (Saenz et al., 2017, Study 4) that investigated the role of students’ grade desires in their prediction information.

Procedure. After providing consent, all participants completed a short demographic form before taking two logical reasoning exams and completing self-evaluation surveys immediately before and after each exam. Of greatest importance, each of these four surveys asked participants to estimate the grade they thought they would get on a test they were about to take (Prediction 1/2: “What grade do you think you will receive on this test? ___%”) or that they had just taken (Postdiction 1/2: “What grade do you think you will receive on this test? ___%”) on a 0% - 100% scale. Predictions were completed on a survey asking a number of metacognitive and academic questions included relevant pre/postdictions, as well as questions about the factors affecting students' predictions (e.g., grade desires, prior test performance; following the procedure used by Saenz et al., 2017, Study 4; Appendix B). Between the two prediction-test-postdiction cycles, participants were asked to, "Ruminate on making the most accurate grade prediction for an exam similar to the one you have just taken". During this Reflection Period, participants were asked to repeatedly make grade predictions, once per minute, for a total of 10 predictions during this 10-minute Reflection Period (e.g., Figure 3). Participants recorded these responses on paper, and were prompted to make each of these predictions aurally by an experiment proctor. Participants were briefly informed about the content tests before their initial predictions, and they were expressly told that second exam was very similar to the first.

Results and Discussion.

Manipulation Check. Before investigating the main questions of the current studies, a manipulation check was performed to observe whether overconfidence was found across the present study. In this, and all further studies, Calibration was calculated as self-evaluation minus performance (self-evaluations include predictions, postdictions, and predictions made during Reflection Periods). Absolute Calibration is the absolute value of Calibration. Indeed, on average all four calibration scores (Prediction 1/2 and Postdiction 1/2) were overconfident (p < .001, one-sample t-test compared to zero), and all absolute calibration scores indicated significant prediction error (p < .001). Furthermore, low performers (first performance quartiles) were significantly less accurate and more overconfident than high performers in terms of both calibration and absolute calibration (p < .001) across Experiment 1. Note, however, that almost no *underconfidence* was observed because, following the experimental design, the average test performance was very low (*M* = 36.657, *SD* = 17.897), such that all participants had some opportunity to change the accuracy of their predictions during the study. Descriptive statistics for Experiment 1 have been summarized in Table 1.

Replicating Preliminary Results: Prediction Accuracy Over Time. Figure 4 outlines the results of Experiment 1, which replicated preliminary data in its three key features including. Self-evaluation confidence decreased from Prediction 1 to Postdiction 1 (Feature 1: $t(162) = 15.681$, $p < .001$, $d_z = 1.228$; $M = 75.13$, $SD = 12.266$; to $M = 53.55$, $SD = 18.817$). Participant's confidence peaked at Reflection Period Prediction 9, which marked a significant increase from Postdiction 1 (Feature 2: $t(163) = -7.364$, $p < .001$, $d_z = -.575$; $M = 53.34$, $SD = 18.941$; to $M = 63.12$, $SD = 19.779$). Finally, Self-evaluative confidence decreased significantly from Reflection Period Prediction 9 to Postdiction 2 (Feature 3: $t(161) = 7.146$, $p < .001$, $d_z = .561$; $M = 63.41$, $SD = 19.606$; to $M = 54.22$, $SD = 18.169$).

Although the second Feature, an increase in confidence between tests, is of most interest, these three Features are the hallmark of the Shifting Focus effect thus far described, and may be affecting the way students perform across their academic careers. The initial decrease may indicate self-evaluative learning associated with taking an exam, wherein students better understand their level of knowledge for that domain, and become more metacognitively accurate (as in the postdiction superiority effect: Pierce & Smith, 2001). The second Feature, an increase in confidence between tests may mark a mental shift wherein people become more confident, and potentially less metacognitively accurate.

The third Feature, a final decrease in confidence, may mimic the decrease of Feature 1 as taking a test generates more self-feedback, and might cause participants to refocus their self-evaluations back to being factually oriented, instead of motivationally driven. Why does Feature 3 begin to show a reduction in self-evaluative confidence *before* Test 2, when no new feedback could have been generated? There is evidence that students are aware of their overconfidence (Miller & Geraci, 2011a), and they may demonstrate that by reducing their predictions very shortly before an exam. Other researchers have found that, when faced with impending feedback (i.e., something that could prove one wrong, such as a medical test), people make more pessimistic metacognitive judgements (Shepperd, Ouellette, & Fernandez, 1996; Shepperd, Findley-Klein, Kwavnick, Walker, & Perez, 2000; Taylor & Shepperd, 1998), possibly because the anxiety of an upcoming test may again shift their predictive focus.

In addition to the three Features, a number of corroborating analyses were run. Paired samples comparisons showed that participants became more confident from Postdiction 1 to Prediction 2 ($t(163) = -6.402$, $p < .001$, $d_z = .500$; $M = 53.34$, $SD = 18.941$; to $M = 60.19$, $SD = 17.920$). Repeated measures analyses corroborated the interpretation of these confidence differences as incremental shifts as both the rise in confidence over the Reflection Period (Feature 2: $F(1,163) = 44.156$, $p < .001$, *partial* $\eta^2 = .213$) and the fall in confidence from Reflection Period Prediction 9 to Postdiction 2 (Feature 3: $F(1,161) = 56.236$, $p < .001$, *partial* $\eta^2 = .259$) were significant, linear outcomes. No repeated measure analysis was necessary for Feature 1 because it only involved 2 time points, Prediction 1 to Postdiction 1. The linear nature of these outcomes suggests that people's self-evaluations change incrementally. If, for example, confidence fell the moment a test was placed in front of a participant, and rose to plateau when it was taken away, one might suspect the presence of the test to be associated with the change in participant confidence. Instead, however, the linear nature of these changes suggests participants become more confident over time, or at least over multiple prediction attempts within a 10 minute span.

To further corroborate these results, a series of paired samples comparisons were run on the supplementary prediction measures: percentile predictions included in Appendix B (item 3), that asked participants to estimate their percentile ranks instead of simple grade predictions. Running these analyses with percentile predictions replicated the first and third Features, significant drops in predictions from before to after each test (Prediction 1 to Postdiction 1: $t(161) = 8.970$, $p < .001$, $d_z = .705$; $M = 58.78$, $SD = 13.309$; to $M = 47.11$, $SD = 17.033$; Prediction 2 to Postdiction 2: ($t(146) = 4.332$, $p < .001$, $d = .357$; $M = 48.61$, $SD = 16.114$; to $M = 44.53$, $SD = 17.413$). Because percentile predictions are thought to produce more extreme results than absolute performance predictions, it may not be surprising that the effect size of the difference from Prediction 1 to Postdiction 1 was greater for a percentile prediction than for an absolute performance prediction (Feature 1), but this relationship was not true for the comparison between Prediction 2 and Postdiction 2 (Feature 3), and there was no significant increase in percentile predictions from Postdiction 1 to Prediction 2 (feature 2: $t(156) = -.948$, $p = .345$, $d_z = -.076$; $M = 47.05$, $SD = 17.111$; to $M = 47.94$, $SD = 16.393$). Note, however, that these measures could not be analyzed in the same way as were absolute performance predictions for Features 2 and 3, because Reflection Period data was not collected for these variables. Further investigation is necessary to verify whether the Features of the preliminary data are entirely consistent across different types of metacognitive questions, but participants did, at least, become more metacognitively accurate from predictions to postdictions. It may be that, because the Reflection Period specifically involved absolute judgements, participants did not think about their performance in relation to other students, and so did not change their relative judgements. To anticipate further studies, percentile prediction results were similar across all studies, generally mirroring those of absolute predictions with reduced effect sizes and as such will not be reported in subsequent studies.

Performance Levels and Prediction Accuracy Over Time. Do both low and high performers exhibit the same change in predictions over time? Because low performers are often considered more inaccurate and overconfident (e.g., Dunning & Kruger 1999), one might suspect that low performers would be more susceptible to increasing prediction confidence during the reflection period. Participants were divided into four roughly equal groups based on average test performance (performance quartiles). Additionally, there is evidence that high performing students are more capable of incorporating feedback to improve the accuracy of their judgements (Drunning & Kruger, 1999, Study 3), so the opposite might also be true: high performers might be less susceptible to regaining overconfidence. Table 2 details

Prediction, Calibration, and Absolute Calibration for low and high performer data. Low performers improved their self-evaluation accuracy to a similar or greater extent (higher effect size) than average for both tests (Feature 1, Prediction 1 to Postdiction 1: $t(45) = 8.738$, $p < .001$, $d_z = 1.288$; $M = 76.52$, $SD = 13.617$; to $M = 50.70$, $SD = 22.252$; Feature 3, Reflection Period Prediction 9 to Postdiction 2: $t(46) = 3.719$, $p = .001$, $d_z = .543$; $M = 61.64$, $SD = 23.869$; to $M = 51.51$, $SD = 19.705$), but also became more drastically inaccurate during the Reflection Period (Feature 2, Postdiction 1 to Reflection Period Prediction 9: $t(46) = -4.242$, $p < .001$, $d_z = -.619$; $M = 50.04$, $SD = 22.460$; to $M = 61.64$, $SD = 23.869$).

High performers still exhibited the three Features from the preliminary data, but to a slightly lesser (in terms of effect size), though still significant degree (Feature 1: $t(46) = 7.876$, $p < .001$, $d_z = 1.149$; $M = 76.53$, $SD = 11.017$; to $M = 59.06$, $SD = 15.495$; Feature 2: $t(46) = -2.797$, $p = .007$, $d_z = -.408$; $M = 65.49$, $SD = 18.891$; to $M = 60.32$, $SD = 16.907$; Feature 3: $t(46) = 2.853$, $p = .006$, $d_z = .416$, $M =65.49$, $SD = 18.891$; to $M = 60.32$, $SD = 16.907$). These results suggest that both low and high performers are susceptible to increasing confidence between tests. Although high performer’s data seemed to be a less extreme version of low performer’s results, they were also in a less extreme case. Both group’s Prediction 1 was an overconfident 76, but high performers were inherently closer to being accurate, and so they would have needed to compensate less at Postdiction 1, and then would have less room to regress during the Reflection Period.

Although high performers are often underconfident, increased confidence between tests may be entirely adaptive, but because low performers are often overconfident, the same behavior may not be adaptive. Regardless, because both groups rose and fell in self-evaluative confidence in similar way, it may be that these changes in confidence are common to everyone, not just the metacognitively inaccurate or overconfident. Although further investigation is needed to confirm these results, to anticipate the results from further experiments in the current writing, both low and high performers seemed to be equally affected by the three features of the preliminary data, except in cases where participant numbers were too low for proper analysis. Further analyses of the relationship between the Shifting Focus effect and performance levels have therefore been omitted for brevity and due to poor and inconsistent sample sizes.

Factors Affecting Performance Predictions. Items 2-3 (Anchoring factors), 5-8 (Educational), and 9-11 (Motivational Factors, Appendix B) were entered as grouped Factors into a series of hierarchical regressions predicting performance predictions, each regression using a prediction, postdiction, or Reflection Period prediction as its dependent variable. These regressions produced a series of coefficients that showed the explanatory value of each of the Factor groups for how participants made their self-evaluations across Experiment 1 (as was done by Saenz et al., 2017).

These groups of variables were entered into a hierarchical regression, with items 5-8 entered as a group (Educational Factors), followed by 2 & 3 entered individually (Anchoring Factors), and 9-11 entered as a group (Motivational Factors). This order was chosen because it may give the most conservative estimate of the

Motivational Factors, and the most liberal estimate of Educational Factors, allotting as much explanatory power to the Educational Factors first. This type of analysis was performed for each grade prediction and postdiction, including those during the Reflection Period.

Educational Factors were expected to have very low explanatory power that would rise and fall according to calibration scores, while Motivational Factors would have an inverse relationship with Calibration and be consistently high, following results from prior experiments (Saenz et al., 2017; Serra & DeMarree, 2016). For example, postdictions are known to be more accurate than predictions (Pierce & Smith, 2001), so one might predict that postdictions would be more strongly associated with Educational Factors, and less strongly associated with Motivational Factors, than predictions. Finally, Anchoring Factors were expected to have significant explanatory power that would stay constant throughout the study, because they were on the same scale as the initial prediction, and because there is some evidence that predictions may be anchored naturally (Ferrel & McGoey, 1980). Grade predictions made during the Reflection Period did not have full self-evaluation surveys associated to them, so instead their regressions were run using data from Postdiction 1. Figure 5 summarizes the expected findings for the relationships between explanatory factors and calibration across Experiment 1.

Figure 6 summarizes the actual findings for explanatory factor and calibration relationships. Anchoring Factors were noticeably more important for predicting grade predictions than either Educational or Motivational Factors across Experiment 1. The relationships between Motivational Factors, Educational Factors, and calibration were not as clear they were expected to be. Whereas both Motivational and Educational Factors followed expected trends at some points throughout the study, they accounted for very little variance in self-evaluations, and their effect sizes seemed mostly flat and low compared to Anchoring Factors'.

To anticipate further studies, analyses of the explanatory factors was not very illuminating as results seemed inconsistent and, at best, only partially consistent with expected outcomes. To simplify this and further explanatory factor analyses, a new set of comparisons was devised. Regressions predicting key self-evaluations (Prediction 1, Postdiction 1, Peak Reflection Period Prediction, and Postdiction 2) were each run twice, once using Educational Factors and once using Motivational Factors as independent variables (grouping, and stepwise conditions were eliminated). The effect size (R^2) of Motivational Factors was then subtracted from that of Educational Factors, to produce a bias quotient. A more positive number indicated that self-evaluations were based more on Motivational Factors, whereas a more negative number indicated that self-evaluations were more based on Educational Factors, with zero indicating a perfect balance between the two types. This quotient would then be plotted alongside the key self-evaluations to check for expected results. Although this representation did not fully align with the expected results or a-priori proposed analyses (i.e., overall correlations between predictions and factors), it presented a simpler picture of the present data, and may suggest a trend in the way in which Motivational and Educational Factor's relationship changes across multiple tests.

Factor analysis data for Experiments 1-3 was summarized in Figure 7. Although Figure 6 did not depict the expected inverse relationship between Motivational Factors and performance predictions, Figure 7 did seem to show Motivational factors becoming less important relative to Educational Factors across Experiment 1. However, while there may be a trend in the expected direction, these results were hazy at best as Motivational Factors did not become more important with rising confidence during the Reflection Period, and changes in the Relative value of Motivational and Educational Factors amounted to a less than 10% difference. Further data is needed to better understand if and how Motivational, Educational, and Anchoring Factors are related to prediction change over time.

Individual Differences in Shifting Focus. An exploratory set of analyses was run to check for individual differences in calibration, as well as calibration change over time including comparisons for: age, gender, race, and first-generation student status. Calibration and Absolute Calibration were averages across all prediction, postdiction, and Reflection Period self-evaluation points, to create an Average Calibration and an Average Absolute Calibration variable. ANOVAs were run using the each of the previously described demographic variables, but there were no significant differences between groups on either Average Calibration ($p \geq .069$) or Average Absolute Calibration ($p \geq .214$). Differences in calibration based on age were closest to being significantly different, but this is most likely due to having very small sample sizes for participants aged 20 and up. People did not seem to be better or worse at self-evaluation based on age, gender, race, or first-generation status. Table 3 summarizes these data across Experiment 1, but in brief, none of the demographic groups seemed to show sizeable differences in their calibration lines across Experiment 1.

Experiment 2

Experiment 2 was designed to replicate and extend the results of Experiment 1 by extending the Reflection Period to a week of time. There are two reasons for extending the Reflection Period. First, this manipulation may help discern whether the passing of time, or the amount of thought on predictions is responsible for the increase in self-evaluation confidence demonstrated in Experiment 1. Experiment 2 gave participants much more time between predictions, and fewer Reflection Period predictions than Experiment 1. If predictions increase in Experiment 2 more than in Experiment 1, then the amount of time between tests may be more important than the quantity of predictions in producing a Shifting Focus effect. If predictions in Experiment 2 do not increase as much as they did in Experiment 1, then the amount of rumination (or quantity of predictions) may be more responsible for change in predictions. Secondly, students in an actual class environment do not experience merely 10 minutes of reflection time between class exams. By extending the amount of time in the Reflection Period, Experiment 2 may provide data for change in grade predictions under a more ecologically valid setting.

Methods.

Participants. Because participants self-initiated grade predictions during the Reflection Period of Experiment 2, some non-compliance and missing data was expected, so 117 undergraduate students were recruited for Experiment 2. Of these, only 88 participants successfully completed the second phase of Experiment 2. The smallest effect size of the main analyses of Experiment 1 was d_z = .561, an effect size with a minimum sample size of 36 (based on power analyses). So the sample size of Experiment 2 was approximately twice as large as it needed to be based on effect sizes from Experiment 1, in the hopes that this would be sufficient to eliminate non-compliance issues. Note that, although not all participants completed all portions of the study, as much data as possible was used so df values may be drastically different in some analyses of this, and all following studies.

Demographic information for the 117 participants in Experiment 2 was similar to that of Experiment 1. Participants were mainly early-college aged (*M* = 18.91, *SD* = 1.70) with 80.3% of participants being either 18 or 19 years of age. Participants were mostly female (N = 80 or 68.4%), and mostly self-identified as either Caucasian (N = 70, 57.8%) or Hispanic (N = 29, 24.8%), with a smaller "other" group of varied race or ethnicity (N = 18). Of the 117 participants, only 18 (15.4%) identified as first generation students.

Materials and procedure. The prediction surveys, logical reasoning tests, and most procedures were the same as those used as in Experiment 1. The Reflection Period for this experiment was extended to one week, so that the experiment was completed in two sessions, each including one test and related prediction surveys. Participants made up to 7 Reflection Period grade predictions during the extended Reflection Period, and were instructed to do so once per day, using an online google form. To ensure consistency in participants' responses, the data were combed for any predictions made inappropriately. For example, participants who made multiple predictions during a single day only have their first prediction of the day included in the analyses. However, all required predictions made during either testing session will be included. No incentives were given for completing the Reflection Period predictions to avoid introducing any novel motivational influence.

Results and Discussion.

Manipulation Check. A manipulation check was again performed to ensure overconfidence levels were similar to those of Experiment 1. Indeed, on average participants were overconfident at Predictions 1 and 2 as well as Postdictions 1 and 2 ($p < .001$). Low performers (bottom 25% of participants by average test performance) were always significantly more confident and less accurate than high performers in both Calibration and Absolute Calibration ($p < .001$).

Replicating Preliminary Results: Prediction Accuracy Over More Time. Because participant numbers change drastically throughout the experiment, a summary of self-evaluations, Calibration, and Absolute Calibration across Experiment 2 has been included in Table 3. Feature 1 replicated perfectly, participants significantly reduced their self-evaluation confidence, and became more metacognitively accurate from Prediction 1 to Postdiction 1 ($t(115) = 14.513$, $p < .001$, $d_z = 1.348$; $M = 74.80$, $SD = 11.693$; to $M = 51.23$, $SD = 18.276$). Self-evaluation confidence peaked on the second day of the Reflection Period, unlike Experiment 1 where confidence rose throughout the majority of the Reflection Period. To double-check the data, analyses for all three Features were run again with Day 2 and Day 7 (the final Reflection Period

Judgement) as the pivoting points for self-evaluations in Experiment 2. Feature 2 was replicated successfully (Postdiction 1 to Day 2: $t(47) = -6.064$, $p < .001$, $d_z = -.875$; $M = 49.96$, $SD = 17.813$; to $M = 59.40$, $SD = 15.789$; Postdiction 1 to Day 7: $t(56) = -2.865$, $p = .006$, $d_z = -.383$; $M = 52.20$, $SD = 16.387$; to $M = 56.89$, $SD = 19.906$), although participants did not have as much self-evaluative confidence on Day 7 of the Reflection Period, so its related effect size was less than the change from Postdiction 1 to Day 2. Results for Feature 3 follow the previous, with a significant drop in self-evaluative confidence from day 2 to Postdiction 2 ($t(56) = -2.865$, $p = .006$, $d_z = .383$; $M = 52.20$, $SD = 16.387$; to $M = 56.89$, $SD = 19.906$), but no significant change from Day 7 to Postdiction 2 ($t(56) = -2.865$, $p = .006$, $d_z = .383$; $M = 52.20$, $SD = 16.387$; to $M = 56.89$, $SD = 19.906$). To corroborate, participants became more confident from Postdiction 1 to Prediction 2 ($t(88) = 3.440$, $p = .001$, $d_z = .365$; $M = 51.83$, $SD = 17.803$; to $M = 55.52$, $SD = 18.000$).

It seems as though the change in self-evaluative confidence is not necessarily higher as the next test approaches. Though confidence did still increase during the Reflection Period, it peaked at Day 2, and seemed to drop slowly until Postdiction 2, with no significant changes happening between any of the intervening self-evaluations. Although results for Experiment 2 Features 2-3 were partially corroborated by linear repeated measures analyses (Feature 2, Postdiction 1 to Day 2: $F(2,33) = 27.532$, $p < .001$, *partial* $\eta^2 = .455$; Feature 3, Day 2 to Postdiction 2: $F(1,8) = 1.165$, $p = .768$, *partial* $\eta^2 = .068$), the data for a repeated measures analysis had an abysmally low quantity of participants (N = 9) due to inconsistent participant compliance throughout

Experiment 2. Nevertheless, Experiment 2 seemed to replicate the results of Experiment 1, and the preliminary data.

Factors Affecting Performance Predictions. Anchoring, Educational, and Motivational Factors were again analyzed in the same fashion as Experiment 1. In brief, these factors did not follow expected trends (Figure 7). Motivational Factors became more important relative to Educational Factors from Postdiction 1 to Postdiction 2, instead of exhibited the expected inverse correlation with predictions.

Individual Differences in Shifting Focus. An exploratory series of analyses were performed to investigate whether individual differences had any effect on metacognitive accuracy. Results replicated those of Experiment 1, however, and no significant differences in either Average Calibration (p = .063) or Average Absolute Calibration (p = .271) were found for any of the demographic variables. The closest analysis to being significant was that comparing Average Calibration between age groups (p = .063), but this apparently marginal result was due to extremely small Ns for participants aged 20+, all other analyses were well above ($p \geq .200$) significance levels.

Experiment 3.1

Experiment 2 replicated the results from Experiment 1 and the preliminary data showing that people become less metacognitively accurate during the time between tests, But Experiment 2 further found this result to be true for a week-long lab study, instead of the 10-minute setting involved in the previous findings. Experiment 3 was designed to replicate the results from Experiments 1-2 in a classroom environment.

Increases in predictions over time in a classroom environment would establish further evidence for the Shifting Focus effect here described, as well as provide its validity in an applied environment. Beyond differences in prediction accuracy, data were gathered to assess a motivational explanation for the Shifting Focus effect. Although the results from Experiments 1 and 2 were rocky at best, the primary area of focus, Motivational and Educational Factors, may come into greater play in an ecologically valid setting, because real students would actually be motivated to perform well, but have educational information at their disposal when making performance predictions. Thus the applied nature of Experiment 3 might have produced clearer associations between explanatory factors and the Shifting Focus effect.

Experiment 3 was divided into two parts because Experiment 3 was performed twice. The first time Experiment 3 was performed (hereafter referred to as Study 3.1) the number of participants was very low, and so was compliance, resulting in very difficult data to analyze. Although these data are here presented, Experiment 3.2 gathered more participants, and improved on participant compliance, resulting in much more complete data. Both data sets are reported in full in the interest of transparency.

Methods.

Participants**.** Participants in this experiment consisted of 36 university students taken from a Psychology of Learning Psychology and Neuroscience course in the summer of 2018. This sample size was limited first by the quantity of students enrolled in the course, and second by the quantity that consented to participate. Demographic information was not collected because the small sample size may have resulted in privacy concerns. However, participants were of similar proportion to Experiments 1-2 in every way.

Materials and Procedure. The procedure for Experiment 3 was similar to that of Experiments 1-2, with exception of necessary adaptations for an actual course. Therefore, instead of taking two LSAT exams, participants completed two in-class exams, spaced a week apart, following the summer-semester schedule of the present university course. Participants completed a self-evaluation survey before each of these two exams using the same form included in the previous studies. Unlike Experiments 1-2, no postdiction self-evaluation survey was completed due to time constraints and instructor preference. This limitation unfortunately reduces the number of analyses that can be run on these data. There were 4 class periods between these two exams, and participants made a grade prediction at the beginning of each of these class days. The two exams consisted of multiple choice questions, and were the third and fourth exams in the course. In addition to the above changes, the self-evaluation survey used in this, and all further experiments was altered slightly to better fit a classroom setting. Item 7 (Appendix B), which asked participants about their prior performance on, “this type of exam” was removed.

Results and Discussion.

Procedure Check. A series of analyses were performed to check for overconfidence levels. Because the sample size was too small to analyze the data in quartiles, results were analyzed as a whole group. On average participants were <u>not</u> overconfident across the experiment (*t*(35) = -.339, *p* = .737, d_z = -.056, *M* = -.500, *SD*

= 8.845). This result may not be surprising as test scores (*M* = 83.14, *SD* = 7.948) may have been slightly higher than other courses of the same level, and because the test scores of the prior two exams were approximately 20 points lower. Nevertheless, there was still a significant amount of inaccuracy in Absolute Calibration scores (*t*(35) = 9.994, *p* < .001, d_z = 1.666, *M* = 8.15, *SD* = 4.890). It is difficult to say how these circumstances may affect results.

Prediction Accuracy over Time. Because no postdictions were gathered in Experiment 3.1, it was not possible to observe how self-evaluative accuracy changed from before to after each exam (Features 1 and 3). In addition, due to poor participant compliance, paired samples comparisons were underpowered, and it was difficult to determine the best way to analyze these data. To give the most positive result, Feature 2 seemed to be replicated in the same was as it was in Experiment 2. Participants seemed to become more confident in their performance predictions from Prediction 1 (*M* = 80.909, *SD* = 10.578) to Reflection Period Prediction 2 (*M* = 85.364, *SD* = 5.653; *t*(21) = -2.138, *p* = .044, d_z = .456), though this confidence did not necessarily lean towards overconfidence. Because data from this experiment was so poor, further analyses were not attempted.

Factors Affecting Performance Predictions. Motivational and Educational Factors were analyzed as in Experiments 1-2. Regression results were precisely opposite of expected results, Motivational Factors were less important to self-evaluations (relative to Educational Factors) at Prediction 1 and the Reflection Period Peak, times when self-evaluations should have been most important.

Experiment 3.2

Experiment 3.1 had a very small sample size, numerous participant compliance issues, no postdictions, and an inter-test interval of only one week (a circumstance that does not mimic most academic settings).Therefore Experiment 3.2 involved a second class worth of data that did not have any of these limitations.

Methods.

Participants. A group of 132 undergraduate students participated in partial completion of their introductory psychology course requirements. As in Experiment 3.1, this was a convenience sample, with sample size determined by the amount of willing participants in the course. Experiments 3.1, 3.2, 4, and 5 each were performed in class, but although some were the same course, each had a different instructing professor. Participants were mainly early-college aged (M = 18.293, SD = .624) with 95.5% of participants being either 18 or 19 years of age. Participants were mostly female (N = 88 or 66.7%), and mostly self-identified as either Caucasian (N = 73, 55.3%) or Hispanic (N = 32, 24.2%), with an assorted “other” group of varied race or ethnicity (N = 27). Of the 132 participants, only 26 (19.5%) identified as first generation students.

Materials and Procedure. The materials and procedure for Experiment 3.2 were very near to those of Experiment 2, and mainly differed from Experiment 3.1 in that Study 3.2 did not suffer from the same limitations as its predecessor. At the approval of the instructor, the same demographic survey from Experiments 1-2 was used in Study 3.2. Although Study 3.2 still used the slightly altered self-evaluation survey for predictions, postdiction data were also gathered in Study 3.2. Study 3.2 was also performed during a regular university course semester, and so the Reflection Period lasted approximately one month, the time between two in-class exams. However, to avoid over-burdening class-time, Reflection Period Predictions were only made once per week, resulting in only 4 Reflection Period Predictions, in addition to Predictions 1/2, and Postdictions 1/2. Finally, compliance issues were reduced in Experiment 3.2 by making bonus research credits available for any participant who completed the experiment in full. Because of this, compliance was noticeably improved, with 61 of 132 (46.2%) participants completing experiment procedures in full. This reward procedure was exclusive to Experiment 3.2, and was *not* used in any other experiments of this dissertation. Despite the incentive, not all participants completed all experiment procedures, so df values may vary from analysis to analysis. Performance quartiles *were* calculated for the current experiment and analyses.

Results and Discussion.

Procedure Check. The data were analyzed to check for overconfidence check was again performed to ensure overconfidence levels were similar to those of experiments 1 and 2. In terms of average Calibration participants were overconfident at Prediction 1 (p = .002) and 2 (p = .001), but they became significantly underconfident at Postdiction 1 (p = .006), and were not significantly inaccurate at Postdiction 2 (p = .409). Absolute Calibration scores were always significantly inaccurate (p < .001). Low performers (bottom 25% of participants by average test performance) were always significantly more confident and less accurate than high performers in both Calibration and Absolute Calibration ($p < .001$). In sum, these data represented a normal set of classroom self-evaluations.

Prediction Accuracy over Time. Self-evaluative confidence was highest at the first Reflection Period Prediction, which occurred approximately one week after the first exam, this may mirror Experiment 2, where confidence was highest at the second day. All three features of the prior experiments were replicated; participants become less confidence from Prediction 1 to Postdiction 1 (Feature 1: ($t(102) = 9.332$, $p < .001$, $d_z = .919$; $M = 78.86$, $SD = 11.758$; to $M = 71.33$, $SD = 14.964$), more confident from Postdiction 1 to Reflection Period 1 (Feature 2: ($t(91)= -11.041$, $p < .001$, $d_z = -1.151$; $M = 72.48$, $SD = 12.759$; to $M = 83.43$, $SD = 7.821$), and less confident from Reflection Period 1 to Postdiction 2 (Feature 3: Reflection Period 1 to Postdiction 2 ($t(90) = 7.203$, $p < .001$, $d_z = .755$; $M = 83.25$, $SD = 7.646$; to $M = 75.48$, $SD = 13.140$). Note that although the point of greatest confidence occurred at the first Reflection Period Prediction, participants also became less confident from Prediction 2 to Postdiction 2 ($t(97) = 2.479$, $p = .015$, $d_z = .250$; $M = 77.94$, $SD = 11.562$; to $M = 76.28$, $SD = 10.787$). Furthermore, a repeated measures analysis showed a significant downwards linear trend in self-evaluative confidence from Reflection Period 1 to Postdiction 2 ($F(1, 67)$, $p < .001$, *partial* $\eta^2 = .499$).

These results suggest that the participating students became more confident over the time between tests. To corroborate this point, a paired samples analysis showed a significant increase from Postdiction 1 to Prediction 2 ($t(95) = -6.437$, $p < .001$, $d_z = .657$; $M = 72.05$, $SD = 14.128$; to $M = 78.41$, $SD = 11.506$), suggesting increased confidence from immediately after the first exam to immediately before the second. Going further, there was no significant difference between Prediction 1 and Prediction 2 (p = .496), nor was Prediction 2 (M = 78.28, SD = 11.009) any more accurate than Prediction 1 (Self-evaluation M = 78.94, SD = 10.669; Calibration paired samples comparison p = .952, Absolute Calibration paired samples comparison p = .777) despite significant accuracy improvements from before to after the first test (Calibration: t(99) = 9.137, p < .001, d_z = .914; M = 41.64, SD = 9.428; to M = 34.160, SD = 12.452; Absolute Calibration: t(99) = 10.952, p < .001, d_z = 1.095; M = 41.640, SD = 9.428; to M = 34.820, SD = 10.443). So, despite having improved their metacognitive accuracy by taking test 1, students' self-evaluative predictions had regressed to their originally inaccurate state by Prediction 2.

Factors Affecting Performance Predictions. Educational, and Motivational Factors were analyzed as in previous studies. Figure 7 depicts the difference between Motivational and Educational Factors compared to self-evaluations across Experiment 3.2. Although some portions of the graph may align with expected results, on the whole an analysis of the Factors we gathered to predict performance predictions did not seem to yield any results that were consistent with prior analyses of this type. Furthermore, any differences in the current analysis were of a very low magnitude (5% or less), as they were in Experiment 1, suggesting results may be entirely spurious in nature.

CHAPTER III

EXPERIMENTS 4-5: INTERVENTIONS

Experiment 4

It is clear that students become less confident about their knowledge after being tested, an outcome that often resulted in more accurate self-evaluations. Unfortunately this metacognitive improvement does not seem to be permanent as people seem to become more confident, and more overconfident, between tests. Across Experiments 1-3 a consistent pattern of results, both from laboratory and classroom studies, showed that people tend to exhibit a zig-zagging self-evaluative confidence level as they complete two tests, replicating the three features of the proposed Shifting Focus effect. Participants have become less confident from before tests to after tests (Feature 1), even when those tests are the second of their type or university semester to be taken (Feature 3). The rise in confidence between tests may be the cause of students' tendency towards overconfidence. Experiment 4 investigated the efficacy of an intervention designed to prevent rising confidence between in-class exams by requiring students to remind themselves of their past test performance and past prediction accuracy as they made predictions about future test performance between tests.

Methods.

Experiments 1-3 repeatedly showed that people become more confident in their self-evaluations during the period between two tests, both in laboratory and classroom conditions. Although the current evidence was rocky at best, it may be that such results occur because people are motivated to feel they will perform well in the future (Klein & Kunda, 1993, 1994; Saenz et al., 2017; Taylor & Brown, 1988) despite prior results to the contrary. Because of this natural confidence, *some* people may be prone to overconfidence, leading them to under prepare for exams (Dunlosky & Hertzog, 1998; Kruger & Dunning, 1999; Metcalfe, 2002; Metcalfe & Finn, 2008b), a dangerous yet common outcome. Experiment 4 was designed to help prevent people from becoming more confident between tests. The intervention consisted of repeatedly reminding students of their past grade predictions and test grades, while simultaneously asking students to make grade predictions during a Reflection Period. By making past calibration salient during the Reflection Period, and at the time of predictions, students may avoid incrementally shifting their grade predictions towards being more confident. There is evidence that this type of feedback may "strong" and salient enough to be effective in preventing overconfidence (Nietfeld et al., 2006; Saenz et al., under review).

Participants. Participants in this experiment consisted of 85 university students taken from an introductory psychology course in the fall of 2018. This sample size was limited first by the quantity of students enrolled in the course, and second by the quantity that consented to participate, but should still be sufficient based on the effect sizes and power analyses from prior studies. Demographic information was not collected and at the request of the instructors for Experiments 4-5. However, participants were of similar proportion to Experiments 1-3 in every way. This was a hybrid course, meaning that students participated in a regular lecture course once per week, and completed 2 class-periods worth of instruction online per week.

Materials and procedure. The materials and procedures were very similar to those used in Experiment 3.2. Although Experiments 3.2, 4, and 5 were each taught by a different professor, all three experiments involved two exams, taken about a month apart, which consisted of multiple-choice questions. These two exams were the third and fourth in the semester for the course. Participants completed the same self-evaluation surveys as in Experiment 3 before and after each of these tests. However, Experiments 4-5 involved an intervention during the Reflection Period. In Experiment 4, instead of simple Reflection Period Predictions, participants completed a larger survey designed to remind them of their prior performance and prediction accuracy. The survey consisted of three questions as follows: "What was your test grade on Test 3 in percent?", "What grade did you predict you would make on the test, right before you took it?", and, "What grade do you think you will get on the next exam?"

At the onset of this study, participants received a brief, verbal overview of the experiment procedures, and were informed of the possible benefit of making accurate grade predictions (e.g., they were informed, in a 5-minute lecture, that more accurate grade predictions may be beneficial to students because they may allow for better preparatory behavior like studying sufficiently). Students were asked to complete these surveys three times a week, following a 3-class per week schedule. Participants were reminded about the repeated predictions procedure once per week to promote participation, and in-class bonus points were offered for completing the majority of the study. Despite this, participant compliance was poor, so much of the available data is flawed. Given the four-week interval between class exams, participants were be able to complete up to 14 grade predictions during this Reflection Period. To put participant compliance into perspective, the 85 participants could have completed a total of 1190 google surveys during the Reflection Period, but a total of 130 (10.9%) were received. As in Experiment 2, repeated predictions made within about 48 hours of another prediction were not analyzed, resulting in 120 successfully completed intervention surveys.

Results and Discussion.

Procedure Checks. A series of analyses were performed to check for overconfidence levels. Participants' average test scores (M = 77.451, SD = 10.967) were similar to (anecdotally) average scores for the departments psychology tests. However, on average participants were not overconfident, producing calibration scores that were not significantly different from zero at Prediction 1(p = .071), or Postdiction 1 (p = .456). Calibration scores were significantly underconfident at Prediction 2 (t(65) = -2.513, p = .014, d_z = .309) and Postdiction 2 (t(65) = -4683, p < .001, d = .576). Although participants seemed to become less accurate and more accurate as the experiment progressed, Absolute Calibration scores were significantly inaccurate across all self-evaluation points (p < .001), suggesting that self-evaluation accuracy merely went from being slightly overconfident at Prediction 1, to slightly underconfident at Postdiction 2. These results suggest that the intervention used in this experiment successfully prevented students from increasing their self-evaluative confidence between tests.

There were not enough participants for analyses comparing low and high performance quartiles (N < 20 per cell), so median split analyses were run instead. The low performing students did not have significantly different Calibration scores from high performing students any point in the experiment ($p \geq .097$). In terms of Absolute Calibration, low performers were less accurate than high performers at Prediction 1 ($F(1,72) = 10.859$, $p = .002$, $\eta^2 =$, $M = 12.257$, $SD = 10.455$, and $M = 6.077$, $SD = 5.0125$) and Postdiction 1 ($F(1,72) = 5.578$, $p = .021$, $\eta^2 =$, $M = 10.853$, $SD = 11.995$, and $M = 5.750$, $SD = 4.789$), but not in either Prediction 2 or Postdiction 2 ($p \geq .501$). Although performance groups were not significantly different in Calibration, they were in Absolute Calibration. Results suggest that, before the intervention, all students were close to being accurate, but that low performers were slightly more variable and thus less accurate. Than high performers. After the intervention (Prediction/Postdiction 2), both low and high performers had achieved self-evaluative accuracy, and this accuracy was similarly variable.

Students' ability to recall their previous test scores and initial performance predictions was critical to the current intervention, so a series of analyses was performed to check for the accuracy of these during the Reflection Period. Paired samples comparisons revealed that there were no significant differences between average test grades, and the average test grade reported during the Reflection Period intervention ($p = .706$; Average Test 1 Grade: $M = 77.700$, $SD = 10.675$; Average

Reported Test 1 Grade: *M* = 77.365, *SD* = 12.745). No significant differences were found between average reported Prediction 1s (*M* = 78.856, *SD* = 10.658) and actual Prediction 1s (*M* = 79.100, *SD* = 15.216; *p* = .935). These data suggest that participants that did complete the intervention did so faithfully, and accurately reported their previous grades and self-evaluations, so they may have been at least somewhat affected by the intervention. Note, however, that participants completed an average of less than 3 Intervention surveys (*M* = 2.76, *SD* = 2.639) during the Reflection Period.

Intervention and Prediction Accuracy Over Time. Feature 1 was replicated (*t*(72) = 2.755, *p* = .007, d_z = .323), students were less confident at Postdiction 1 (*M* = 75.07, *SD* = 15.68) than at Prediction 1 (*M* = 78.67, *SD* = 13.119). Due to poor participant compliance, Reflection Period survey data varied between 1 to 22 responses at any one time point, so paired samples comparisons were difficult. The highest confidence points during the Reflection Period coincided with very low response rates (N < 10), so the Reflection Period peak was instead found through a series of paired samples comparisons. Indeed, the *only* significant change in self-evaluations from Postdiction 1 to a point during the Reflection Period (Feature 2), occurred at Reflection Period Prediction 1 (*t*(17) = -4.160, *p* = .001, d_z = -.980; *M* = 75.22, *SD* = 13.256; to *M* = 83.56, *SD* = 8.793). Although confidence did decrease significantly from the Reflection Period Peak to Postdiction 2 (Feature 3: *t*(16) = 4.847, *p* < .001, d_z = 1.176; *M* = 84.35, *SD* = 8.366; to *M* = 72.41, *SD* = 14.689), predictions did not seem to become more confident from Postdiction 1 to Prediction 2 (*t*(60) = -.422, *p* = .675, d_z = -.054; *M* = 75.74, *SD* = 16.300; to *M* = 76.48, *SD* = 13.070). Furthermore, students were actually *less* confident at Prediction 2 than at Prediction 1 (*t*(17) = -4.160, *p* = .001, d_z = -.980; *M* = 75.22, *SD* = 13.256; to *M* = 83.56, *SD* = 8.793).

The current results seem to align with expectations. Just a few days after their first test, students had begun to become more confident about their future performance. However, after the first intervention survey, students seemed to lose this non-adaptive confidence, and ceased gaining confidence for the remainder of the Reflection Period. Although these students did not necessarily make more accurate predictions at Test 2 than at Test 1 (Absolute Calibration paired samples comparison: *t*(17) = -4.160, *p* = .001, d_z = -.980; *M* = 75.22, *SD* = 13.256; to *M* = 83.56, *SD* = 8.793), they *did* have less overconfident at Test 2 than at Test 1 (Calibration paired samples comparison: *t*(17) = -4.160, *p* = .001, d_z = -.980; *M* = 75.22, *SD* = 13.256; to *M* = 83.56, *SD* = 8.793), and Test 2 scores were higher than Test 1 (*t*(17) = -4.160, *p* = .001, d_z = -.980; *M* = 75.22, *SD* = 13.256; to *M* = 83.56, *SD* = 8.793). Whereas different test scores cannot be directly attributed to the current intervention, it is possible that the reduced confidence elicited by the intervention may have increased study behavior in those students that participated. Further research might investigate a more causal link between this type of intervention and study choice behavior (but see research suggesting a link between metacognitive confidence and study behavior: Dunlosky & Hertzog, 1998; Metcalfe, 2002; Metcalfe & Finn, 2008b).

Key Aspect of the Intervention. A series of regressions were run to examine which aspect(s) of the intervention were critical to its effects. The difference between Postdiction 1 and Prediction 2 was used as the dependent variable, because it described the amount of confidence change students exhibited between tests. The first regression included only the quantity of intervention surveys successfully completed as an independent variable, but no significant relationship was found (*F*(1, 65) = .828, *p* = .366, R^2 = .013). Of the students who completed Postdiction 1 and Prediction 2, only about half (30 of 61) successfully completed any intervention surveys, so if completion of at least one or more interventions did not prevent confidence change, could the current results be merely spurious? A second regression was run to check what, if any, of the intervention aspects was responsible for preventing confidence change between tests. Four independent variables were entered into this second analysis: quantity of predictions, Absolute Calibration of estimated prior test performance, Absolute Calibration of estimated prior prediction, and Test Performance average. Estimations of prior test performance and predictions were entered because they were an integral part of the intervention surveys. However, these items were entered as absolute calibrations, or the absolute value of the difference between themselves and the actual data point, because participants may not have been entirely accurate in their estimations or recollections. In other words, these items represented student's accuracy in reporting their prior test performance and predictions during the intervention. Finally, average test performance was entered because there is evidence that higher performing students may be more susceptible to feedback and interventions than lower performing students (Kruger & Dunning, 1999). Of these four variables, only the Absolute Calibration of estimated prior test performance was significant (*unstandardized* β = .900, *p* = .016).

This suggests that the *accuracy* with which participants reported their prior test performance was crucial to the effect of the intervention.

Test Performance. A paired-samples comparison indicated that students obtained higher test scores after the intervention ($t(62) = -2.069$, $p = .042$, $d_z = .228$, $M = 76.220$, $SD = 11.151$; to $M = 78.683$, $SD = 13.204$). A regression was run using the difference in test scores as a dependent variable, to check what aspect of the intervention, if any, could explain test-performance improvement. The difference between Prediction 1 and 2, quantity of predictions, Absolute Calibration of estimated prior test performance, Absolute Calibration of estimated prior prediction, and Test Performance average were each entered into a regression, and nonsignificant predictors were removed one-by-one until only significant factors remained. The resulting regression model was significant ($F(2, 37) = 5.950$, $p = .006$, $R^2 = .243$), indicating that average Test Performance ($t = 3.421$, $p = .002$, Std. $\beta = .693$) and the Absolute Calibration of estimated prior test performance ($t = -2.111$, $p = .042$, Std. $\beta = -.427$) each accounted for a unique portion of test improvement. Whereas it may be unsurprising that higher performing students were more able to improve their test grades, it was promising to find that accurate recollection of one’s past test performance played a significant role in both preventing rising confidence, and improving test grades above and beyond test performance. Further investigation into the association between accurate performance recollection, accurate metacognition, and improving performance needed to determine a causal relationship. Although the change in test performance was not large in either number (about 2 points) or in effect size, it was promising to find improvement in test performance especially after such poor intervention compliance. Further investigation would be needed to determine the consistency of this outcome, and the size of its effect under fully compliant circumstances.

Factors Affecting Performance Predictions. Educational, and Motivational Factors were analyzed as in previous experiments (Figure 7). The results of this analysis were the closest to expectations across all studies. The relative influence of Motivational Factors seemed to be correlated with students' self-evaluative confidence. Although these results are promising, because they are only one in six experiments to follow expected results, they may only be due to chance.

Experiment 5

Experiment 4 investigated an intervention designed to prevent rising confidence in students between tests. Results suggested that participants who accurately recalled their prior test performance while estimating their future performance were less likely to become increasingly confident between tests. Although this intervention may have improved prediction accuracy and test scores, further data is required to establish the consistency and requirements of this outcome. Although the intervention in Experiment 4 was not particularly involved, it seemed to produce results despite poor participant compliance. Could a minimalist intervention, designed to help students associate prior test scores to their grade predictions produce similar results? Experiment 5 investigated this possibility by simplifying the intervention used in Experiment 4.

Methods.

Participants. Participants in this experiment consisted of 59 university students taken from a cognitive psychology course in the fall of 2018. This sample size was limited first by the quantity of students enrolled in the course, and second by the quantity that consented to participate, but should still be sufficient based on the effect sizes and power analyses from prior studies. Demographic information was not collected and at the request of the instructors for Experiments 4-5. However, participants were of similar proportion to Experiments 1-3 in every way. This was regular lecture-based course for upper-level students.

Materials and Procedure. Experiment 5 can be conceived of as an extremely minimalistic version of Experiment 4. Participants completed two in-class tests as part of their regular course schedule. Participants completed four self-evaluation surveys, one immediately before and after of the two tests, again described as Prediction 1/2 and Postdiction 1/2. The surveys were the same used in Experiments 3-4 (very similar to Appendix B). Between the two exams, students had their grade predictions (Prediction 1) presented alongside their test grades using the online system normally used to communicate grades to students. This was the only intervention measure, and students completed a series of predictions alongside the next exam in their class.

Results and Discussion.

Procedure Check. A series of analyses were performed to check for overconfidence levels and appropriate manipulation conditions. In terms of Calibration, students were not significantly inaccurate at any self-evaluation point ($p \geq 1.01$) except

Postdiction 1, where participants were significantly *underconfident* (*t*(43) = -3.864, *p* < .001, d_z = -.583, *M* = -4.729, *SD* = 8.118). Indeed, students were underconfident throughout the study, potentially because test scores (test 1: M = 83.692, *SD* = 8.360; test 2: *M* = 84.578, *SD* = 7.716) were anecdotally higher than average test scores for the department and course level (usually closer to 77). Despite presenting underconfidence, the current results seem to align with expected outcomes, as described in the following sections. Significant Absolute Calibration inaccuracy was found at all self-evaluation points in the current experiment (*p* < .001).

Intervention and Prediction Accuracy over Time. To evaluate the efficacy of the current intervention, predictions were first analyzed to check for the three features of the Shifting Focus effect. Students replicated prior experiments in exhibiting a significant decrease in confidence from Prediction 1 to Postdiction 1 (*t*(45) = 4.599, *p* < .001, d_z = .678 *M* = 82.35, *SD* = 10.800 to *M* = 77.78, *SD* = 12.094). No Reflection Period Predictions were made during Experiment 5, so Features 2 and 3 could not be analyzed in the same way as prior studies. However, students did not become more confident from Postdiction 1 to Prediction 2 (*t*(32) = -1.102, *p* = .279, d_z = -.192, *M* = 79.61, *SD* = 10.335 to *M* = 81.55, *SD* = 10.583) or become less confident from Prediction 2 to Postdiction 2 (*t*(35) = -.359, *p* = .722, d_z = -.060, *M* = 80.97, *SD* = 10.750 to *M* = 81.39, *SD* = 13.085). Students did not become significantly more confident between in-class tests. Further analyses corroborated these results, indicating that participants did not become less accurate from Postdiction 1 to Prediction 2 in terms of either Calibration (*t*(31) = -1.303, *p* = .202, d_z = .231, *M* = -4.535, *SD* = 8.721

to *M* = -2.223, *SD* = 10.036) or Absolute Calibration (*t*(31) = -.159, *p* = .874, d_z = .028, *M* = 7.387, *SD* = 6.399 to *M* = 7.588, *SD* = 6.811).

Checking One's Grade. The current intervention involved presenting prior grade predictions alongside prior test grades such that a demonstration of students' prediction accuracy may have been salient whenever they thought or viewed their test grades. Because this information was presented online, data were gathered on how many times students checked their grades. Although about half of the participating students did not log in to view their grades (N = 24 of 52 valid data points), students averaged about three views per person (*M* = 3.06, *SD* = 5.177; although these data were positively skewed, and two participants were excluded as extreme outliers (viewed their grades 54+ times). Postdiction 1 was subtracted from Prediction 2 to measure the change over time, and ran a correlational analysis to the resulting variable with the amount of grade views. After the removal of outliers (two or more SD's outside average in grade views or prediction change) no significant relationship was found (*r* = .064, *p* = .736, N = 30).

Test Performance. A paired-samples comparison did not find higher test scores after the intervention for the full sample (*t*(55) = -1.089, *p* = .281, d_z = .147, *M* = 83.692, *SD* = 8.360; to *M* = 84.849, *SD* = 7.514) or for participants who checked their grade at least once (*p* = .351). Although not predicted, this outcome was not entirely unsurprising as the effect size, and sample size of Experiment 5 were lower than Experiment 4. Further investigation is necessary to determine whether a minimalistic intervention can help improve test performance as well as metacognitive accuracy.

CHAPTER IV

GENERAL DISCUSSION

At the end of every semester, instructors and academic advisors must work with students who did not achieve their academic goals. Losing a scholarship, bombing a course, or even missing graduation, these innumerable students join the millions of yearly college dropouts who fail to achieve their higher education goals (data from the U.S. Dept. of Education). The current experiments investigated one of the possible causes for these outcomes: a tendency for people to become more confident between tests, potentially leading to overconfidence and underperformance in the classroom.

When people predict how well they will perform a task, they often exhibit overconfidence (Bol et al., 2005; Foster et al., 2016; Hacker et al., 2008; Miller & Geraci, 2011a, 2011b; Nietfeld et al., 2005). In students, this overconfidence has been prevalently observed and linked to low test performance (Everson & Tobias, 1998; Kelemen et al., 2007; Shepperd, 1993). In response, researchers have investigated a variety of interventions designed to improve student's self-evaluative accuracy, with inconsistent success (Hacker et al., 2008; Foster et al., 2016; Nietfeld et al., 2005, 2006). Ideally, interventions for improving self-evaluative accuracy should have staying power, that is, they should help students improve their self-evaluations over multiple tests, and ideally, across multiple semesters or domains of knowledge. One problem with these interventions is our limited understanding of the way in which these self-evaluations are made. How do students go about predicting or postdicting their test performance? And what circumstances or mental strategies lead students to potentially harmful overconfidence? Going further, the human mind is extremely adept at learning, and the proper evaluation of behavioral results should be key to this learning ability, so why would people exhibit a tendency towards overconfident, inaccurate self-evaluations?

I propose that overconfidence during performance predictions may actually be an adaptive mental habit that allows for both accurate self-assessment after a test, as well as sufficient confidence in one's abilities to try again during the next test. Imagine a novice juggler, who fumbles his first performance. Should this person evaluate themselves accurately, they would correctly identify their incompetence, but also lose the confidence to attempt a second performance in the future, and give up practicing, believing themselves a failure. Should the person incorrectly evaluate themselves to be a master (ignoring their results) they might have the confidence to continue performing but may neglect further practice, leading to a different failure state of perpetual incompetence. It may be then, that a fluctuating level of confidence is necessary for people to improve themselves, to prepare for future performance despite past failures. As time passes, people may *shift the focus* of their self-evaluations from past failures, to future desires.

Research Questions and Summarized Results

Why are test-performance predictions inaccurate and often overconfident? Can the pattern of prediction accuracy fluctuation found in the preliminary data (the Shifting Focus effect) be replicated? The current experiments provide evidence for a fluctuating, zig-zagging self-evaluative confidence level both in the laboratory and in the classroom (e.g., Figure 4). These fluctuations present three distinct and consistent features: a decrease in confidence from predictions to postdictions for an initial test (Feature 1: Prediction 1 to Postdiction 1), an increase in confidence from an initial test to a point before a subsequent test (Feature 2: Postdiction 1 to Reflection Period Peak), and a final drop in confidence from a point between two tests, to after the second exam (Feature 3: Reflection Period Peak to Postdiction 2). These three features comprise a Shifting Focus effect, and were replicated across Experiments 1-3. Additional comparisons may be made by comparing predictions for two tests, which might exhibit no change, and therefore indicate students did not learn anything from taking the first test. One may also compare postdictions from one test to predictions from a subsequent test, which might exhibit increased confidence, and indicate that metacognitive learning from one test was somehow lost by the next one. These comparisons were considered secondary because they may be more conservative and or inconsistent than the three primary features. They may be more conservative or inconsistent because they require increases in confidence to increase to the same or a greater level than confidence before the first test, whereas Features 2 and 3 are more liberal because they can detect any increase in confidence during the Reflection Period.

Across Experiments 1-3, participants made more confident and less accurate predictions than postdictions, suggesting that they learned something about themselves by taking a test. However, self-evaluations were no different before the second test than before the first test, indicating that participants somehow lost the accuracy improvement gained from taking the first exam. Experiments 1-3 also showed that people incrementally raised their self-evaluative confidence between tests, until arriving at the same confidence level they had before the first test.

Are these results robust to the particular prediction type (absolute vs percentile)? The current experiment focused on absolute predictions, or self-evaluations where one estimates a grade. Another way to evaluate oneself is to estimate one's performance in terms of percentile ranking. Although percentile predictions followed similar patterns to absolute predictions, percentile predictions were not consistently of greater or lesser magnitude than absolute predictions. These results may be unsurprising as existing data suggest that percentile predictions may be more variable, and potentially more difficult to successfully produce than absolute predictions (Hartwig & Dunlosky, 2014; Tirso et al., 2019). In addition, a major quality of the Reflection Period (the time between tests) was that participants were encouraged to think about their future test performance in terms of absolute estimates. It may be that shifts in percentile evaluations would follow from asking people to think about their future performance in terms of percentile ranking as well.

What factors affect these fluctuations? Prior research shows that motivational information, such as the ideal grades, is associated to greater confidence in self-evaluations (Saenz et al., 2017; Serra & Demarree, 2016). Contrarily, educational factors like past test performance should be related to better accuracy, because past test performance is predictive of future test performance (Jensen & Barron, 2014). Although the fluctuation of confidence across multiple tests paints a slightly more complete picture of self-evaluative confidence, it does not inherently describe what mental processes are behind it. So, the current experiments investigated whether these fluctuations could be explained by changes in the factors one focuses on when making self-evaluations. Figure 7 summarizes these data. Results did not reveal any consistent trend in the relationship of motivational factors, educational factors, and self-evaluations. Despite this, it is difficult to rule of the role of these factors entirely, given the findings of prior research. The current investigation into these factors was limited in several ways. Data on these factors was not gathered during Reflection Periods, so analyses for these time points was flawed. Furthermore, the prior research involving these factors mainly focused on in-class studies, and should only be valid for experiments without interventions, conditions that were only met in Experiment 3. Although there are many available methods for assessing the contributing factors of self-evaluative judgements, there were very few instances of data where these methods may be compared. Further research may focus on using validated ways of measuring motivational, educational, another factors to examine their relationship to self-evaluations.

The educational factors analyzed in the current experiments may not have been entirely reasonable. These educational factors involved students' studying habits and class attendance, information that *seems* related to grade predictions, but may not necessarily be, especially in a laboratory setting. Furthermore, the data from the current Shifting Focus effect replications indicates that confidence changes as time elapses from a prior test, so instances where participants may have chosen to study, or attend class were not involved. Instead, further study may focus on the fluency, and recollection of past test performance and feedback, information that *must have been available* to participants across all current experiments (at least in the form of self-feedback during testing). Other possible explanations for the Shifting Focus effect are further described in the "Connecting the Shifting Focus effect to Existing Literature" section.

An additional set of items, Anchoring Factors, was included in the analyses of this experiment to observe the extent to which self-evaluations might be made arbitrarily. These items were not designed to observe whether participants made their self-evaluations randomly, but instead, to observe the extent to which reasonable, though impersonal information, like perceived class-test-averages, might inform or anchor self-evaluations. Although it makes sense to base one's self-evaluation around the class average, that number is not necessarily associated to one's own performance level, and may lead participants to make metacognitive mistakes. Although the current experiments were, at best, only an initial investigation of Anchoring Factors, they *were* found to be consistently strongly related to self-evaluations to the extent that they may have overshadowed other factor groups. Further investigations might research the extent to which self-evaluations are made arbitrarily, and the circumstances that may affect or elicit anchoring behavior in self-evaluations and other metacognitive judgements.

Are the answers to these questions consistent across different laboratory and applied settings? Are changes in prediction accuracy more strongly related to the amount of *time* between predictions, or the amount of *thought* put into those predictions? Figure 8 summarizes the effect sizes of the three main Shifting Focus features, as well as the difference between Prediction 1 and Prediction 2. In brief, the features of the Shifting Focus effect were replicated in both laboratory and classroom studies. Interestingly, the Shifting Focus effect (particularly Feature 2) was most pronounced in the classroom setting, with the longest Reflection Period, and the fewest Reflection Period Predictions. Additionally, it was difficult to say that either time or thought (amount of predictions) was more impactful. Although longer periods of time, and fewer predictions had greater effect sizes, these outcomes had two caveats. First, the Reflection periods occurred during the in-class experiments, so that may confabulate results. Second, Confidence seemed to peak somewhere between 2-7 days into the reflection period. However, this rising period may be variable based on the reflection period size, as a sort of peak and plateau was reached in experiment 1. Furthermore, data were not gathered daily in experiments 2-3, so the exact "reflection period peak" cannot be determined with the current data. Implications and possible explanations for these outcomes are further discussed in the effect size comparisons section below.

Can interventions to improve self-evaluations across multiple tests be developed based on what we know about the causes of inaccuracy in predictions? Can such interventions be implemented with minimal interference in the classroom? Experiments 4-5 investigated this question, and their evidence suggests that increases in confidence between tests can be prevented by the presentation, recall, and/or association of prior test scores to future performance predictions. Of note, results suggested that the key factor in these interventions was the pairing of *accurate* prior tests results to prior or future self-evaluations. In other words, people who incorrectly recalled their test scores as being higher than they actually were, did not benefit from thinking about their test scores as an intervention, and went on to become more confident (See also the discussion on discounting feedback in following sections). Interestingly, the minimalistic intervention in Experiment 5 may have avoided this issue, by directly pairing past test performance with past self-evaluations: students who checked their grades were required to notice the accuracy of their predictions. Experiments 4-5 were severely limited by participant compliance, and many of the analyses therein may be underpowered, so further study and replication is needed to verify their results and implications.

Do such interventions interact with the underlying factors in expected ways? As in Experiments 1-3, analyses of the Motivational and Educational Factors involved in self-evaluations across Experiments 4-5 did not yield either consistent or expected results. Summarized in Figure 7, regression analyses for Experiment 4 seemed to follow exactly hypothesized trends: participant’s self-evaluations seemed to be more motivationally oriented when they were more confident, and more educationally oriented when they were less confident. However, the same analysis in Experiment 5 yielded an exactly inverse relationship. It is difficult to make any inferences about these factors based on these data, though it is interesting to note that the most dramatic changes in these factors occurred in Experiment 4, and aligned with expected results.

Did the interventions in Experiments 4-5 improve test performance? Although such an outcome would be ideal, it is difficult to make this conclusion. No evidence of such a change was found in Experiment 5, and although there was improvement in Experiment 4, the improvement was small. Still, being applied studies, it is difficult to interpret data from these experiments definitively. Results may have been entirely due to differences in test difficulty, rather than intervention conditions. It was promising to find test score recollection accuracy was a consistently important factor in Experiment 4, both for prediction accuracy, and for test score improvement. Some studies have found that more accurately reported SAT/GPA scores are associated with higher academic performance (Everson & Tobias, 1998; Kelemen et al., 2007; Thiede, 1999; Thiede, Anderson, & Therriault, 2003). These data may therefore support a Memory for Past Test performance account of shifting metacognitive confidence (further discussed below).

Connecting the Shifting Focus Effect to Existing Literature

Accuracy vs. Confidence.

To oversimplify the Shifting Focus effect, people seem to become less confident after a test of their abilities, but regain that confidence before the next test. Parts of this effect has been observed and evidenced in many ways, and many of these related effects focus on the accuracy of self-evaluations instead of their absolute level of confidence. For example, the Postdiction Superiority effect proposes that postdictions are more accurate than predictions when estimating one’s knowledge after reading a given text (Pierce & Smith, 2001). The Underconfidence with Practice effect (Koriat & Ma’ayan, 2002; See also the Memory for Past Test heuristic: MPT, Finn & Metcalfe, 2007) focuses more so on memory, and purports that people become underconfident after memory tests involving feedback. Effects like these describe a circumstance where people become more accurate in their self-evaluations after being tested, often accompanied by a reduction in confidence relative to performance (i.e., from overconfidence towards underconfidence).

The current Shifting Focus effect does not focus on self-evaluative accuracy (Calibration or Absolute Calibration), but instead on how people to evaluate themselves across time (changes in confidence). Although over- or under-confidence may often be associated to lower and higher metacognitive accuracy respectively, the two items are not necessarily linked (e.g., if everyone is underconfident, higher confidence would be more accurate). This distinction was evidenced by the fact that the Shifting Focus effect did not seem to interact with test performance. For example, lower and higher performing participants were affected equally by the three Features of the Shifting Focus effect, and during an intervention to eliminate the Shifting Focus effect (Experiment 4), test performance levels were not a moderating factor.

Much metacognitive research has focused on lower and higher performing students because lower performing students tend to exhibit greater overconfidence, and less self-evaluative accuracy (Bol et al., 2005; Hacker et al., 2008; Miller & Geraci, 2011a, 2011b; Kruger & Dunning, 1999; Saenz et al., 2017). Many researchers have focused on the accuracy of self-evaluations because it presents a readily apparent goal: achieving accuracy. However, because the current experiments showed that confidence may change irrespective of self-evaluative accuracy, and because the current and prior studies show that self-evaluations may be affected by anchoring (Ferrel & McGoey, 1980; Burson et al., 2006; Klayman et al., 1999), future investigations may seek to interpret existing metacognitive research in terms of simple self-evaluations or predictions, instead of focusing on accuracy. Though I do not propose that accuracy based accounts are inaccurate, I do propose they may be incomplete[2].

[2] Theories and hypotheses like the Postdiction Superiority Effect and the Underconfidence with Practice Effect specify different reasons for change in self-evaluative accuracy, but do not describe how self-evaluative behavior changes, and instead focus on the level of function or dysfunction of metacognitive behavior. By analyzing changes in self-evaluative accuracy (such as by Calibration or Gamma Correlations) instead of overall confidence level, one must necessarily bundle two different data points: people's confidence and people's performance scores. Some of the research in this area involves multiple testing trials on the exact same material, which may result in greater changes in test performance than self-evaluations. In other words, effects like the Underconfidence with Practice Effect might rely, at least partially, on a resistance to change in self-evaluations, where performance over multiple trials improved more drastically than people were willing to change their predictions (e.g. Figure 9, from Koriat 1997). These types of apparently inaccurate metacognitive judgements may therefore be considered statistical artefacts by some (i.e., Kelemen et al., 2007) or a form of inaccuracy through anchoring.

An interesting comparison one may make to these theories of metacognitive judgements is to consider how certain pieces of information affect self-evaluations and their accuracy. For example, one explanation for the Underconfidence with Practice Effect is that people become more metacognitively accurate, and less confident, as they are made more aware of their past test performance (the Memory for Past Test Performance Effect). Indeed, such an explanation fits with other accuracy changing outcomes. It seems reasonable that *postdictions* might be *superior* because when making a postdiction, people have just experienced a test, so their memory for their past test performance should be fresh and salient (even if that memory relies on self-evaluated performance during the test, i.e., “how well you felt you did”). Feature 2 of the Shifting Focus Effect may therefore result from the opposite condition. As mental processes work, or as time passes, memory for past test performance may deteriorate. However, there is evidence that people, and students in particular, do remember their past test scores accurately (e.g., Study 4), or are at least somewhat aware of their prior test scores (Miller & Geraci, 2011a). Instead, it may be that the fluency, or saliency, or mental connections between past test performance and future performance may weaken as a function of something that happens during the time between tests, the Reflection Period. Indeed, the interventions in Experiments 4-5 may have been effective specifically because they strengthened this connection or fluency during the Reflection Period. Further investigation is required to ascertain whether the mechanism causing these two types of effects (Features 1 and 2)can be attributed to the same broader concept.

Effect Size Comparisons.

Increases in self-evaluative confidence between tests were found across multiple settings in Experiments 1-3. Experiment 1 gave participants 10 minutes to make 10 predictions in a laboratory environment, Experiment 2 gave participants one week to make 7 predictions in a laboratory/home environment, and Experiment 3 gave participants four weeks to make 4 predictions in a classroom environment. Although Experiments 1-3 each found significant changes in grade predictions for all three of the features hereto described, the effect sizes of these features may help us understand how they work. Figure 8 outlines the effect sizes for Experiments 1-5 and shows a few interesting comparisons. The first Feature, a decrease in confidence from Prediction 1 to Postdiction 1 seems greater in Experiments 1 and 2 than in Experiments 3, 4, or 5. This difference seems to be most likely a demand characteristic of these experiments as the test used in Experiments 1-2 (the laboratory studies) was made purposefully difficult to elicit greater overconfidence, and so should result in a greater change in predictions. Additionally, Experiments 3-5 were performed in the classroom, a setting where participants would naturally have a better idea of the test difficulty and content. The second feature of these experiments was more interesting, however, as it encapsulated the rise of self-evaluative confidence during the Reflection Period, specifically from Postdiction 1 to the highest Reflection Period confidence point. The effect size of Feature 2 seemed to become larger from Experiments 1 to 3, and the specific conditions of those experiments might explain this outcome.

I propose three possibilities. First, Experiments 1 to 3.2 each involved incrementally longer Reflection Periods, and fewer self-evaluations to make during the Reflection Period, so it may be that time is a driving force behind increases in self-evaluative confidence, and that the amount of predictions is less important. It may be that people become more *self-confident* as *more time* passes but not necessarily as they *made more predictions* or thought harder on their predictions (but see Buratti & Allwood, 2012, 2013). However, this explanation is flawed as in Experiment 2 and 3.2, self-evaluative confidence hit its peak early on in the Reflection Period (Experiment 2: day 2; Experiment 3.2: Week 1), suggesting that the length of time that passes is important, but only as long as people are given at least 2-7 days to consider their judgements, with longer periods of time either lowering or not affecting confidence during the Reflection Period. At best, this explanation can explain why Experiment 1 has the lowest effect size, but not why Experiment 3.2 has a higher effect size than Experiment 2.

Second, it may be that the confidence changes between tests are most strongly affected by the importance and motivational circumstances of those tests. That is, Experiment 3.2 was conducted in a classroom, and a motivational account of the Shifting Focus effect would expect changes in confidence to be caused a greater focus on motivations when making self-evaluations. Put simply, because participants in Experiment 3.2 were students, they may have experienced greater motivation to feel confident that participants in Experiments 1-2, leading them to experience a greater shift in confidence during their Reflection Period. Although the effect sizes of Feature

3 seem slightly difference between Experiments 1-2 and Experiment 3, such differences may be explained by the difference in Feature 2. That is, if participants' confidence rose more in Experiment 3 during the Reflection Period (Feature 2), then they have more room to lose confidence after their test (Feature 3). But this explanation may also be flawed as the analysis of factors in the previous experiments has provided minimal (at best) evidence for an explanation based on changes in the motivational or educational nature of self-evaluations. Furthermore, Experiments 4-5 did not seem to show these motivational differences either. Although prior research does seem to suggest a significant role of Motivational Factors in the inaccuracy of performance predictions, it may be that this role is in addition to, or aside from changes in prediction accuracy over time. Although this explanation might explain the difference between Experiments 2 and 3.2, the analysis of prediction Factors in the previous and the following results sections does not seem to support this explanation (but see following discussion on how different types of Motivational Factors may be important).

Third, and potentially most parsimonious, it may be that variance in predictions was simply lesser in the classroom, resulting in a greater apparent impact of prediction change. Generally, laboratory studies are performed because they help eliminate unnecessary variability in data, but the current experimental procedures may have actually elicited a more variable dependent measure in the laboratory setting. Indeed, the standard deviations for self-evaluations in Experiment 3.2 were numerically smaller than those of Experiments 1-2, potentially because they involved real students and tests. It seems reasonable that a prediction for a test you have studies for would be less arbitrary than a prediction for a test you are unfamiliar with, such as those we administered in the laboratory. Further investigation would be needed to confirm this difference in variability, so it is difficult to conclude that the "artificiality" of the prior studies can be used to explain the differences in effect size between them. Although there may be other explanations for these differences in effect size, it also seems reasonable that a *combination* of the above mentioned may be the best explanation. It is also important to note that interpreting differences in effect size from a single set of experiments may be premature.

Still, effect sizes of the features in Experiments 4-5 may corroborate these and other interpretations of the current studies. The effect size of Experiment 4, Feature 2 fell somewhere between Experiments 2 and 3.2[3]. This result may be unsurprising because, although participants did have a Reflection Period of at least 2-7 days, and would have been affected by the motivation of their applied setting, most participants did not much of the intervention procedure for Experiment 4. One might consider that a high and significant effect size for Features 2-3 in Experiment 4 indicates a failed intervention. Why did Experiment 4 have a Feature 2, despite having an apparently successful intervention? While, yes, confidence reached a "peak" in Experiment 4, it also gradually declined before the following test. Any confidence gained in the first week of the Reflection Period, was lost by the time of the next test. Furthermore, the negative effect size of the Postdiction 1 – Prediction 2 comparison suggests a successful prevention (or maybe quelling?) of rising confidence during the Reflection Period. Indeed, the Reflection Period Peak for Experiment 4 occurred, as in

Experiments 2-3, approximately 2-7 days after Postdiction 1, due to the self-initiated nature of the intervention used therein. It may be that the effect size of Experiment 4, Feature 2 shows how students reached a peak level of metacognitive confidence before the intervention began to affect them (as the first of intervention procedures would have been completed, at best, just moments before that peak confidence was reported).

Motivation as a Reason for Confidence.

A number of literatures have connected motivation and confidence. Some of these have already been described in supporting hypotheses and expectations for the current studies, but still others can aid in interpreting the current results. For example, an externalization account of overconfidence (Hacker & Bol, 2004; Bol et al., 2005; Hacker et al., 2008a, 2008b) states that lower performing students are overconfident because they blame their poor performance on outside factors. Evidence indicates that low performers are more likely to endorse statements like, "The test didn't really cover the things we covered in class" than "I didn't study the right things in the right way." These studies propose that because lower performing students attribute poor performance to external factors, those students tend to discount feedback, like prior test scores, leading them to maintain overconfidence on later tests. This self-fulfilling cycle of poor performance, and discounted feedback has been suggested and evidenced in a number of ways, such as the external locus of control literature (e.g., Seligman, 1991), and the Mindset literature (Dweck, 2008; See also Elliott & Dweck, 1988 for a look at connecting goals and motivation and Martin, Bostwick, Collie, & Tarbetsky, 2017 for a review). Relatedly, the Memory for Past Test performance literature (i.e., MPT, Finn &

Metcalfe, 2007) proposes that people become more accurate in their metacognitive judgements (and eventually "underconfident with practice", i.e., Koriat & Ma'ayan, 2002) as their memory for past test performances and feedback is reinforced or made fluent through repetition and framing. Together, these findings suggest that a forgetting or discounting of feedback explanation can produce overconfidence.

How do these ideas fit in with current findings? On one hand, it is difficult to say that the current entirely data can accommodate a simple discounting explanation because Experiments 1-2 showed a gradual increase in confidence, so participants probably did not ignore their prior test performance outright. On the other hand, one cannot completely ignore the possibility that people discount negative feedback. It seems possible that people might discount negative feedback over a period of time, letting the "sting" of failure fade away. Data from Experiment 4, where only people who accurately remembered their prior test scores improved their metacognition, also supports the notion of a discounting hypothesis. Indeed, because the motivational and educational factors gathered as part of the current dissertation did not load onto confidence change consistently, a gradual discounting of feedback might be a better explanation.

It is important to note strong parallels between the Self-efficacy /-regulation literature, the currently described motivational account, and the Shifting Focus effect. Self-efficacy encompasses a wide range of ideas, and has been described as, "people's beliefs in their ability to influence events that affect their lives" (Bandura, 2010, pp. 1) as well as, "judgements of one's capabilities to do [an] academic task" (Pintrich, 1999, pp. 462). Despite the wide-ranging ideas encompassed by Self-efficacy, it is clear that it should inform discourse on students' confidence in their learning. People with high levels of Self-efficacy tend to "approach" difficult tasks, and attribute their own failure to poor preparation or effort. In other words, people with better self-efficacy tend towards metacognitive accuracy, and people with worse self-efficacy tend towards overconfidence. Another parallel is that poor self-efficacy may be associated to a maladapted form of motivations that encourage one to focus on optimistic possibilities (Bandura, 2010).

Self-regulation focuses on the extent to which people are able to regulate aspects of their learning, and encompasses the choices students make about studying, as well as *how* they make those choices. For example, research into Self-regulation has found a number of "best practices" for providing students with feedback (Nicol & Macfarlane-Dick, 2006). These suggestions have focused on encouraging and improving Self-reflection, such as performing self-assessments that seem to mirror the grade predictions participants completed in the current dissertation. Whereas Experiments 1-3 asked participants to focus on prior test performance, it may instead be beneficial to encourage students to reflect on their current learning during the time between tests (i.e., Nietfeld et al., 2006). By focusing on current learning, motivational biases (desired grades and/or externalizing) may also be avoided, resulting in more accurate metacognition (See also Schunk & Zimmerman, 1994; Steadman, 1998).

CHAPTER V

CONCLUSIONS

Students are known to be consistently overconfident in their self-evaluations or metacognition about their future test performance (Bol, et al., 2005). This difference between students' self-evaluations, and their actual performance (termed calibration) may be critically harmful to students as it may lead them to underprepare for exams. Many investigators have designed interventions aimed at improving students' calibration, with the ultimate goal of improving students' test performance. Unfortunately, these interventions have had inconsistent success, and have not elicited lasting improvement (Hacker et al., 2008). Further, these interventions may be limited by a poor understanding of the causes of overconfidence in students (Saenz et al., 2017).

My preliminary data showed that student's grade predictions may rise between tests, without any new information being learned, so people may be *Shifting the Focus* of their grade predictions during the time between tests (Figure 2). Furthermore, recently published evidence shows that students make overconfident predictions at least in part because they use optimistic, motivationally oriented information instead of realistic academically based information (Saenz et al., 2017). In combination, these findings suggest a Shifting Focus effect may be causing overconfident performance estimations, and that this Shifting Focus phenomena may be associated to motivational biases. The current experiments replicated this Shifting Focus effect, explored the possibility of a motivational explanation, and tested interventions targeting overconfidence, by preventing Shifting Focus effects.

Experiments 1-3 replicated the preliminary data, showing that people become more confident in their self-evaluations during the period between tests. These results were robust to a number of circumstances, and were found both in and out of the classroom, as well as over periods of time ranging from minutes to weeks. Although results did not suggest that greater confidence was associated to a greater motivational bias and a weaker influence of educational information, other possible explanations for the features of the Shifting Focus Effect seemed to be reinforced by the data. For example, two interventions (Experiments 4-5) showed that reminding students of their prior test grades and connecting that information to self-evaluations could successfully prevent increases in self-evaluative confidence between tests. It may therefore be that the time between tests evokes a process, wherein people become more confident in their future test performance by disassociating (or discounting) past test performance from self-evaluations. Such a process may be unsurprising and adaptive because it would allow people to evaluate themselves accurately, and maybe even pessimistically after a test or performance, but become more confident over time so that they can try again in the future. In apparent corroboration of this possibility, rises in confidence seemed to peak about 2 days after an initial test, a process not-unlike the activation and return to baseline of ions in a neuron. It may be that people are normally confident to a certain degree, such that they are willing to attempt tests of their knowledge or ability. Taking these tests or doing these performances might jolt them out of this confidence, like an electrical impulse in a neuron. Similarly the resulting state may be imbalanced, and like the ions in a neuron, people might naturally become more confident until they are ready to fire again (i.e., take another test). Unfortunately for delayed-result behavior, the same confidence that allows one to re-attempt a test, might dissuade one from practicing for that test; if a person feels they are prepared to perform a task, why would they increase their preparatory behavior?

These results may have some major implications for metacognitive research, particularly as it relates to education. First, the expected results may require future investigations into metacognition to consider that overconfidence is not caused by faulty metacognition, but an adaptive form of motivated optimism. For example, existing theories explaining inaccurate metacognition such as the unskilled and unaware theory (a.k.a. the Dunning-Kruger theory), must either be amended to include evidence for a Shifting Focus effect in metacognitive judgements, or somehow account for changes in self-evaluations over time. Further intervention investigations may need to pay attention to the *time* in which interventions are presented, in addition to their design. This is not to say that the proposed intervention designs may be the only way to improve calibration in students. However, future interventions may also need to first prevent a shift towards using motivational information, before attempting to improve self-evaluation efficacy in other ways.

REFERENCES

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In Alicke, M. D., Dunning, D., & Krueger, J. I. (Eds.) *The Self in Social Judgment*, 85-106.

Alicke, M. D., Vredenburg, D. S., Hiatt, M., & Govorun, O. (2001). The “better than myself” effect. *Motivation and Emotion*, 25, 7–22

Bandura, A. (2010). Self-efficacy. *The Corsini encyclopedia of psychology*, 1-3.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117, 871-915.

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

Buratti, S., & Allwood, C. M. (2012). Improved realism of confidence for an episodic memory event. *Judgment & Decision Making*, *7*.

Buratti, S., & Allwood, C. M. (2013). The effects of advice and “try more” instructions on improving realism of confidence. *Acta psychologica*, *144*, 136-144.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*, 60-77. DOI:10.1037/0022-3514.90.1.60

Cambria, J., & Guthrie, J. T. (2010). Motivating and engaging students in reading. *New England Reading Association Journal*, *46*, 16-29.

Devolder, P. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults. *Psychology and Aging*, *5*, 291. DOI: 10.1037//0882-7974.5.2.291

Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy production? *Psychology and aging*, *13*, 597. DOI: 10.1037/0882-7974.13.4.597

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

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

Dweck, C. S. (2008). *Mindset: The new psychology of success*. NY: Random House Digital, Inc.

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of personality and social psychology*, *54*(1), 5.

Everson, H. T., & Tobias, S. (1998). The ability to estimate knowledge and performance in college: A metacognitive analysis. *Instructional Science*, *26*, 65-79. DOI: 10.1037/0022-3514.77.6.1121

Feather, N. T. (1968). Change in confidence following success or failure as a predictor of subsequent performance. *Journal of Personality and Social Psychology*, *9,* 38.

Feather, N. T. (1969). Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance. *Journal of Personality and Social Psychology*, *13,* 129.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26*, 32-53. DOI: 10.1016/0030-5073(80)90045-8

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 238.

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: 10.1007/s11409-016-9158-6

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11,* 702. DOI: 10.1037/0278-7393.11.1-4.702

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, *55,* 96-135. DOI: 10.1257/jel.20151245

Hacker, D. J., & Bol, L. (2004). Considering the Social-Cognitive Influences. *Big theories revisited*, 275.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*, 101-121. DOI: 10.1007/s11409-008-9021-5

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

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. *Handbook of metamemory and memory*, *429-455*.

Hartwig, M. K., & Dunlosky, J. (2014). The contribution 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

Jensen, P. A., & Barron, J. N. (2014). Midterm and first-exam grades predict final grades in biology courses. *Journal of College Science Teaching*, *44*, 82-89.

Kelemen, W. L., Winningham, R. G., & Weaver III, C. A. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology*, *19*, 689-717. DOI: 10.1080/09541440701326170

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

Human Decision Processes, *79*, 216-247. DOI: 10.1016/0001-6918(91)90036-Y

Klein, W. M., & Kunda, Z. (1993). Maintaining self-serving social comparisons:

Biased reconstruction of one’s past behaviors. *Personality and Social Psychology Bulletin, 19*, 732–739.

Klein, W. M., & Kunda, Z. (1994). Exaggerated self-assessments and the preference for controllable risks. *Organizational Behavior and Human Decision Processes, 59*, 410–427. 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.

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

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. DOI: 10.1037/0022-3514.77.6.1121

Klein, W. M., & Weinstein, N. D. (1997). Social comparison and unrealistic optimism about personal risk. In B. P. Buunk & F. X. Gibbons (Eds.), *Health, coping, and well– being: Perspectives from social comparison theory* (pp. 25–61). Hillsdale, NJ: Erlbaum.

Lench, H. C., & Bench, S. W. (2012). Automatic optimism: Why people assume their futures will be bright. *Social and Personality Psychology Compass, 6*, 347-360. DOI: 10.1111/j.1751-9004.2012.00430.x

Lin, L. M., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22*, 111-128. DOI: 10.1080/02702710119125

Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, *26*, 959-964. DOI: 10.3758/BF03201176

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review*, *1*, 126-129. DOI: 10.3758/BF03200769

Maki, R. H., & Serra, M. (1992). The basis of test predictions for text material. *Journal of experimental psychology: Learning, memory, and cognition, 18,* 116. DOI: 10.1037/0278-7393.18.1.116

Martin, A. J., Bostwick, K., Collie, R. J., & Tarbetsky, A. (2016). Implicit theories of intelligence. In V. Zeigler-Hill & T.K. Shackelford (Eds.) *Encyclopedia of personality and individual differences*. New York: Springer

Maslow, A. H. (1943). A theory of human motivation. *Psychological review*, *50*, 370.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*, 349. DOI: 10.1037//0096-3445.131.3.349

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

Middleton, W., Harris, P., & Surman, M. (1996). Give'em enough rope: Perceptions of health and safety risks in bungee jumpers. *Journal of Social and Clinical Psychology, 15,* 68–79.

Miller, T. M., & Geraci, L. (2011a). Unskilled but aware: reinterpreting overconfidence in low-performing students. *Journal of experimental psychology: learning, memory, and cognition*, *37,* 502. DOI: 10.1037/a0021802

Miller, T. M., & Geraci, L. (2011b). 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. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and cognition*, *29*, 131-140. DOI: 10.1016/j.concog.2014.08.008

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Educational*, 7-28.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, *1,* 159. DOI: 10.1007/s10409-006-9595-6

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, *31*(2), 199-218.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition*, *29*, 62-67. DOI: 10.3758/BF03195741

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International journal of educational research*, *31*, 459-470.

Wolters, C. A., & Baxter, G. P. (2000). 2. Assessing Metacognition and Self-Regulated Learning. In G. Schraw & J. C. Impara (Eds.) *Issues in the Measurement of Metacognition*, Buros Center for Testing, University of Nebraska, Lincon.

Putwain, D. W., & Sander, P. (2016). Does the confidence of first-year undergraduate students change over time according to achievement goal profile? *Studies in Higher Education*, *41*, 381-398.

Rawson, K. A., O'neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied*, *17,* 288. DOI: 10.1037/a0024749

Regan, P. C., Snyder, M., & Kassin, S. M. (1995). Unrealistic optimism: Self-enhancement or person positivity? *Personality and Social Psychology Bulletin, 21,* 1073–1082.

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

Saenz, G. D., Geraci, L., & Tirso, R. (2019) Improving metacognition: A comparison of interventions. Applied Cognitive Psychology, 2019, 1-12/ DOI: 10.1002/acp.3556

Schunk, D. H., & Zimmerman, B. J. (1994). *Self-regulation of learning and performance: Issues and educational applications*. New York, NY:Lawrence Erlbaum Associates, Inc.

Sedikides, C., & Gregg, A.P. (2003). Portraits of the self. In M. A. Hogg & J. Cooper (Eds.), *Sage handbook of social psychology* (pp. 110–138). London: Sage.

Seligman, M. E. P. (1991). Learned optimism. New York: Knopf.

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. (1993). Productivity loss in performance groups: A motivation analysis. *Psychological bulletin*, *113*, 67. DOI: 10.1037/0033-2909.113.1.67

Shepperd, J. A., Findley-Klein, C., Kwavnick, K. D., Walker, D., & Perez, S. (2000). Bracing for loss. *Journal of personality and social psychology*, *78*, 620.

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

Steadman, M. (1998) Using classroom assessment to change both learning and teaching. *New Directions for Teaching and Learning, 75*, 23–35.

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition*, *3,* 161-164. DOI: 10.1016/j.jarmac.2014.02.001

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, *103,* 193.

Taylor, K. M., & Shepperd, J. A. (1998). Bracing for the worst: Severity, testing, and feedback timing as moderators of the optimistic bias. *Personality and Social Psychology Bulletin*, *24,* 915-926. DOI: 10.1177/0146167298249001

Taylor, S. E., Wayment, H. A., & Collins, M. A. (1993). Positive illusions and affect regulation. In D. M. Wegner, & J. W. Pennebaker (Eds.), *Handbook of mental control* (pp. 325–343). Upper Saddle River, NJ: Prentice-Hall.

Tirso, R., Geraci, L., Saenz, G., D. (2019, in press) Examining Underconfidence Among High-performing Students: A Test of the False Consensus Hypothesis. *Journal of Applied Research in Memory and Cognition*

Ucbasaran, D., Westhead, P., & Wright, M. (2006). Habitual entrepreneurs experiencing failure: overconfidence and the motivation to try again. In *Entrepreneurship: Frameworks And Empirical Investigations From Forthcoming Leaders Of European Research,* 9-28. Emerald Group Publishing Limited.

United Stated Department of Education (2018; accessed February 2019). *College Scorecard Data* https://collegescorecard.ed.gov/data/

Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. *Handbook of self-regulation of learning and performance*, 15-32.

Weinstein, N. D., & Klein, W. M. (1995). Resistance of personal risk perceptions to debiasing interventions. *Health Psychology, 14*, 132–140.

APPENDIX A

LOGICAL REASONING EXAM SAMPLE QUESTIONS

Questions 1–2

A company employee generates a series of five-digit product codes in accordance with the following rules:

The codes use the digits 0, 1, 2, 3, and 4, and no others. Each digit occurs exactly once in any code. The second digit has a value exactly twice that of the first digit.

The value of the third digit is less than the value of the fifth digit.

1. If the last digit of an acceptable product code is 1, it must be true that the

    (A) first digit is 2
    (B) second digit is 0
    (C) third digit is 3
    (D) fourth digit is 4
    (E) fourth digit is 0

2. Which one of the following must be true about any acceptable product code?

    (A) The digit 1 appears in some position before the digit 2.
    (B) The digit 1 appears in some position before the digit 3.
    (C) The digit 2 appears in some position before the digit 3.
    (D) The digit 3 appears in some position before the digit 0.
    (E) The digit 4 appears in some position before the digit 3.

Question 3

3. Situation: Someone living in a cold climate buys a winter coat that is stylish but not warm in order to appear sophisticated. Analysis: People are sometimes willing to sacrifice sensual comfort or pleasure for the sake of appearances. The analysis provided for the situation above is most appropriate for which one of the following situations?

    (A) A person buys an automobile to commute to work even though public transportation is quick and reliable.
    (B) A parent buys a car seat for a young child because it is more colorful and more comfortable for the child than the other car seats on the market, though no safer.
    (C) A couple buys a particular wine even though their favorite wine is less expensive and better tasting because they think it will impress their dinner guests.
    (D) A person sets her thermostat at a low temperature during the winter because she is concerned about the environmental damage caused by using fossil fuels to heat her home.
    (E) An acrobat convinces the circus that employs him to purchase an expensive outfit for him so that he can wear it during his act to impress the audience.

APPENDIX B

SAMPLE PREDICTION SHEET

Prediction Sheet 1

Please answer all questions as honestly and accurately as possible. Please answer questions 1-4 with a percentage (NOT a letter) between 0% and 100%.

1. What grade do you think you will receive on this test? _____________%
2. What is your goal to earn on this test? ____________%
3. What do you think will be the average grade on this test? __________%
4. Compared to the rest of the participants, I think I will perform better than _________% of the participants. (e.g. If you consider yourself average, you should indicate 50%, if you think you are very above average, you may say 90%)

Please indicate how much you agree or disagree with the following statements.

5. I based my grade predictions on my academic background

Strongly Disagree 1 2 3 4 5 6 7 8 Strongly Agree

6. I based my predictions on my preparation or familiarity with this kind of exam

Strongly Disagree 1 2 3 4 5 6 7 8 Strongly Agree

7. I based my predictions on my prior performance on this type of exam

Strongly Disagree 1 2 3 4 5 6 7 8 Strongly Agree

8. I based my predictions on my prior academic performance

Strongly Disagree 1 2 3 4 5 6 7 8 Strongly Agree

9. I based my predictions on the lowest grade I would be happy with on this test

Strongly Disagree 1 2 3 4 5 6 7 8 Strongly Agree

10. I based my predictions on my ideal grade (considering my individual efforts for this test).

Strongly Disagree 1 2 3 4 5 6 7 8 Strongly Agree

11. I based my predictions on what I consider to be an “okay” or “acceptable” grade

Strongly Disagree 1 2 3 4 5 6 7 8 Strongly Agree

APPENDIX C

TABLES

Table 1

Descriptive statistics for experiment 1.

				Calibration				
	Prediction 1		Postdiction 1		Prediction 2		Postdiction 2	
Performance Quartiles	n	M(SD)	n	M(SD)	n	M(SD)	n	M(SD)
1	46	52.50(16.32)	47	26.00(24.10)	47	37.36(24.13)	47	31.19(19.74)
2	39	38.41(15.03)	39	18.33(19.41)	39	26.59(19.66)	38	20.34(18.81)
3	42	33.02(11.79)	42	10.50(16.97)	42	18.74(18.16)	41	11.24(18.90)
4	36	21.03(14.91)	36	4.31(15.86)	36	6.28(20.82)	36	2.44(18.27)
Average	163	37.16(18.43)	164	15.45(21.12)	164	23.21(23.65)	162	17.21(21.70)
				Absolute Calibration				
	Prediction 1		Postdiction 1		Prediction 2		Postdiction 2	
Performance Quartiles	n	M(SD)	n	M(SD)	n	M(SD)	n	M(SD)
1	46	52.50(16.32)	47	28.98(20.33)	47	38.77(21.75)	47	32.17(18.06)
2	39	38.41(15.03)	39	21.92(15.11)	39	29.05(15.68)	38	23.50(14.54)
3	42	33.02(11.79)	42	15.50(12.45)	42	22.07(13.80)	41	17.10(13.68)
4	36	23.25(11.00)	36	13.47(9.17)	36	18.78(10.55)	36	14.67(10.89)
Average	163	37.65(17.40)	164	20.45(16.29)	164	27.79(18.00)	162	22.43(16.21)

Table 2

Individual difference data for calibration across experiment 1 as M(SD).

		n	Average Calibration	Average Absolute Calibration	Prediction 1 Calibration	Postdiction 1 Calibration	Prediction 2 Calibration	Postdiction 2 Calibration
Age Groups	18	101	-11.093 (11.992)	28.168 (15.31)	36.119 (16.693)	14.158 (20.483)	23.663 (22.781)	16.396 (21.313)
	19	41	-11.735 (12.785)	30.661 (18.027)	42.550 (20.633)	20.781 (23.837)	24.707 (24.502)	21.300 (22.216)
	20	10	-12.721 (11.717)	28.150 (16.730)	39.200 (15.483)	18.500 (13.754)	28.300 (18.227)	21.500 (15.995)
	21	7	-9.082 (14.011)	27.571 (12.269)	25.000 (22.730)	13.000 (15.853)	17.143 (25.635)	10.714 (22.253)
	22	5	4.166 (19.451)	26.318 (16.058)	28.000 (25.150)	5.000 (17.678)	0.000 (35.355)	2.500 (32.275)
Gender	Male	49	-11.973 (14.595)	31.075 (15.674)	34.878 (20.579)	12.755 (22.617)	22.899 (27.371)	14.833 (23.180)
	Female	115	-10.303 (11.753)	27.700 (15.934)	38.140 (17.431)	16.591 (20.443)	23.339 (22.005)	18.211 (21.078)
Race	Caucasian	98	-10.511 (11.458)	27.236 (15.075)	35.979 (15.992)	15.776 (19.626)	23.663 (20.707)	17.975 (23.177)
	Hispanic	40	-12.754 (13.689)	32.464 (16.400)	41.325 (18.821)	18.800 (22.243)	26.000 (24.999)	17.975 (23.177)
	Other	26	-8.897 (15.192)	28.479 (17.684)	35.154 (18.430)	15.445 (21.119)	23.207 (23.650)	17.210 (21.704)
Student Generation	First-Generation	28	-10.263 (12.465)	27.852 (16.010)	39.929 (25.690)	14.893 (20.993)	24.429 (24.619)	13.286 (22.213)
	Non-First-Generation	135	-12.783 (12.612)	28.537 (15.781)	36.485 (16.626)	15.378 (21.194)	22.682 (23.400)	17.790 (21.491)

Table 3

Summarized results for experiment 2.

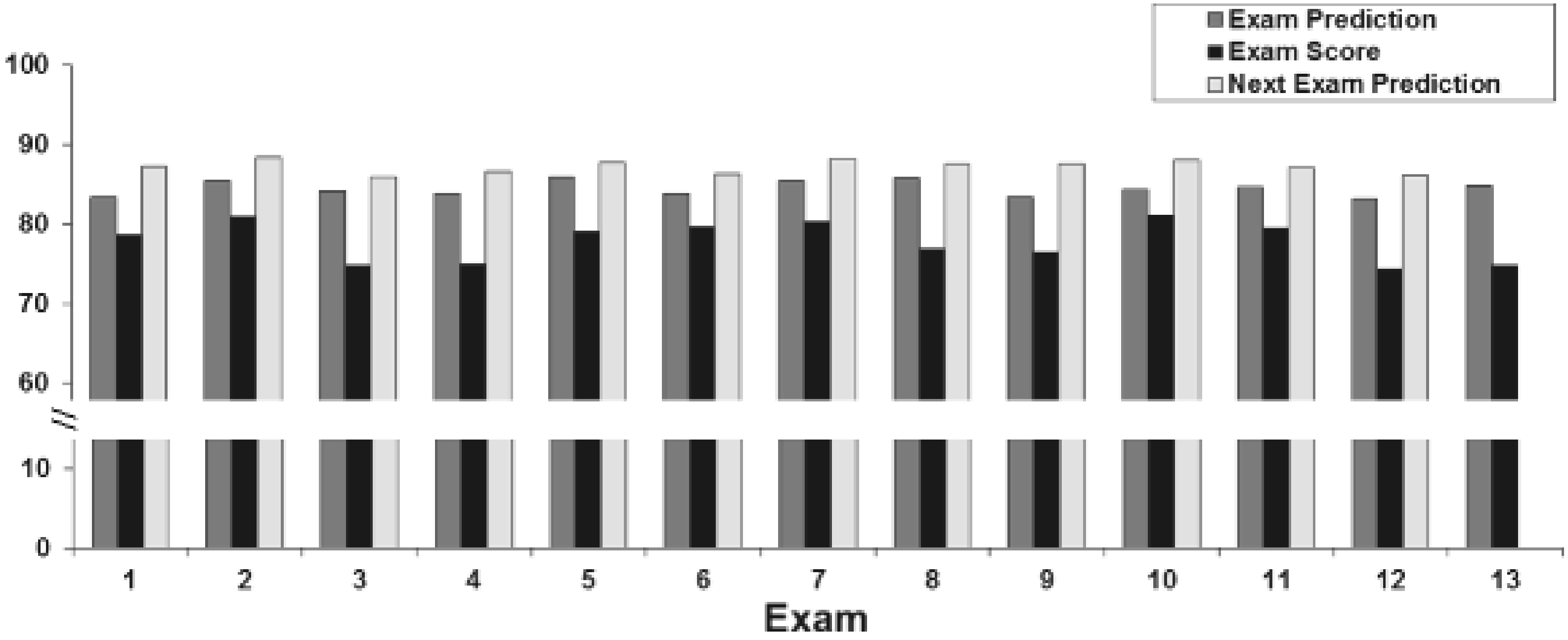
	Test 2		Reflection Period							Test 2	
	Pre-diction 1	Post-diction 1	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Pre-diction 2	Post-diction 2
N	116	117	71	48	48	46	45	41	56	89	89
Self-Evaluation	74.8 (11.693)	51.44 (18.33)	56.17 (19.904)	59.4 (15.789)	55.25 (20.52)	56.13 (18.464)	56.67 (17.419)	53.32 (19.701)	56.89 (19.906)	55.52 (18)	53.42 (19.024)
Calibration	36.18 (19.042)	12.85 (19.571)	12.54 (24.259)	16.55 (19.915)	11.79 (24.915)	11.3 (22.664)	11.79 (20.916)	9.29 (22.853)	12.28 (23.272)	13.48 (22.556)	11.18 (21.141)
Absolute Calibration	37.58 (16.084)	19.63 (12.677)	22.02 (15.976)	20.55 (15.634)	22.26 (15.999)	21.05 (13.79)	20 (13.028)	19.57 (14.722)	21.68 (14.692)	21.32 (15.258)	19.59 (13.609)

APPENDIX D

FIGURES

Figure 1

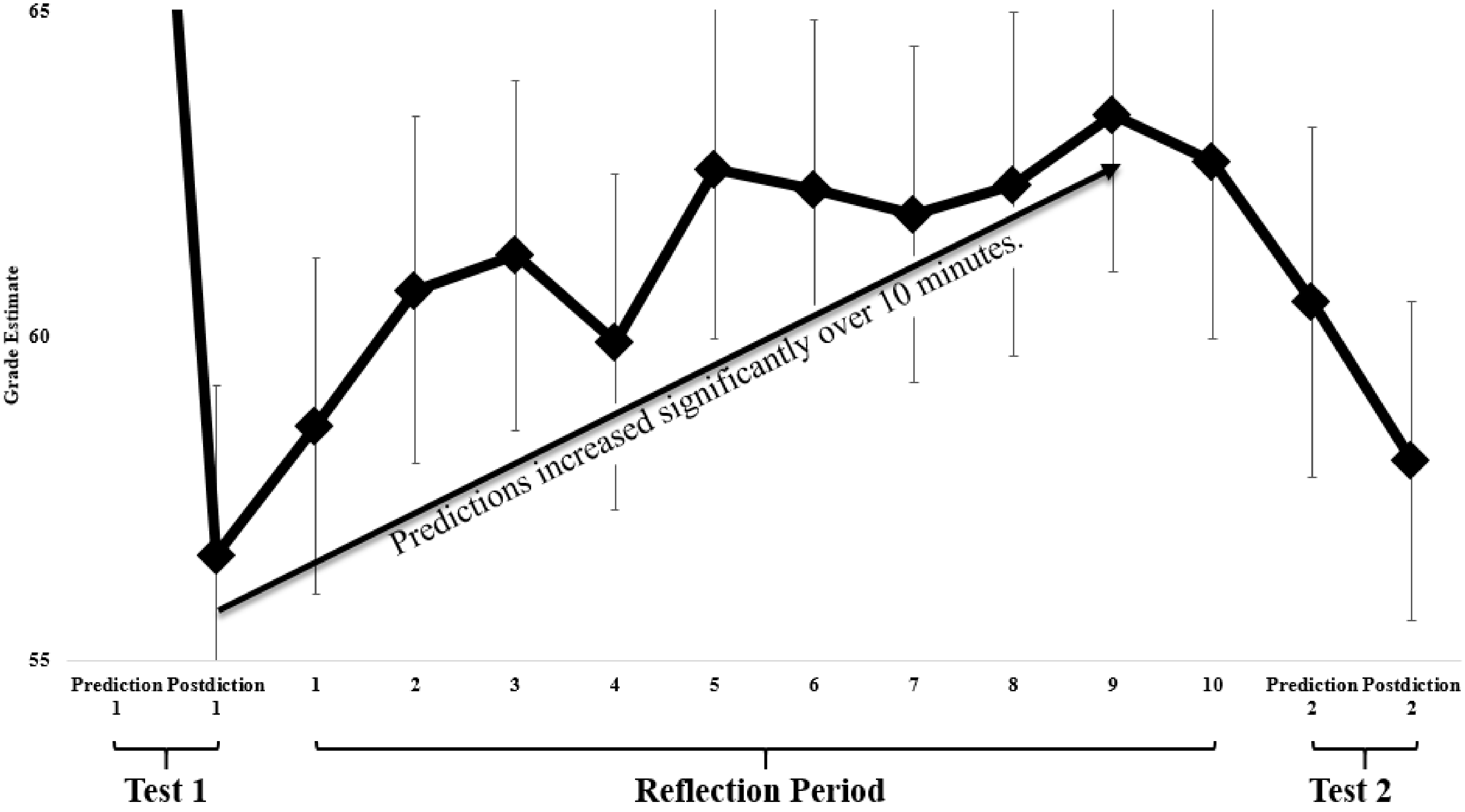
Average predictions and scores across thirteen exams over a university semester.



Note: Reprinted from “Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions”, by Foster et. al., 2017. Prediction accuracy did not improve over time.

Figure 2

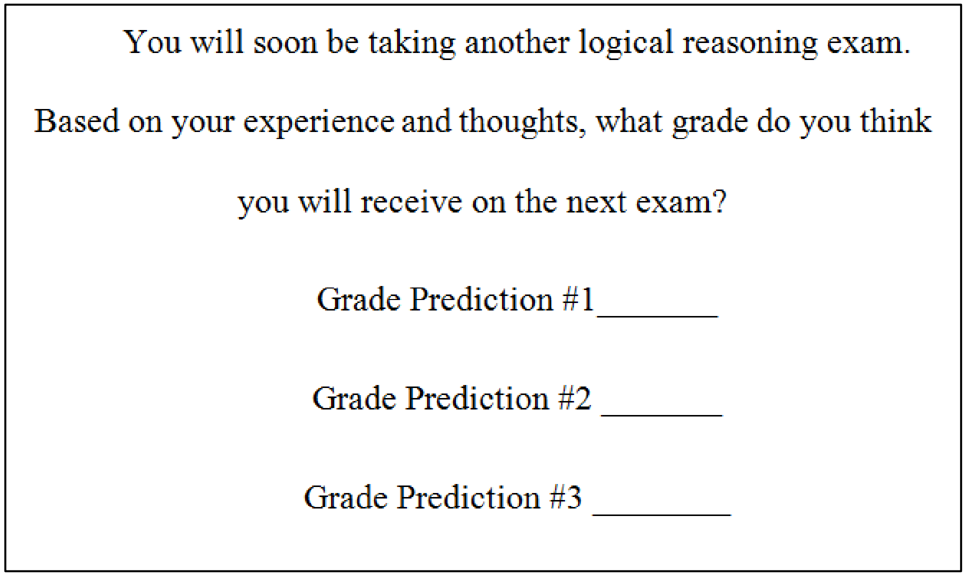
Preliminary data: repeated predictions intervention between two exams.



Note: These data were gathered as part of Saenz, Geraci, and Tirso (2019). Predictions rose significantly from the postdiction, to the tenth repeated prediction made about 10 minutes after the test (*t*(42) = -2.697, *p* = .010, *d* = -.23). Error bars represent Std. error of the mean.

Figure 3

Partial repeated predictions form used during the reflection period of experiment 1.



You will soon be taking another logical reasoning exam.

Based on your experience and thoughts, what grade do you think you will receive on the next exam?

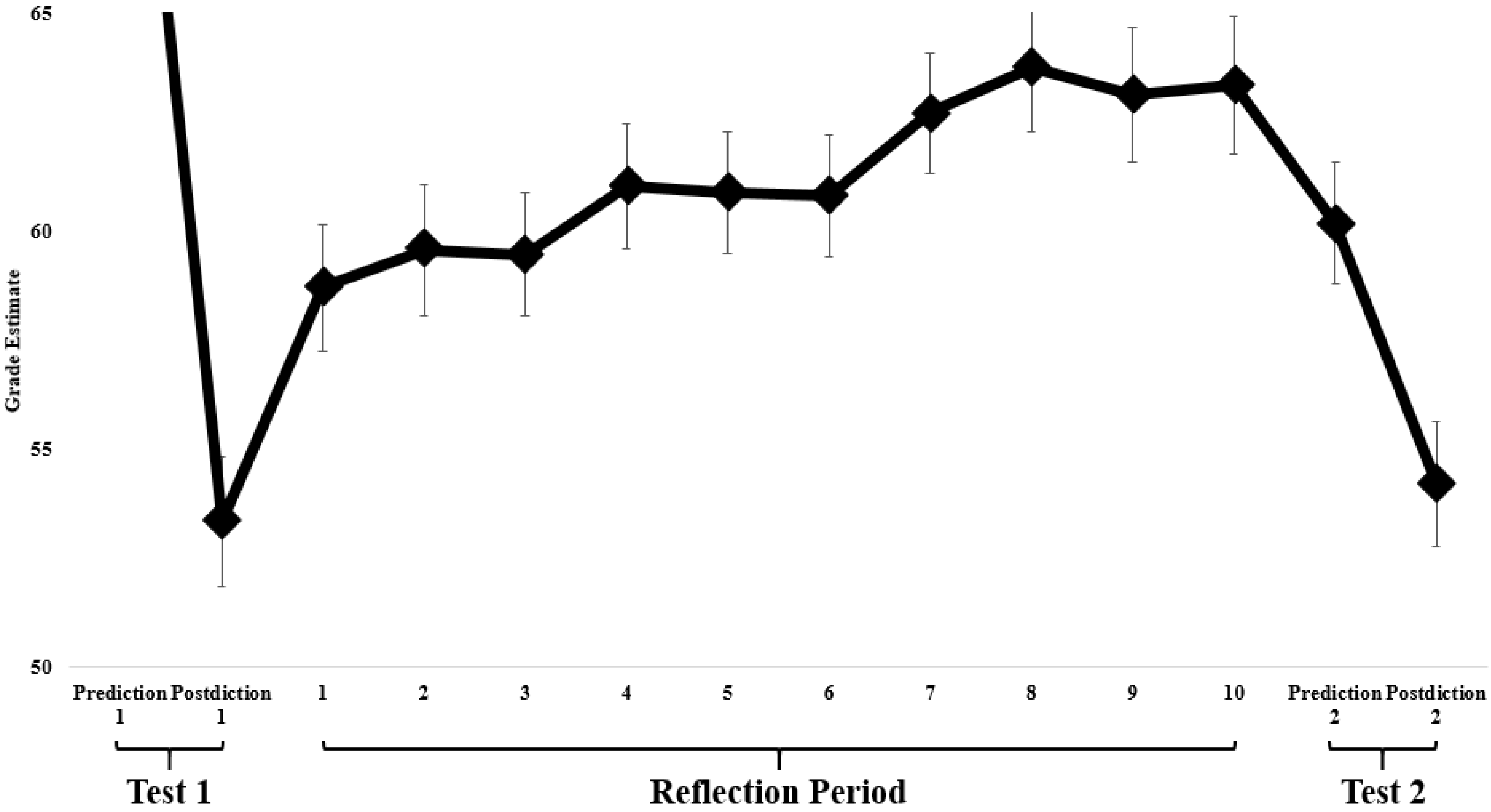
Grade Prediction #1_______

Grade Prediction #2 _______

Grade Prediction #3 ________

Figure 4

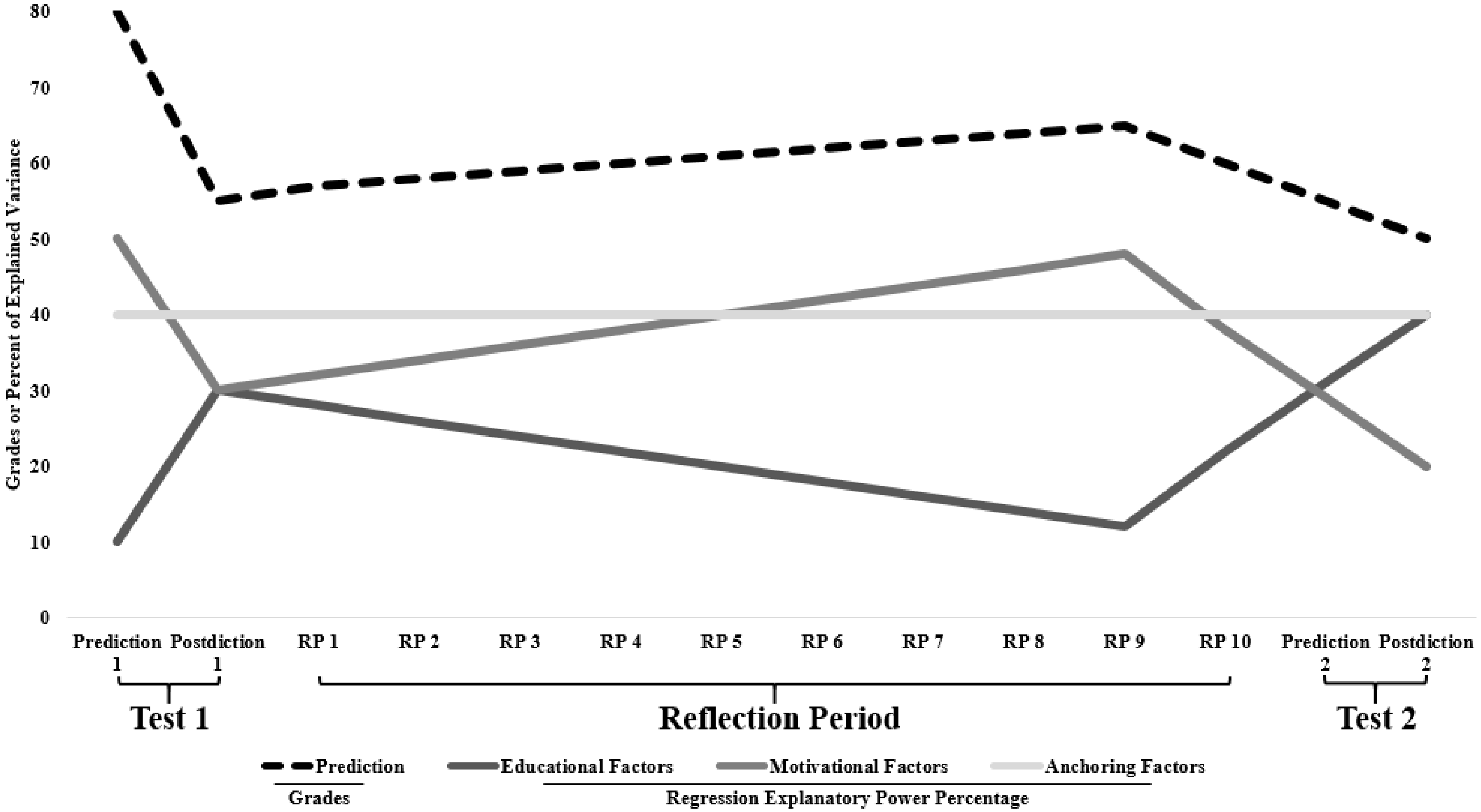
Performance predictions across experiment 1.



Note: Error bars represent Std. error of the mean.

Figure 5

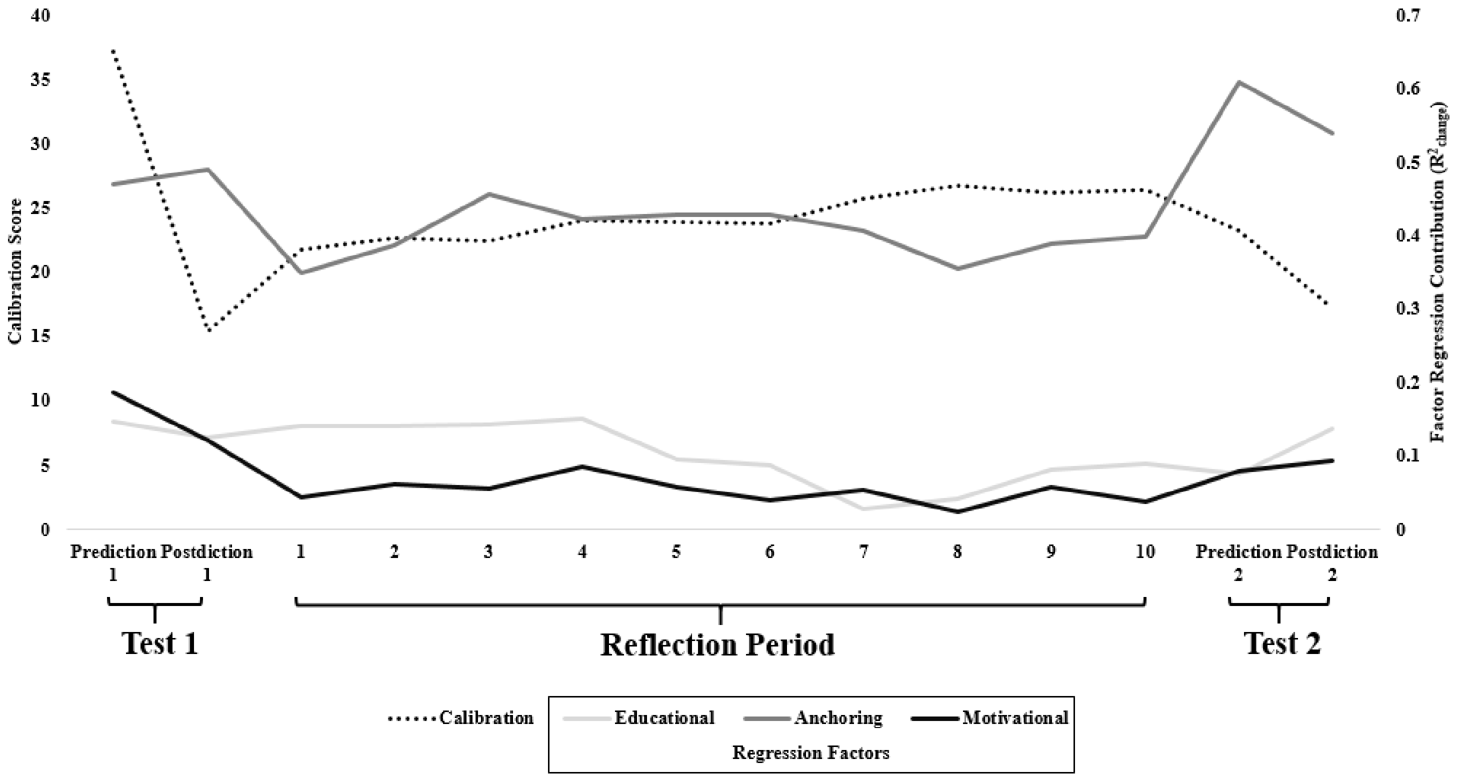
Experiment 1: Summary of expected results.



Summary of expected findings for Experiment 1-3 across tests and Reflection Period. Predictions are expressed as grade percentages. Educational, Motivational, and Anchoring Factors are expressed in terms of percentage of explained variance accounted for, when predicting grade predictions. These do not represent actual data, but are based on preliminary results.

Figure 6

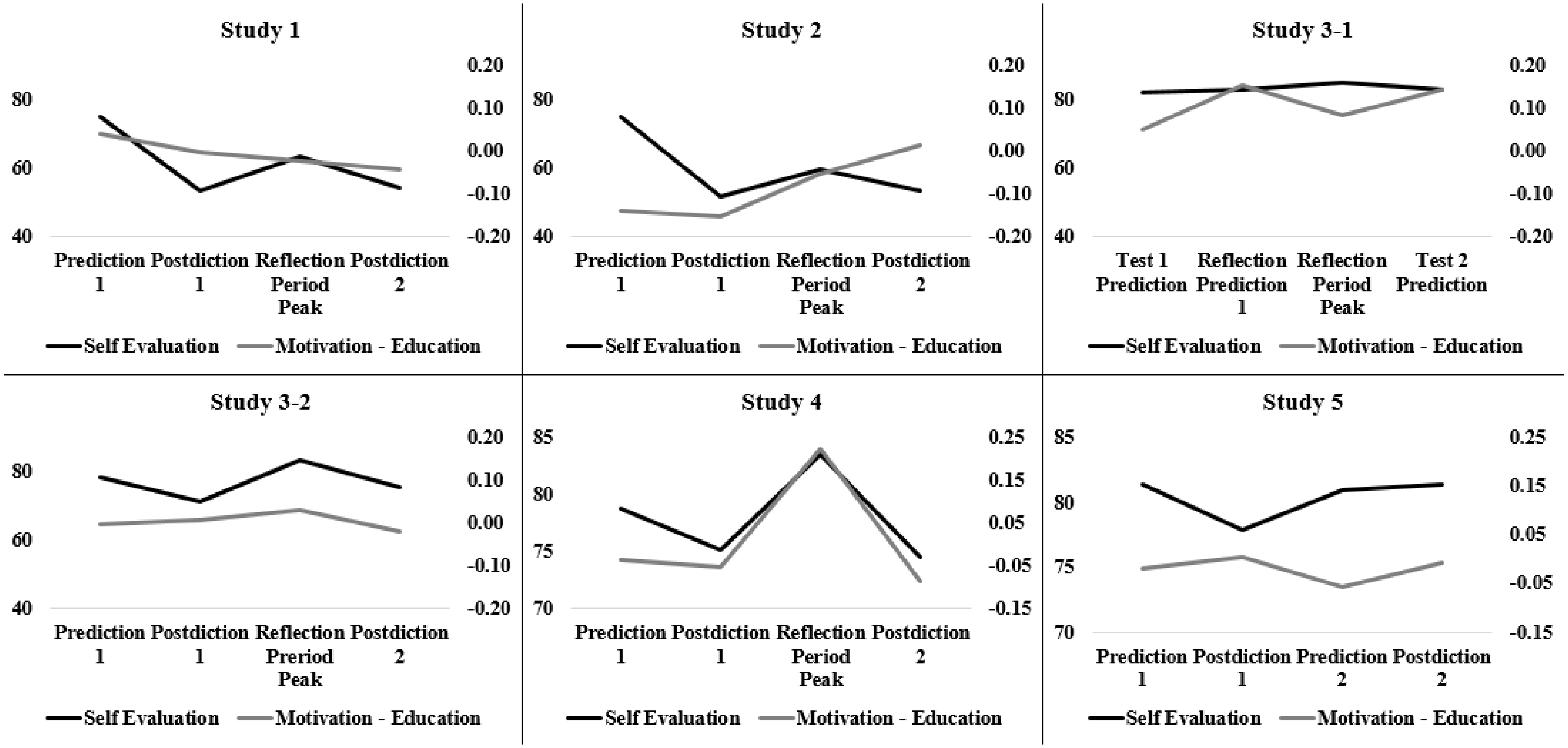
Calibration and explanatory factors across experiment 1.



Summary of results from Experiment 1. Calibration uses the left vertical axis and shows average performance prediction accuracy across Experiment 1, with perfect accuracy at 0, and more positive numbers indicating greater overconfidence. Educational, Anchoring, and Motivational Regression Factors use the right vertical axis, and show the explanatory power for each factor group (R^2_{change}) using each self estimation point (horizontal axis) as its dependent variable. Factors were only surveyed at Prediction 1/2 and Postdiction 1/2, so Reflection Period data points use factors surveyed from Postdiction 2.

Figure 7

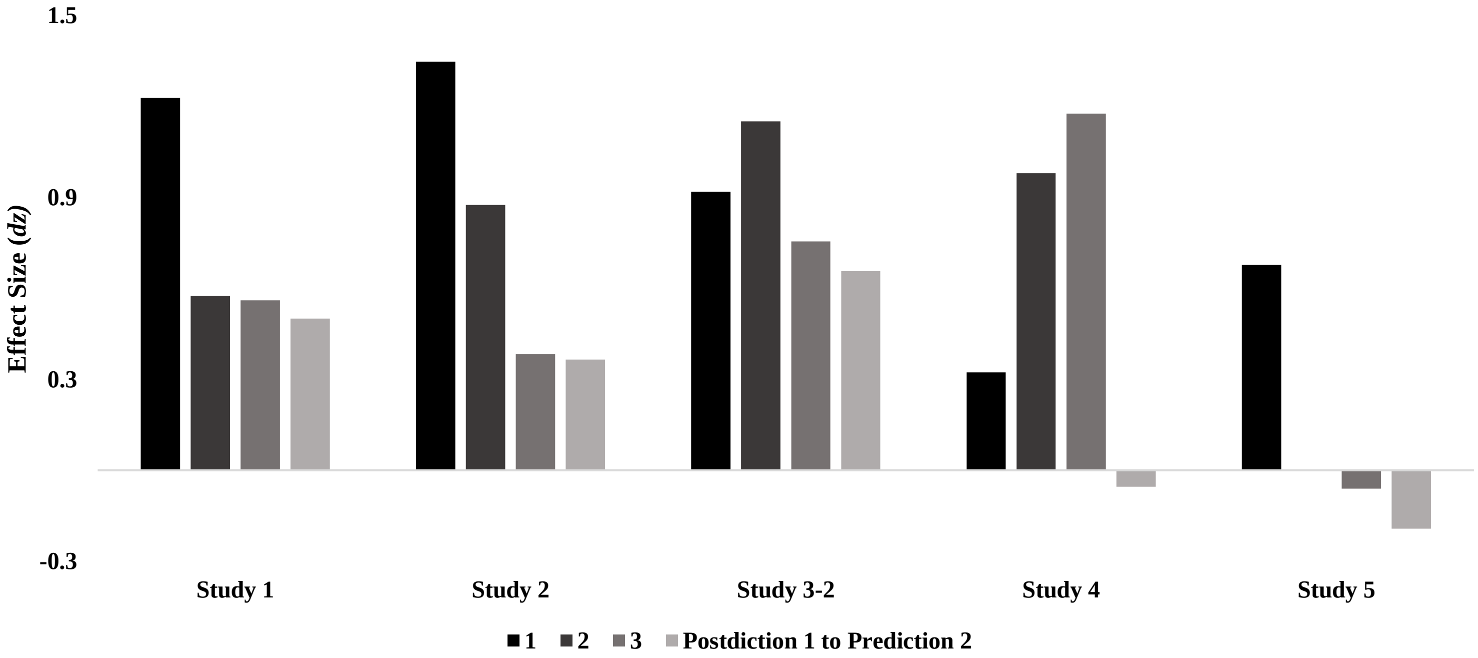
Motivational and educational factors' relative Association to self-evaluations.



Self-evaluations use the left vertical axis and show self-evaluative confidence at different points throughout Experiments 1-3. Motivation – Education uses the right vertical axis, and shows the explanatory power of the Motivational Factor group (R^2) minus that of the Educational Factor Group for explaining self-evaluations using regressions. Motivational and Educational Factors were regressed on self-evaluations in two separate analyses. More positive right-axis values indicate more motivationally based self-evaluations, while more negative numbers indicate more educationally based self-evaluations. Zero indicates equally motivationally and educationally based self-evaluations. Experiments 4-5 involved an intervention, between Test 1 and Test 2. No Reflection Period self-evaluations were gathered for Experiment 5.

Figure 8

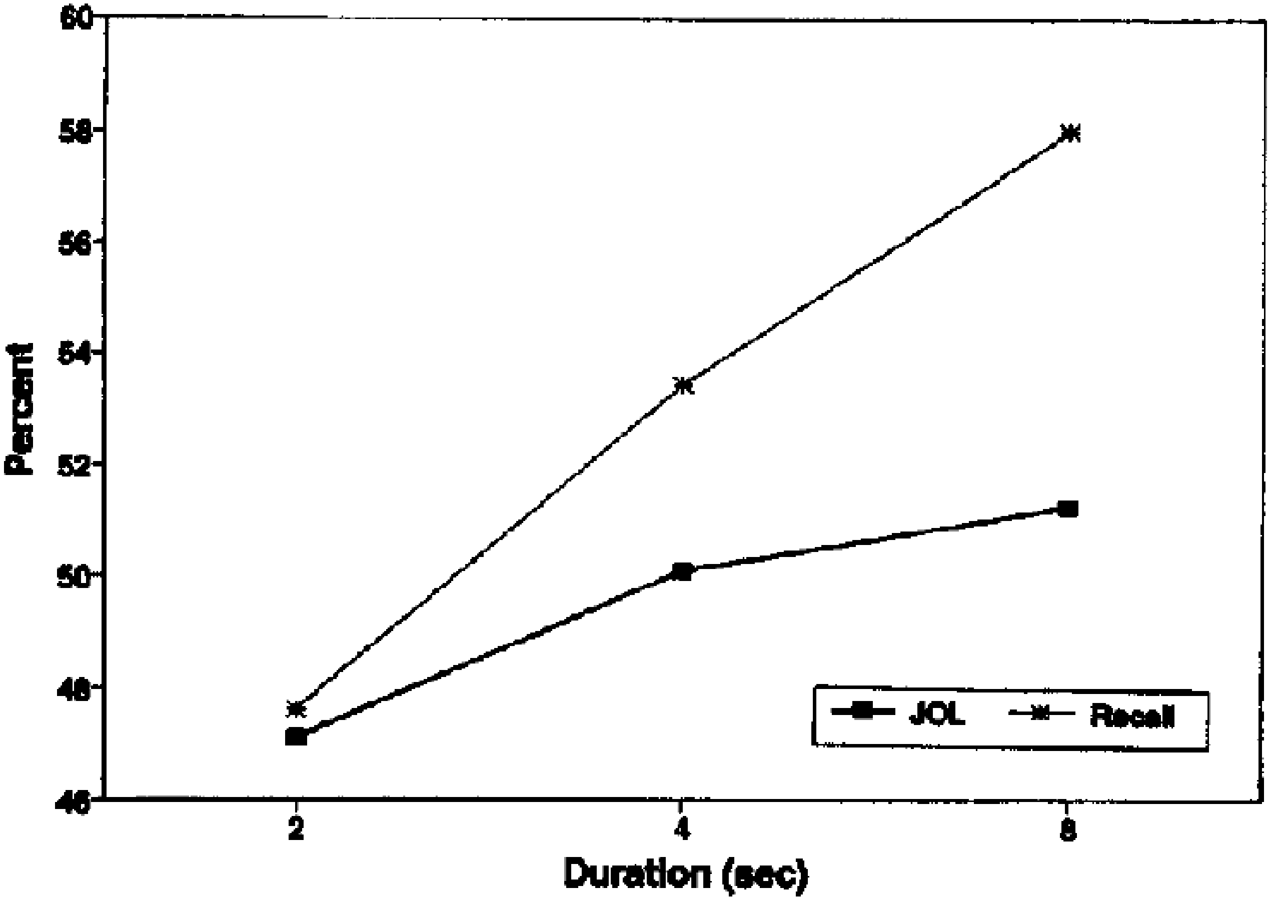
Effect size comparison for prediction changes across studies.



A comparison of effect sizes for the change in self-evaluative confidence across Experiments 1-3. Feature 1 compares participants' confidence before and after the first test (Prediction 1 to Postdiction 1). Feature 2 compares confidence from Postdiction 1 to the highest confidence point in the Reflection Period, this was the 9th judgement in Experiment 1, the 2nd day in Experiment 2, and the 1st week in Experiment 3.2. Data for Experiment 3.1 were omitted due to poor participant compliance. Feature 3 compared the highest Reflection Period Point to Postdiction 2. Experiment 1 lasted 10 minutes and had 10 Reflection Period Predictions, Experiment 2 lasted 1 week and had 7 Reflection Period Predictions, and Experiment 3.2 lasted 1 month and had 4 Reflection Period Predictions. Experiments 1-2 were laboratory studies, whereas Experiment 3.2 occurred in a university classroom. Experiments 4-5 involved an intervention to reduce the change in Postdiction 1 to Prediction 2. No data were collected for Experiment 5, Feature 2.

Figure 9

Self-evaluations may vary less than recall.



Reprinted from Figure 9 in Koriat, 1997, in this case, JOLs (Judgements of Learning) may be considered a type of self-evaluation.