DECONSTRUCTING THE UNSKILLED-AND-UNAWARE PROBLEM: EXAMINING THE EFFECT OF FEEDBACK ON MISESTIMATION WHILE DISENTANGLING COGNITIVE BIAS FROM STATISTICAL ARTIFACT

A Thesis
By
J. CHASE HOOD

Submitted to the Graduate School
at Appalachian State University
in partial fulfillment of the requirements for the degree of
MASTER OF ARTS

August 2015
Department of Psychology

DECONSTRUCTING THE UNSKILLED-AND-UNAWARE PROBLEM: EXAMINING THE EFFECT OF FEEDBACK ON MISESTIMATION WHILE DISENTANGLING COGNITIVE BIAS FROM STATISTICAL ARTIFACT

A Thesis
By
J. CHASE HOOD
August 2015

APPROVED BY:

_____
Rose Mary Webb, Ph.D.
Chairperson, Thesis Committee

_____
Andrew R. Smith, Ph.D.
Member, Thesis Committee

_____
Twila Wingrove, Ph.D.
Member, Thesis Committee

_____
James C. Denniston, Ph.D.
Chairperson, Department of Psychology

_____
Max C. Poole, Ph.D.
Dean, Cratis D. Williams School of Graduate Studies

Abstract

DECONSTRUCTING THE UNSKILLED-AND-UNAWARE PROBLEM:
EXAMINING THE EFFECT OF FEEDBACK ON MISESTIMATION
WHILE DISENTANGLING COGNITIVE BIAS FROM STATISTICAL ARTIFACT

J. Chase Hood
B.A., Furman University
M.A., Appalachian State University

Chairperson: Rose Mary Webb, Ph.D.

The Unskilled-and-Unaware Problem (UUP), sometimes referred to as the Dunning-Kruger effect for its original discoverers (Kruger & Dunning, 1999), describes the phenomenon where poor performers tend to overestimate their abilities while high performers underestimate theirs. Dunning (2011) argues that the misestimations of low and high performers result from two separate deficits in metacognitive ability. An alternate explanation provided by Krueger and Mueller (2002) suggests that the observed results are a product of the better-than-average (BTA) heuristic and the statistical artifact of regression to the mean. This study controlled for this statistical artifact by reducing measurement error. I also tested the metacognition explanation by examining the effects of different forms of feedback on misestimation. Results generally replicated previous research in that there was a strong negative relationship between performance and misestimation. Additionally, misestimations lessened somewhat over time, demonstrating a calibration effect that was greatest for participants with more extreme scores. However, calibration was not attributable to direct feedback, and calibration was seen even for participants who did not receive

feedback, suggesting that how misestimations change over time in combination with feedback may be more complicated than previously expected. Given the findings that direct feedback on performance had a minimal effect on calibration, results imply that individuals may need feedback more verisimilar to that of their everyday experiences to better utilize it to inform their self-assessments and reduce misestimation. This has implications for improving our interactions in instructional and other performance-based settings.

Keywords: Unskilled-and-Unaware Problem, Dunning-Kruger effect, regression to the mean, better-than-average heuristic, metacognition, feedback, ICAR

Acknowledgments

There is no doubting that without the continuous support of my committee that I would not have been able to see this project to fruition. I would like to thank my mentor and committee chair, Dr. Rose Mary Webb, for her constant encouragement, her companionship, her tutelage, and (perhaps most importantly) her near saint-like patience with me despite whatever road-blocks or missed deadlines came our way. It was entirely by her assistance that I was able to design, conduct, analyze, and disseminate a study of a scale and complexity that I would have never been able to realize on my own. Without her aid and advice in this and other areas, my time at App State would not have been a fraction as enlightening, enjoyable, or (dare I say) sardonic. I would like to thank Dr. Andrew Smith, a committee member, for his constant challenging of my ideas and assumptions. His questionings not only helped me to make the most use of the resources I had for my Thesis, but helped me to (in the words of Dr. Charles Brewer) "eschew [my own] intellectual flabbiness." I would also like to thank him for his open-door policy; a policy I made quite a practice of abusing. I also thank Dr. Twila Wingrove, a committee member, for her aid in the development of my understanding of structural equation modeling and related advanced statistics without which many of the current and most of the future analyses involved in this project would not be possible. Collectively, I would like to thank all of my committee members for consistently going above and beyond in their efforts to help my peers and myself grow.

Dedication

This work is dedicated to those who question everything; the world, others, and themselves. It is through your thoughtful in(tro)spection that improvements are made, growth is realized, and potentials are met.

Table of Contents

List of Tables

List of Figures

Deconstructing the Unskilled-and-Unaware Problem:

Examining the Effect of Feedback on Misestimation

While Disentangling Cognitive Bias from Statistical Artifact

J. Chase Hood

Appalachian State University

Abstract

The Unskilled-and-Unaware Problem (UUP), sometimes referred to as the Dunning-Kruger effect for its original discoverers (Kruger & Dunning, 1999), describes the phenomenon where poor performers tend to overestimate their abilities while high performers underestimate theirs. Dunning (2011) argues that the misestimations of low and high performers result from two separate deficits in metacognitive ability. An alternate explanation provided by Krueger and Mueller (2002) suggests that the observed results are a product of the better-than-average (BTA) heuristic and the statistical artifact of regression to the mean. This study controlled for this statistical artifact by reducing measurement error. I also tested the metacognition explanation by examining the effects of different forms of feedback on misestimation. Results generally replicated previous research in that there was a strong negative relationship between performance and misestimation. Additionally, misestimations lessened somewhat over time, demonstrating a calibration effect that was greatest for participants with more extreme scores. However, calibration was not attributable to direct feedback, and calibration was seen even for participants who did not receive feedback, suggesting that how misestimations change over time in combination with feedback may be more complicated than previously expected. Given the findings that direct feedback on performance had a minimal effect on calibration, results imply that individuals may need feedback more verisimilar to that of their everyday experiences to better utilize it to inform their self-assessments and reduce misestimation. This has implications for improving our interactions in instructional and other performance-based settings.

Introduction

Deconstructing the Unskilled-and-Unaware Problem:
Examining the Effect of Feedback on Misestimation
While Disentangling Cognitive Bias from Statistical Artifact

People can be poor judges of their own abilities and often 1) miscalibrate their performance on a task and 2) provide inaccurate estimates of their relative abilities or attributes compared to their peers (Alicke, 1985; Kruger, 1999; Kruger & Dunning, 1999). Kruger and Dunning (1999) noticed that this phenomenon followed a pattern; low scoring participants generally overestimated their performance and high scoring participants sometimes underestimated their performance. Furthermore, the further participants' scores were from the mean, the greater their misestimation was, with poor performers reliably showing greater misestimation overall. They coined this phenomenon the Dunning-Kruger effect, but a broader literature refers to it as the Unskilled-and-Unaware Problem (UUP; Hartwig & Dunlosky,; Krajč, & Ortmann, 2008; Ryvkin, Krajč, & Ortmann, 2012). Kruger and Dunning suggested that the reason low scoring individuals overestimated their abilities was due to the lack of metacognitive skills necessary to recognize their own incompetence. High performers, however, lack information in a different metacognitive area in that they overestimate the performance of others, leading them to fail to recognize exactly how well they actually did perform in relation to other participants.

In the following sections, I first provide a basic description of a well-known cognitive bias called the better-than-average (BTA) heuristic to provide a context for explaining the UUP. I then describe the archetypal design of the studies examining the UUP followed by the general pattern of results seen thus far. Following this, I describe the metacognitive explanation provided by Dunning, Kruger, and colleagues and focus on the research

concerning feedback that both supports and challenges this explanation. The next sections discuss an alternative explanation of the UUP and Kruger and Dunning's reply to this alternate account. Following this review of the literature, I state the rationale for this study and describe the procedure and study design. I then detail the analyses used to test the hypotheses provided as well as the results of these analyses. Lastly, I discuss the conclusions derived from these results, their implications, limitations of the study, and future directions.

Better-than-Average Heuristic

Considering the BTA heuristic is helpful for understanding the UUP. The BTA heuristic, described generally, is a tendency for individuals to estimate that they are above average in a positive domain, be it concerning a skill or trait or otherwise (Alicke, 1985; Guenther & Alicke, 2010; Krueger & Mueller, 2002). For example, previous research on the UUP has found that the majority of participants will estimate that their performance or ability is above average in areas as diverse as humor, grammar, logic (Kruger & Dunning, 1999), gun-safety knowledge (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008), and emotional intelligence (Sheldon, Dunning, & Ames, 2014; discussed more below). These examples do not exhaust the list of areas the BTA heuristic has been studied in, but it should be noted that the BTA heuristic is considered to be a robust and reliable effect (Guenther & Alicke, 2010).

How does the BTA heuristic relate to the UUP? While the bias resulting from the BTA heuristic is common and well-documented, it seems that, at least in some cases, a more complex pattern of results emerges that cannot be explained completely by a general trend to overestimate one's abilities. While people do tend to overestimate their abilities on average, it appears that sometimes, as will be discussed more below, a small portion of individuals,

those who score the highest on a task, actually underestimate their abilities on average (Ehrlinger et al., 2008; Kruger & Dunning, 1999). This suggests that everyone does not use the BTA heuristic universally. Furthermore, it appears that the overall magnitudes of misestimation of the high and low performers of a task are asymmetrical, suggesting that low performers are misestimating their performance more than high performers. These two observations suggest that a more complex theory may be needed to explain this pattern of results. The UUP has been offered to address these previous limitations.

The Typical UUP Design

While researchers have examined the UUP in many different scenarios using various tasks and procedures, the experiments described by Kruger and Dunning (1999) provide a prototypical design model that has largely persisted throughout the subsequent work. It allows for a concise but still complete description of the basic procedures used and the reasoning behind these procedures without the need to describe every experiment in detail. Almost every experiment examining the UUP consists of three fundamental aspects: an ability measure, the performance estimations made by participants, and an experimental manipulation.

Ability measure. Fundamental to every UUP study is an ability measure. The researchers studying this effect have utilized numerous types of tasks measuring different abilities from logic, grammar, and humor (Kruger & Dunning, 1999); to college exam performance, debate tournament performance, and gun safety knowledge (Ehrlinger et al., 2008); to emotional intelligence (Sheldon et al., 2014), among others. While the ability in question is not of particular concern, what is paramount is that in each study there is some objective ability criterion on which participants can be scored. This provides an objective

point of reference (referred to here as the *absolute score*) to which the participants' estimations of their performance can be compared. Additionally, the absolute scores of participants can then be compared to each other to calculate a participant's relative standing (referred to here as their *relative score*).

Performance estimations. The second crucial element to all of the studies conducted is the performance estimations provided by the participants after completion of the ability measure. In the majority of the studies on the UUP, three performance estimations are collected and analyzed: the absolute score estimation, the relative score estimation, and the relative ability estimation.[1] The absolute score estimation is simply the participant's estimation of how well he or she performed on the ability measure. This estimation is usually presented in the form of the number of questions on the ability measure answered correctly or the percentage of the items answered correctly. The absolute score estimation is directly compared to the absolute score in most studies. The amount of misestimation in an estimated absolute score is the discrepancy between the estimated absolute score and the absolute score on the ability measure itself.

The relative score estimation is the estimation of how well the participant thinks he or she performed the task compared with others who performed the task. In previous research, this estimation has been represented by percentile rank. Similar to above, the relative score estimation can be compared to the participant's true relative score. In this case, the amount of misestimation in the relative score estimation is the discrepancy between the estimated relative score and the actual relative score.

[1] The original researchers did not designate specific terms for the performance estimates. More recent researchers, however, have begun using these terms to facilitate communication and I have chosen to emulate them in this (Krajč, & Ortmann, 2008; Ryvkin, Krajč, & Ortmann, 2012)

The last estimation present in the previous research is that of relative ability. The term relative ability refers to the actual ability of the participant that the task is attempting to measure and how this ability compares to others. Unlike the relative score estimation, which asks how well the participant performed the task in question compared to others, the relative ability estimation asks the participant to estimate how his or her skill in that ability (e.g., logic itself) compares to others as a whole. As such, misestimation in relative ability cannot be measured as it is impossible to know exactly what the true ability of an individual is.

Experimental manipulations. While an exhaustive review of all the experimental manipulations used in the past to study the nuances of the UUP is beyond the scope of this thesis, a general description of how they are implemented is warranted. Generally the researchers establish that a particular ability measure is appropriate to elicit the general pattern of results (discussed below) that characterizes the UUP in an initial study. After this is done, they may add an experimental condition in a follow up study to examining the effect of this condition on the UUP. Some examples include the effect of a feedback condition (which will be utilized in this thesis; Kruger & Dunning, 1999), training (Kruger & Dunning, 1999), monetary incentives for estimation accuracy (Ehrlinger et al., 2008), and constructive feedback on how to improve one's performance (Sheldon et al., 2014). These experimental manipulations provide additional insight into the discrepancy between the magnitude and direction of misestimation of high and low performers.

General Pattern of Results in UUP Research

A strong pattern of results has emerged from the body of research on the UUP. While not every study replicates these results exactly, there is clearly a pattern that generalizes across the studies that can be summarized in four main points. The first is that the majority of

participants overestimate their performance and this overestimation often exceeds the sample's mean performance. This point reflects the BTA heuristic discussed above. The second is that low performers tend to overestimate their scores and high performers tend to underestimate their scores. The third is that, for both high and low performers, the further away from the mean their absolute score falls, the greater their misestimation is on average. The fourth is that the magnitude of the misestimation is greater for poor performers than it is for high performers. These four points can be easily seen in Figure 1, reproduced from Dunning and Kruger (1999).

Metacognitive Explanation for the UUP

As an explanation for the UUP, Dunning (2011) argues that the misestimations of the scores of bottom and top performers are caused by a lack of metacognitive ability; that is, knowing about knowing. This lack of relevant information, however, stems from two separate sources of knowledge for the two groups: The bottom performers lack knowledge about their own incompetence whereas the top performers lack knowledge about the incompetence of others. The bottom performers, Dunning argues, suffer what he calls a "double burden." The first burden is that their general lack of ability in the domain being tested leads them to make numerous mistakes. The second burden is that the same factors that cause these individuals to perform poorly make it extremely difficult for them to identify their mistakes even upon being informed that they are incorrect. This is because the ability to identify a mistake, according to Dunning, requires the ability to differentiate the mistaken answer from the correct answer. In other words, to identify a mistake, one must possess the requisite skills needed to answer the question correctly in the first place. High performers, however, suffer from a different, and arguably lesser, cognitive bias. The high performers

mistakenly assume that, because the test was relatively easy for them, that it was most likely relatively easy for their peers (Dunning, 2011; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger et al., 2008; Kruger & Dunning, 1999). This bias is typically referred to as the false-consensus effect (Ross, Greene, & House, 1977).

The varied effect of feedback on high and low performers' misestimation as support for the metacognition explanation. Dunning (2011) argues that high and low performers' misestimations are caused by a lack of knowledge stemming from two different sources. The low performers lack accurate knowledge of their absolute score. They consistently overestimate their absolute score because they lack the requisite ability to identify when they have made a mistake. This overestimation results from poor performers assuming that, even when they are unsure of what the correct answer or choice in a situation is, their guess is a reasonable one or is likely correct. Poor performers then go on to use this incorrect assessment of their performance to estimate how well they are doing relative to others. By doing so, they tend to also overestimate their relative scores, a pattern that has been demonstrated several times (Ehrlinger et al., 2008; Kruger & Dunning, 1999; Sheldon et al., 2014). High performers, however, are consistently accurate in judging their absolute score, at least compared to low performers. This is because they do not lack the requisite skill to identify their own mistakes. High performers lack only accurate knowledge of how others performed, making them unable to judge their own performance relative to others. High performers mistakenly overestimate the average performance and ability of others and, consequently, underestimate their relative score and ability.

Some preliminary evidence for these two hypotheses comes from the results of Kruger and Dunning (1999)'s third experiment. In this experiment, participants were asked to

complete a grammar test and then to estimate their absolute score, their relative score, and their relative ability. A few weeks later, participants in the top and bottom quartiles, based on their previous performance, were asked to return to the laboratory under the cover story that they were to grade the tests of five other participants. The tests that they graded reflected the range of scores, the mean, and the variability of the actual scores of the full sample, a fact that was explained to the participants. After grading, participants were shown their test again and asked to re-estimate their absolute and relative scores along with their relative ability. The effect of this additional information on their re-estimations differed for high and low performers.

High performers, who had initially underestimated their relative score and ability, re-estimated their relative score and ability much more accurately after receiving feedback, providing evidence that they suffered from a false-consensus bias. They did not change their absolute score estimations, which were already fairly accurate, but this was not expected to happen as the false-consensus bias should only concern one's relative standing. Poor performers, however, failed to more accurately re-estimate their performance or ability. Additionally, they were significantly less accurate in grading the tests than the high performers, providing evidence that their lack of ability to correctly answer the questions prevented them from recognizing both their own and others' errors and prevented them from using the new information to make accurate relative score and ability estimates.

Ehrlinger et al. (2008) documented a similar trend in an analysis of four earlier studies examining the UUP. Using counterfactual regression analyses, the authors attempted to estimate what the misestimation of high and low performers would have been had they been given feedback on their absolute or relative score. It is important to note that the

feedback being discussed here is qualitatively different than that given by Kruger and Dunning (1999) who allowed their participants to glean feedback by examining the errors of others. The feedback proposed in Ehrlinger et al. refers to accuracy-based feedback that would be given directly to the participants, not feedback based on infereces that participants would have to construct themselves. However, it is critical to note that this feedback was not actually given; instead, the analysis simulated what the researchers expected to happen if the feedback had been given.

From this analysis the authors concluded that for poor performers, who greatly overestimated their absolute and relative scores on average, receiving feedback about their absolute score (i.e., knowing how many questions they had answered correctly) would have allowed them to better estimate their relative score (i.e., how well they did compared to other participants). Receiving feedback about the relative scores of others, however, may have led poor performers to further overestimate their own performance. This is because poor performers overestimated the relative scores of other participants (i.e., poor performers estimated that the other participants did better on average than they actually did). In light of this, the poor performers' relative score estimates were artificially attenuated by their overestimation of the performance of other participants on average. Had this overestimation of the performance of others been corrected by actual feedback on their relative score, poor performers may have further overestimated their own relative score. The effect of feedback on misestimation is thought to be different for high performers, however. The authors propose that feedback concerning either high performers' absolute or relative scores would have led to a reduction in misestimation overall because high performers both underestimated their own absolute score and overestimated the absolute scores of the other participants.

These interpretations have not been universally supported, however. Ryvkin et al. (2012) conducted two studies to assess how feedback affected the UUP. In their first study, the researchers examined how feedback affected estimations of future exam performance. They asked students to predict their absolute and relative scores on a midterm and a final exam. Students' predictions of their midterm performance were collected at the beginning of the semester when students had little knowledge about the course and again minutes before taking the midterm. Students were not provided with any kind of formal feedback (e.g., exam grades) between their first and second midterm predictions, but were provided with opportunities for *environmental* feedback over the intervening weeks consisting of their impressions of the course material and its overall difficulty. The students then made a third prediction about their absolute and relative score on their final exam minutes before taking the exam. Feedback differed between the second midterm prediction and the final prediction in that students were provided with direct feedback (i.e., grade on their midterm and relative standing in the class) after the midterm along with continuing environmental feedback.

Previous work on the UUP would suggest that the unskilled in the class, while they might benefit some from direct feedback, would not benefit from the environmental feedback present between the first and second midterm predictions in this study. This is because it is thought that the poorest performing individuals are incapable of recognizing their own incompetence and therefore fail to properly incorporate the feedback they receive in their predictions of absolute and relative performance, especially when they must construct this feedback for themselves from their experience in the course. However, this did not seem to be the case. The environmental feedback that the participants received between the first and second midterm predictions seemed to not only result in better calibration in predicted exam

performance overall, but had the greatest effect on the poorest performers, possibly because these students had the most room to improve. This trend continued after the midterm for the final exam prediction when participants received both environmental and direct feedback; see Figure 2, reproduced from Ryvkin et al. (2012). The results of this study showed that feedback, especially direct feedback, can largely mitigate the UUP.

To exert more control and to address some of the limitation of their first study, the researchers conducted an experiment using the same participants from Study 1. In this experiment, participants were asked to complete two tasks, each with two stages. The first task was a simple math exercise where participants had to sum as many 2-digit numbers as they could in a 3-minute time limit both in Stage 1 and Stage 2. The second task asked participants to compare the populations of 20 random pairs of European countries in Stage 1 and the populations of the of 40 out of the 50 most populated countries at Stage 2. After each stage for each task, participants estimated their absolute and relative scores in a manner more similar to previous research than the predictions used in Study 1. At Stage 2 of each task, half of the participants received feedback about their absolute and relative performance from Stage 1 before making their absolute and relative score estimations, while the other half received no feedback.

For the first task, the misestimations of high and low performers at Stage 1 mirrored that of previous research, as can be seen in Figure 3, reproduced from Ryvkin et al. (2012). Low performers were overconfident in their absolute and relative scores, while high performers accurately estimate their absolute score and slightly underestimate their relative score. This was not true of the second task, where the previous pattern of results was only replicated for the relative score estimates. Of most importance is the effect of feedback on

performance estimates. This experiment replicates the findings of Study 1 by demonstrating that feedback about one's absolute and relative score can reduce misestimation, especially for poor performers, and it does so using a design comparable to previous research on the UUP. Participants who did not receive feedback showed no improvement in the calibration of their estimations. This provides some evidence for the utility of environmental feedback demonstrated in Study 1.

This experiment unfortunately suffered from a few important limitations. The first was the difficulty level of the two tasks used. Performance on the first task demonstrated differences in ability among participants, but this difference was apparent only in the speed of completion (i.e., the number of problems solved) and not in the overall level of accuracy. While speed of completion of math problems is a valid assessment of ability, it is not necessarily comparable to other ability measures in regard to estimating one's absolute score. Task 2 suffered from a pronounced ceiling effect, reducing its utility to examine the relationship between performance and estimations.

Additionally, this experiment does not allow for the comparison of different forms of feedback. In their counterfactual regression analysis, Ehrlinger et al. (2008) proposed that receiving feedback about one's absolute and relative scores will have different effects on high and low performers, but that may only be examined by separating the two types of feedback. Finally, Ryvkin et al. (2012) noted that the restriction of range in ability of their population, along with the fact that students are intrinsically motivated to increase their academic achievement, might have affected their results. The current study addressed these three limitations, as discussed below.

An Alternate Explanation for the UUP: Regression to the Mean and the Better-Than-Average Heuristic

An alternative to the metacognition explanation given for the pattern of results observed thus far argues that, coupled with the BTA heuristic, regression to the mean explains the misestimations observed and that, when controlled for, the asymmetry in the misestimations disappears (Krueger & Mueller, 2002). The rationale behind this explanation is based on the observation that research on the UUP has frequently utilized ability measures with low levels of reliability and validity (i.e., instruments with much measurement error) and extreme group selection (usually top and bottom quartiles). When reliabilities are low and extreme performers are focused on, misestimations may be exaggerated. When scores on a measure are normally distributed, the likelihood that participants' observed scores are closer to the mean than their true score (i.e., what their score would be if no random measurement error existed) is equal to the likelihood that their observed scores are further from the mean than their true scores, even with measures with low reliability. However, when extreme scores are examined in isolation from the full distribution (as has been the case in most of the UUP literature), this introduces a systematic inclusion bias, such that some individuals with more extreme observed scores relative to their more moderate true scores (a difference attributable to the random error of the measurement) are more likely to be included in each extreme group, but individuals with more moderate observed scores relative to their true scores are less likely to be included in the extreme groups. The extreme group selection process may then result in exaggerated misestimations between participants' observed scores and their estimated scores, relative to the differences between their true scores and their estimated scores. This artifact of measurement error, coupled with the well-

documented BTA effect, may account for the general pattern of UUP results demonstrating overestimation for poor performers and underestimation for high performers.

However, it is not clear from any single set of results which explanation is preferred, especially because, up until this point, they have only been compared independently. Therefore, it seemed prudent to design a study specifically intended to measure the effect of regression to the mean on the accuracy of estimations of ability. Dunning, Kruger, and colleagues acknowledge that a regression to the mean effect may be present in the misestimations of performance, though they minimize its potential importance. Those who argue that the misestimation of ability is primarily the result of a lack of metacognitive abilities claim that the regression to the mean effect is negligible as long as reasonably reliable measures of ability are used (Dunning, 2011; Dunning et al., 2003; Ehrlinger et al., 2008; Kruger & Dunning, 1999).

One clear way to measure and control for the effect of regression to the mean is to employ a repeated-measure design using multiple alternate forms of a measure. While it is true that the regression to the mean effect is caused in large part by the inherent unreliability of a test, and while it is true that alternate forms of a test can never be perfectly equivalent in either their ability to measure a construct or their reliabilities, the regression to the mean effect should still be distributed relatively equally across participants if all participants take all forms of the test, provided enough versions of the test are used. Moreover, when the scores across all tests are averaged, the regression to the mean effect should be reduced. In short, participants who score in the extremes on one form of the test (i.e., those who have been the focus of previous studies) should be more likely to score closer to the mean on the alternative forms of the test, even though their absolute skill in the area being tested is not

changing. To what extent this occurs is of particular interest in resolving the conflict between the alternate explanations described above.

Rationale of the Study and General Description of Procedures

The goals of this experiment were manifold. The first goal was to replicate the pattern of results that characterize the UUP in a large, representative sample of adults rather than the typical small student sample. While adult non-students have been used in a smaller scale study before (e.g., Ehrlinger et al., 2008, Experiment 3), the risk for restriction of range of the ability being measured was still present, as the sample comprised adults at a skeet tournament and the ability measure used was a test of gun safety knowledge. Using a more representative sample allowed me to address some limitations that the restriction of range of cognitive abilities found in college populations might have had on previous results. This is especially pertinent considering I used a cognitive measure, the International Cognitive Ability Resource (ICAR; Condon & Revelle, 2014), as my ability measure.

The second goal of this experiment was to more acutely examine the effect of regression to the mean on misestimations. While some previous studies have addressed whether or not the UUP was the result of this statistical artifact, their conclusions were limited. Because no measure is perfectly accurate, and thus all measures are affected by regression to the mean to some extent, I am instead concerned with measuring how much this artifact affects misestimation. To address this, I employed a repeated-measure design to measure the extent to which each individual's score regresses to the mean over time. By doing so, I hope to accurately disentangle the effect of this artifact from the UUP.

The third and fourth goals of this experiment were interrelated. Research on the effect of feedback on misestimation by Ryvkin et al. (2012) contradicts predictions made by

previous researchers of the UUP (Ehrlinger et al., 2008; Kruger & Dunning, 1999). However, this latter research suffered from a few limitations, namely, the lack of control and the limitations of the measures used. This experiment sought to simultaneously address the limitations of the work of Ryvkin et al. (2012) while addressing the inconsistency of their results with the predictions made by Kruger and Dunning (1999) and Ehrlinger et al. (2008). I did this by employing a more standardized and better validated ability measure with known psychometric qualities and comparing the misestimations of four feedback conditions over four trials. Participants in this study completed the four subtests and after each one provided estimations of their performance. Some participants received one or both of two different kinds of feedback and the effect of this feedback over time was compared across groups.

Hypotheses. My hypotheses are best considered in two sections: those for which most previous researchers (Ehrlinger et al., 2008; Kruger & Dunning, 1999; Ryvkin et al., 2012) agree and those for which these researchers disagree (See Tables 1 and 2 for a visual representation of the hypotheses and each associated statistic).

Converging hypotheses.

Hypothesis 1a. Participants will overestimate their absolute and relative performance on average.

Hypothesis 1b. The magnitude of this overestimation will be negatively related to participants' performance overall, replicating the UUP.

Hypothesis 2a. Poor performing participants will overestimate their absolute and relative scores.

Hypothesis 2b. High performing individuals will underestimate their relative scores but not their absolute scores (or if they do so, only very slightly).

Hypothesis 3. Poor performers will exhibit more absolute and relative overestimation than high performers.

Unexplored feedback hypothesis: Current Researcher.

Hypothesis 4.0. Feedback condition will have no effect on participants' actual scores.

Competing feedback hypotheses: Ehrlinger et al. (2008) and Kruger and Dunning (1999).

Hypothesis 4.1. High, but not low, performers will be able to use environmental (i.e., indirect) feedback to better calibrate their absolute and relative performance estimates over time.

Hypothesis 5.1a. Direct feedback on absolute performance will improve both absolute and relative estimations for high and low performers.

Hypothesis 5.1b. Direct feedback on relative performance will improve relative estimations for high performers but exacerbate relative misestimations for poor performers.

Competing feedback hypotheses: Ryvkin et al. (2012).

Hypothesis 4.2a. The absolute and relative estimations of both high and low performers may benefit from indirect, environmental feedback over time (i.e., trials).

Hypothesis 4.2b. If absolute and relative estimations benefit from environmental feedback over time, the greatest improvement in calibration will occur for poor performers.

Hypothesis 5.2. (No clear differentiation of the effect of absolute and relative feedback is clear from this study.)

Competing regression to the mean hypothesis: Ehrlinger et al. (2008) and Kruger and Dunning (1999).

Hypothesis 6.1. The extent to which participants' absolute scores will regress to the mean across trials will be negligible.

Competing regression to the mean hypothesis: Krueger and Mueller (2002).

Hypothesis 6.2a. Participants' absolute scores will tend to regress to the mean across trials.

Hypothesis 6.2b. Participants with the most extreme absolute scores will show the greatest regression effect.

Method

Participants

A statistical power analysis was conducted in order to determine how many participants would be needed to provide sufficient power to detect differences between the groups in this study, should differences actually exist (Faul, Erdfelder, Lang, & Buchner, 2007). Because previous research has not reported effect sizes for the UUP, a very conservative effect size was used to ensure that my sample would be large enough to detect small effects. This power analysis indicated that a sample of at least 308 participants would be required to detect small differences between groups ($d = 0.20$, $\alpha = .05$, $\beta = .10$) for my core statistical tests. Therefore, to allow for potential losses associated with poor data quality and the additional planned analyses, I decided to recruit approximately 450 participants through Amazon Mechanical Turk (MTurk). Research by Holden, Dennie, and Hicks (2013) has shown that the quality of data provided by MTurk users is comparable to that of traditional research subject pools (e.g., undergraduates). The only restrictions that were placed on participant recruitment were that participants must be at least 18 years of age or older at the time of the study and from the United States. Participants voluntarily agreed to

complete the measure through MTurk using Qualtrics survey software. Participants were informed that their participation was voluntary and that they could choose to end their participation at any time. Participants were paid $1 upon completion of the task. I received approval from the Institutional Review Board (IRB) at Appalachian State University on 12/08/2014 to conduct this experiment (see Appendix A).

Two listwise deletion criteria were used in this study in an attempt to remove participants who were likely not taking the study seriously. The first was the removal of participants with incomplete data. The second was the removal of participants who had a score of 0 of 10 on two or more of the subtests. Of the 483 participants with at least partial data, 38 failed one or both of these exclusion criteria and were thus dropped, leaving 445 participants in my sample (234 female, 210 male, 1 unanswered) ranging from 18 to 77 years of age (M = 36.06, SD = 12.41, median = 33 years). Of the participants in my sample, 15.1% had a high school diploma or less, 35.3% had some college education, 37.3% had a Bachelors degree, and 12.1% had a graduate degree (one participant chose not to answer).

Design

The design of this experiment was a 4x4 mixed design with four between-group feedback conditions and four within-group ten-item ICAR subtests. The four feedback conditions were the no feedback condition, the absolute feedback condition, the relative feedback condition, and the combined feedback condition (described more below).

Materials

Test and subtest development. Four subtests consisting of 10 multiple-choice items each were developed using items from the 60-item version of the ICAR (Condon & Revelle,

2014).[2] The ICAR is a newly developed public-domain measure of cognitive ability. The items in the ICAR are grouped into four categories: 9 Letter and Number Series items, 11 Matrix Reasoning items, 16 Verbal Reasoning items, and 24 Three-Dimensional Rotation items. Of the 24 Three-Dimensional Rotation items, 12 were discarded in formation of my measure due to the relatively high difficulty of these items compared to the 36 remaining items. As a result, 3 Three-Dimensional Rotation items appeared in each subtest. Likewise, 8 of the 9 Letter and Number series items were used to allow two to appear in each subtest. Additionally, 8 of the 11 Matrix reasoning items were used placing two in each subtest. Finally, 12 of the Verbal reasoning items were used, resulting in 3 appearing in each subtest for a total of 10 items in each subtest. I selected the Three-Dimensional Rotation items with the lowest item difficulty to maintain a more similar difficulty with the other items. The rest of the items were selected to maximize the variability of item difficulties while maintaining a balance of easier and harder items. Items were allocated to each subtest in such a way as to maintain a standard overall level of difficulty and a standard variability of difficulty. Because the normed item-level difficulties of the ICAR are published, I was able to nearly equate the tests in difficulty (at optimal levels around .50, mean = .504, range: .493 to .512) and standard deviation (0.211 to 0.224). Reliability analyses indicated that subtest reliabilities were comparable; Subtest 1 had an alpha of .69, Subtest 2 had an alpha of .63, Subtest 3 had an alpha of .74, and Subtest 4 had an alpha of .63. The 40-item full test had a much higher alpha of .90. Additionally, scores on the four subtests were highly correlated (average r = .70, range = .66 - .73, all ps <.001).

[2] For more information on the ICAR, visit its companion website icar-project.org.

Procedures

Upon opting to participate in the experiment on MTurk, participants read a description of the study informing them that they would complete four 10-item multiple-choice tests and that, upon completion of the study, they would be compensated $1 for their participation. They were also informed that, after the completion of each test, they would be asked to provide estimations of their performance on each test and that, after they provide these estimations, they may or may not receive feedback on their performance. After this, the participants were then assured that their information would be kept confidential and that they were free to withdraw from the study at any point without consequence. Participants then acknowledged that they understood these terms. Upon accepting these terms, participants read more detailed instructions about the experiment including instructions about how to answer the test questions, how to provide their estimations of their performance, and how to interpret the feedback that is given to them (for the exact wording of these instructions and examples provided to the participants, see Appendix B). Once the participants indicated that they understood these instructions, they completed the four subtests and provide estimations on their performance (described more below). Participants were then asked to answer demographic questions about their age, gender, and education level, debriefed, thanked for their participation, and provided a code they could use to collect their compensation.

Feedback conditions. This study utilized four feedback conditions: a control no feedback condition, an absolute feedback condition, a relative feedback condition, and a combined absolute and relative feedback condition. The procedure common to all four conditions is that, after completing a subtest, participants provided an initial estimation of their absolute and relative score on that subtest. This allowed for an initial comparison to be

made across all four conditions after the completion of the first subtest and score estimations to insure that there were no differences between the conditions at this point.

Then, depending on which condition they were in, participants may or may not have received some form of feedback. The first condition was the no feedback condition and participants did not receive any feedback. Participants in this condition simply completed the first subtest, provided an estimation of their absolute and relative scores, and then repeated the process with the three following subtests. The next condition was the absolute feedback condition. In this condition, participants completed a subtest and provided an initial estimation of their absolute and relative scores. Following this, they were informed of how many of the 10 questions they answered correctly. This process was repeated with each subsequent subtest. The relative feedback condition was identical to the absolute feedback condition except that instead of receiving feedback about the number of questions they answered correctly, participants were provided information on how their performance answering these questions compared with that of the normed ICAR population. Participants in this condition were informed of what percentage of people they scored higher than on each subtest. Participants in the combined absolute and relative feedback condition followed the same procedure but received both forms of feedback after each subtest.

Results

The purpose of this results section is not only to present the findings of this experiment, but also to compare the pattern of results found in most of the previous research where only a single observation was used to those of this study where multiple observations were examined. To do this, this section is organized primarily by the hypotheses detailed

above. When hypotheses are being examined using *t*-tests, bivariate correlations, and One-Way ANOVAs, three separate comparisons are made for each hypothesis.

The first serves to emulate most of the previous research by examining only performance at Trial 1 (T1) and (mis)estimations of performance at T1 (summed values for absolute estimations, averaged values for relative estimations). That is, only the scores and estimations for the first test are examined. The second analysis then examines performance at T1 and total estimations (i.e., perceptions of performance across all four trials). These analyses illustrate how the relationship between performance and estimations differs when the same grouping as the previous analyses (T1 quartiles) are maintained but are compared to total estimations. Lastly, the third analysis examines the relationship of total score with total estimations. This process is presented twice for each hypothesis: one set of comparisons for absolute estimations and one for relative estimations. When Mixed-ANOVAs are used, only the second and third of these comparisons are made. When multilevel modeling is used, only the last comparison is made (See Tables 3 and 4 for *M*s and *SD*s).

Validity Checks

While the effect of feedback on actual performance has not been examined specifically in the context of the UUP, it seemed prudent to perform a validity check to examine if the total scores (i.e., the actual scores of the participants) differed between feedback conditions (Hypothesis 4.0). Results of a one-way ANOVA confirm that feedback had no effect on actual performance, $F(3, 441) = 0.10$, $p = .96$, $\eta_p^2 = .00$.

To examine whether participants understood how to properly estimate their absolute and relative performance, the correlation of the sum of their absolute estimations across subtests and their average relative estimations was calculated. The result of this bivariate

correlation showed that absolute and relative estimations were highly and positively correlated, $r(443) = .64$, $p < .001$, indicating that participants who estimated that they scored highly on the tests also estimated that their score was higher than most other participants. This suggests that the participants did understand how to report their estimates of their absolute and relative performance. When the correlation between absolute and relative estimates were examined for each subtest separately (e.g., absolute and relative estimations for Subtest 1, absolute and relative estimations for Subtest 2), a similar pattern of results was found (average $r = .57$, range = .54 - .57, all ps <.001).

Converging Hypotheses

Hypothesis 1a. The hypothesis that participants would overestimate their scores on average examined the reliance on the BTA heuristic in my sample. Paired-samples *t*-tests compared participants' actual scores to their estimated absolute scores. Results of these analyses demonstrated that the participants' average estimated absolute scores at T1 were significantly higher than their average absolute scores at T1, $t(444) = 5.17$, $p < .001$, $d = 0.25$. When examining participants' performance at T1 to their total absolute estimations, their average estimated absolute scores at T1 were significantly higher than their average absolute scores overall (when total score was converted to a comparable metric), $t(444) = 5.79$, $p < .001$, $d = 0.28$. Lastly, participants' average total estimated absolute scores were significantly higher than their total average absolute scores, $t(444) = 18.71$, $p < .001$, $d = 0.68$. These results replicate previous research on the UUP and the BTA heuristic and show that a stronger overestimation effect was observed when comparing estimates to performance across all four trials.

Additionally, Hypothesis 1a predicted that participants would overestimate their relative scores on average. To test this hypothesis, I conducted paired-sample *t*-tests to determine if participants' average estimated relative scores differed significantly from their true average relative scores. Results of the first analysis showed that participants' estimated relative scores at T1 were significantly lower on average than their relative scores at T1, $t(444) = -3.94$, $p < .001$, $d = -0.19$. When scores at T1 were compared to total relative estimations, I found that participants' total estimated relative scores were still significantly lower on average than their relative scores at T1, $t(444) = -5.45$, $p < .001$, $d = -0.26$. Lastly, participants' total estimated relative scores were significantly lower on average than their true relative average, $t(444) = -6.72$, $p < .001$, $d = -0.27$. This does not support my directional hypothesis and suggests that, on average, the participants in this sample believed they performed worse than the average participant.

Hypothesis 1b. Previous research suggests that there should be a negative relationship between participants' absolute scores and the extent to which they misestimate their absolute performance, such that the poorest performers exhibit the most misestimation. To test this hypothesis, I first calculated the extent to which participants misestimated their absolute scores by finding the difference between their estimated absolute scores and their true absolute scores such that positive values represent overestimation of absolute performance and negative values indicate underestimation. Results of a bivariate correlation found that there was indeed a significant negative correlation between absolute misestimation at T1 and absolute score at T1, $r(443) = -.46$, $p < .001$. The same was true when total absolute misestimations were correlated to scores at T1, $r(443) = -.35$, $p < .001$. Total absolute score and absolute misestimation showed a similar negative relationship, $r(443) = -$

.41, $p < .001$. These results support my hypothesis, replicate the findings of previous researchers, and suggest that the better participants performed, the less they overestimated their absolute scores. However, the strengths of the relationships between scores and absolute misestimations in these three analyses did not differ from each other, though this difference did approach significance, $Z = -1.63$, $p = .052$.

Similarly, Hypothesis 1b also predicted that there would be a negative relationship between participants' actual performance and the extent to which they misestimate their relative scores. To test this hypothesis I first calculated the participants' percentile rank (i.e., relative scores) from their actual scores. I then found the difference between the participants' estimated relative scores and their actual relative scores such that positive numbers indicate overestimation relative to the performance of other participants in the sample. Results of a bivariate correlation between participants' total relative scores and their relative misestimation found that there was indeed a strong significant negative correlation between relative misestimation at T1 and performance at T1, $r(443) = -.62$, $p < .001$. A similar relationship was found between total relative misestimation and score at T1, $r(443) = -.57$, $p < .001$. This pattern persisted when examining the relationship between total relative misestimation and total performance, $r(443) = -.69$, $p < .001$. These results support my hypothesis, replicate the findings of previous researchers, and suggest that the poorest performing participants exhibited the most relative misestimation on average. These results also indicate that the strengths of the relationships between performance and relative misestimations did not differ from each other, $Z = -0.97$, $p = .166$.

Hypothesis 2a. To more closely examine participants' misestimation, and to replicate previous research, misestimations of low and high performers were examined separately.

Previous research has reliably shown that poor performers consistently overestimate their performance and this was hypothesized in the current study. Three paired-samples *t*-tests compared the absolute and estimated scores of participants whose scores fell below the 25th percentile. Results of the first analysis indicated that participants in the bottom quartile on T1 estimated that, on average, they answered significantly more questions correctly than they actually did, $t(93) = 9.02$, $p < .001$, $d = 0.93$. This pattern continued when these participants' (those in the bottom quartile at T1) scores were compared to their total absolute estimates $t(93) = 16.26$, $p < .001$, $d = 1.68$. When comparing the total absolute estimations to the total score of participants in the bottom quartile based on their total score, poor performing participants estimated that they answered significantly more questions correctly than they actually did, $t(99) = 13.53$, $p < .001$, $d = 1.86$. These results indicate that not only did poor performing participants overestimate their absolute scores on average, this overestimation was greatest when all four trials were examined.

Additionally, Hypothesis 2a predicts that poor performing participants would overestimate their relative scores as well. To test this hypothesis, I conducted three paired-samples *t*-tests and found that poor performing participants' (those falling below the 25th percentile on T1) relative score estimations for T1 were significantly higher than their actual average relative score on T1, $t(93) = 7.95$, $p < .001$, $d = 0.82$. This pattern continued when participants' total relative score estimations were compared to their T1 performance, $t(93) = 8.96$, $p < .001$, $d = 0.92$. Similarly, participants whose total score fell below the 25th percentile overestimated their average total relative score, $t(99) = 8.77$, $p < .001$, $d = 0.88$. These results indicate that poor performers overestimated their relative scores on average but this overestimation remained constant across analyses.

Hypothesis 2b. Unlike poor performers, previous research regarding high performers (those whose scores were in the top 25%) is mixed: They sometimes significantly *underestimate* their absolute scores, sometimes slightly *overestimate* their absolute scores, and are sometimes able to accurately estimate their absolute scores (i.e., they do not significantly over- or underestimate their absolute scores). Therefore, Hypothesis 2b proposed a difference for relative estimates only, and the analysis was more exploratory for absolute estimates. Three paired-samples *t*-tests revealed that estimated absolute scores of participants in the top quartile on T1 were significantly lower than their actual absolute score on T1, $t(130) = -3.03$, $p = .003$, $d = -0.26$, indicating that they underestimated their absolute scores on T1. When the absolute scores on T1 of these participants were compared to their total absolute estimates, results indicated that there was no significant difference, $t(130) = -1.67$, $p = .10$, $d = -0.15$. Lastly, when examining total absolute scores and estimations, results indicated that the estimated absolute scores of participants in the top quartile were significantly higher than their actual absolute score, $t(114) = 3.96$, $p < .001$, $d = 0.39$. These results suggest that, in this sample, high performers underestimated their absolute scores on T1. However, they overestimated their absolute scores on average over all subtests (trial was not examined here).

When examining relative scores, however, a different pattern emerged. Results of three paired-samples *t*-tests comparing the relative and estimated relative scores indicated that, for T1, high performers underestimated their relative score compared to their actual relative score, $t(130) = -14.52$, $p < .001$, $d = -1.27$. When comparing total relative estimates and performance on T1, high performers continued to underestimate their performance, $t(130) = -17.39$, $p < .001$, $d = -1.52$. Lastly, when examining total relative scores and

estimates, results indicate that high performers persisted in underestimating their relative performance, $t(114) = -16.46$, $p < .001$, $d = -1.54$. These results suggest that high performers consistently underestimated their relative performance across all four trials.

Hypothesis 3. Previous research suggests that the magnitude of both absolute and relative misestimation will be greater for poor performers than for high performers. To test this hypothesis, I first conducted three one-way ANOVAs examining the differences in absolute misestimation among the four performance quartiles similarly to the sets of analyses above. In these analyses, I also specifically compared the differences in misestimation between the top and bottom quartiles using linear contrasts. Results of the first analysis comparing absolute misestimation at T1 indicated that absolute misestimation differed significantly across quartiles (ranked by T1 scores), $F(3, 441) = 28.98$ $p < .001$, $\eta_p^2 = .17$. Additionally, linear contrasts revealed that absolute misestimation was significantly greater for participants in the bottom quartile than those in the top quartile, $t(441) = 9.23$, $p < .001$, $d = 0.44$. The second analysis examining total absolute misestimation indicated that absolute misestimation differed significantly across T1 quartiles, $F(3, 441) = 17.70$, $p < .001$, $\eta_p^2 = .11$ Additionally, linear contrasts revealed that absolute misestimation was significantly greater for participants in the bottom quartile than those in the top quartile, $t(441) = 7.25$, $p < .001$, $d = 0.34$. The last analysis indicated that total absolute misestimation differed significantly across quartiles (ranked by total scores), $F(3, 441) = 27.12$, $p < .001$, $\eta_p^2 = .16$. Linear contrasts revealed that absolute misestimation was significantly greater for participants in the bottom quartile than those in the top quartile, $t(441) = 8.82$, $p < .001$, $d = 0.42$. These results taken together support my hypothesis that absolute misestimation would be higher for poor performers and replicate previous research.

The same analyses as above were performed on the relative misestimations of participants. Because Levene's statistic indicated a lack of homogeneity of variance, all Fs(3, 441) ≥ 2.73, ps ≤ .043, values that have been adjusted for this violation of assumptions are reported (Brown-Forsythe). Results of the first one-way ANOVA in this set of analyses indicate that relative misestimations also differed significantly across the four quartiles at T1, $F(3, 310.95) = 87.73$, $p < .001$, $\eta_p^2 = .38$. Linear contrasts also revealed that, similar to absolute misestimations, relative misestimations were higher for participants in the bottom quartile than in the top quartile, $t(166.97) = 14.58$, $p < .001$, $d = 1.13$. The second analysis also identified a difference between the total relative misestimations of participants grouped by T1 quartiles, $F(3, 299.01) = 65.01$, $p < .001$, $\eta_p^2 = .32$. Linear contrasts revealed again that relative misestimations were higher for participants in the bottom quartile than in the top quartile, $t(161.73) = 12.69$, $p < .001$, $d = 1.06$. Results of the last one-way ANOVA in this set of analyses indicate that total relative misestimations differed significantly across the four quartiles, $F(3, 399.86) = 122.77$, $p < .001$, $\eta_p^2 = .46$. Linear contrasts revealed that relative misestimations were higher for participants in the bottom quartile than in the top quartile, $t(185.62) = 16.75$, $p < .001$, $d = 1.23$. These results also support my hypothesis that poor performing participants would show greater relative misestimation than high performing participants and replicate previous findings.

Competing Hypotheses

Hypotheses 4.1, 4.2a, & 4.2b. To test the competing hypotheses that environmental feedback (i.e., indirect feedback gleaned across trials that was not explicitly given to the participants) would either aid only the estimations of high performers (Hypothesis 4.1) or would aid the estimations of both high and poor performers (Hypothesis 4.2a) with the

greatest calibration seen for poor performers (Hypothesis 4.2b), I conducted pairs of Mixed ANOVAs analogous to the analyses above. It is important to note that, because misestimations across all four trials are being examined, the first comparison made in each set of the analyses above (comparing T1 scores to T1 estimations) cannot be made here. It is also important to note that, because I am not testing the effect of direct feedback, only the no feedback condition will be examined in these analyses. The first (4 quartile x 4 trial) Mixed ANOVA examined the change in absolute misestimation over the four trials while controlling for T1 quartile rank. The results of this analysis indicated that there was a significant main effect of T1 quartile rank on absolute misestimation, $F(3, 105) = 8.19$, $p < .001$, $\eta_p^2 = .19$, a significant interaction between T1 quartile rank and trial, $F(9, 99) = 1.94$, $p = .046$, $\eta_p^2 = .05$, but no main effect of trial $F(3, 105) = ,40$, $p = .75$, $\eta_p^2 = .00$ (see Figure 4, bottom panel). These results overall indicate that T1 performance significantly affected absolute misestimation and the degree to which misestimation changed across time differed depending on participants' performance such that participants with more extreme scores exhibited greater calibration over time.

To analyze this relationship more closely and to replicate previous analyses, between-subjects contrasts comparing the bottom and top T1 quartiles indicated that low performers reported significantly higher absolute misestimations than high performers, $F(1, 107) = 24.12$, $p < .001$, $\eta_p^2 = .19$. Polynomial within-subjects contrasts examining the interaction between T1 quartile and trial identified a linear, $F(3, 105) = 3.08$, $p = .031$, $\eta_p^2 = .08$, relationship. This contrast did not reveal any effect of trial when examining them independently, $Fs < 1.10$, $ps > .296$. These results partially replicate previous findings and suggest that while there was a negative relationship between T1 performance and absolute

misestimation, there was also a negative relationship between T1 performance and change in absolute misestimation such that poor performing participants showed the most calibration on average. These results repudiate Hypothesis 4.1, partially support Hypothesis 4.2a, and fully support Hypothesis 4.2b.

The second (4 quartile x 4 trial) Mixed ANOVA examining the change in absolute misestimation over the four trials while controlling for total score quartile rank indicated that there was a significant main effect of total score quartile rank on absolute misestimation, $F(3, 108) = 18.92$, $p < .001$, $\eta_p^2 = .15$, no significant interaction between total score quartile rank and trial, $F(9, 99) = 1.69$ $p = .090$, $\eta_p^2 = .05$, and no main effect of trial $F(3, 105) = 0.40$, $p = .756$, $\eta_p^2 = .00$ (see Figure 4, top panel). These results overall indicate that total performance was significantly related to absolute misestimation, but, unlike when T1 quartiles are examined, not the degree to which misestimation changed across time.

Similar to when T1 quartiles are examined, between-subjects contrasts comparing the bottom and top total score quartiles indicated that poor performing participants exhibited greater absolute misestimation than high scoring participants, $F(1, 107) = 25.20$, $p < .001$, $\eta_p^2 = .20$. Polynomial within-subjects contrasts did not detect any differences in the change in absolute misestimation across trials between the top and bottom quartiles, all Fs $< .0.91$, $ps > .342$. These results are somewhat consistent with previous research and the above analysis in that absolute misestimation was negatively related to performance. However, there was no significant calibration of absolute estimations over time across all participants (partially supporting Hypothesis 4.1).

The third (4 quartile x 4 trial) Mixed ANOVA examined the change in relative misestimation in the no feedback condition over the four trials while controlling for T1

quartile rank. The results of this analysis indicated that there was a significant main effect of T1 quartile rank on relative misestimation, $F(3, 105) = 16.20$, $p < .001$, $\eta_p^2 = .32$, no significant main effect of trial $F(3, 105) = 0.59$, $p = .623$, $\eta_p^2 = .01$, and no significant interaction between T1 quartile rank and trial, $F(9, 99) = 1.46$, $p = .162$, $\eta_p^2 = .04$ (see Figure 5, bottom panel). These results overall indicate that T1 performance was significantly related to relative misestimation but not the degree to which misestimation changed across time differed depended on participants' performance. Additionally, like with absolute misestimations, relative misestimations did not change over time.

Between-subjects contrasts comparing the bottom and top T1 quartiles indicated that low performers overestimated their relative estimations compared to high performers, $F(3, 107) = 39.92$, $p < .001$, $\eta_p^2 = .28$. Polynomial within-subjects contrasts identified no significant effects of trial on relative misestimation, all Fs < 1.14, ps > .288. When examining the interaction between T1 quartile and trial, these contrasts identified a marginally nonsignificant linear relationship between T1 quartile and trial, $F(3, 105) = 2.63$, $p = .054$, $\eta_p^2 = .07$. These results partially replicate previous findings and suggest that there was a negative relationship between T1 performance and relative misestimation. While there was no significant main effect of trial nor a significant interaction between T1 quartile and trial (supporting Hypothesis 4.1), the marginally nonsignificant linear relationship between T1 quartile and trial suggests that Hypotheses 4.2a & b be examined further before being discounted.

The fourth (4 quartile x 4 trial) Mixed ANOVA examining the change in relative misestimations of the no feedback group over the four trials while controlling for total score quartile rank indicated that there was a significant main effect of total score quartile rank on

relative misestimation, $F(3, 105) = 23.11$, $p < .001$, $\eta_p^2 = .40$, and a significant interaction between total score quartile and trial, $F(9, 99) = 2.56$, $p = .008$, $\eta_p^2 = .07$. However, similar to the previous analysis, there was no main effect of trial $F(3, 05) = 0.61$, $p = .609$, $\eta_p^2 = .01$ (see Figure 5, top panel). These results are somewhat consistent with previous research and the above analyses in that relative misestimation was negatively related to performance. The significant interaction between total score quartile and trial, coupled with the lack of a significant effect of trial, suggests that, overall, even though relative misestimations did not change over time, they did significantly change for some quartiles (see below).

Between-subjects contrasts comparing the bottom and top total score quartiles indicated that low performers overestimated their relative estimations compared to high performers, $F(1, 108) = 58.07$, $p < .001$, $\eta_p^2 = .36$. Polynomial within-subjects contrasts examining the change in absolute misestimation across trials did not identify any trends, all $Fs < 0.81$, $ps > .371$. The only trend that significantly described the interaction between total score quartile and relative misestimation was a cubic one, $F(3, 105) = 5.77$, $p = .001$, $\eta_p^2 = .14$. These results are somewhat consistent with previous research in that absolute misestimation was negatively related to performance. Additionally, even though no significant main effect of trial on relative misestimation was identified (tentatively supporting Hypothesis 4.1), the cubic interaction between total score quartile and trial suggests that the further participants' score were from the mean, the more they calibrated their scores over trial (supporting Hypothesis 4.2a & b).

Hypotheses 5.1a & 5.1b. To simultaneously examine the effects of differing levels of performance and feedback on misestimation as well as how these effects varied over time, two multilevel models (MLM) were constructed; one model to predict absolute

misestimation, one model to predict relative misestimation. These models were constructed using the Mixed Process in SPSS 22 software. In each analysis, the first predictor entered was the participants' performance as this has consistently been the strongest predictor of misestimation in the literature to date (Ehrlinger et al., 2008, Kruger & Dunning, 1999, Ryvkin et al., 2012). The models were built from this point one step at a time adding first absolute feedback, then relative feedback, and then the repeated measure effect of trial. The decision to include absolute feedback before relative feedback was somewhat arbitrary; however, it seemed prudent to include both forms of feedback before the repeated measures effect to facilitate interpretation of the results. At each step the fit of the model was compared to that of the previous model by examining the change in the -2 Log Likelihood (-2LL). Because the feedback conditions and trial were experimental manipulations in this study, an *a priori* rule was used in model development such that these predictors would be added to the model as fixed effects regardless of whether they significantly improved model fit or not. After these were added, however, the rule that model development would cease at the point that the change in the -2LL was no longer statistically significant based on the χ^2 test was used. Both models used Maximum Likelihood (ML) estimations (to compare model fit of subsequent models) as well as unstructured covariance matrices. Fixed intercepts and slopes were modeled because in every case they produced greater model fit than random effects did. Inferential tests of main effects are derived from the best-fitting model. Tables 5 and 6 (upper panels) outline the overall model statistics. Group-mean and Grand-mean centering were not utilized in these analyses because all predictors and outcomes are in a meaningful metric, and the data are not nested within groups (only within subjects).

As can be seen in Tables 5 and 6, when predicting both absolute misestimation and relative misestimation, the best fitting models are the ones that include all four predictors entered as fixed main effects without any random intercepts, slopes, or interactions. When the dependent nature of this within-subject effect is accounted for in the models (by specifying trial as a repeated measure effect), and data are nested within participants, the model fit is significantly improved in both cases. Additionally, it is important to note that more complex models including random intercepts, slopes, and interactions were tested against the (comparatively) simple models presented above. Specifically, once the four fixed main effects were entered and trial specified as a repeated measure predictor, I tested whether or not allowing the intercepts to vary (i.e., specifying them as random) improved model fit. When it did not for either outcome, I then tested whether or not allowing the slopes of the predictors to vary by specifying them as random effects would improve model fit. Model fit was not improved in this case either. This pattern of nonsignificant improvement to model fit continued when I added the two-way interactions (and subsequently, three- and four-way interactions) of the predictors to the model. At this point, model development and testing ceased. It is important to note that, even though many of the more complex models fit the data equally well (i.e., their fit indices were not *worse* than the simpler model), they did not significantly improve model fit and therefore were less parsimonious.

The two resulting models largely replicate the analyses described above (cf., the two Mixed-ANOVAs examining total score quartile rank and trial). Where the above analyses are somewhat limited in that they can only examine a few select factors at any given time (e.g., the effect of performance across time on misestimation) without a significant loss in power, one of MLM's strengths is that it can simultaneous account for many sources of variance

when predicting an outcome. In light of this, when performance, both forms of feedback, and trial are all considered, there is still a strong overarching negative relationship between performance and each type of misestimation such that the lowest performers exhibit the most misestimation. However, unlike in the above analyses, when the effect of feedback is taken into account, there is a significant negative relationship between trial and absolute misestimation and a negative relationship between trial and relative misestimation that is approaching significance (see bottoms of Tables 5 & 6). These results indicate that misestimations were changing over time but that this change could only be detected when the effect of feedback was taken into consideration. As for the effect of feedback on misestimation, the only effect approaching significance was for absolute feedback on relative misestimation (p = .07). Overall these results support my hypothesis that there would be a negative relationship between performance and misestimation (Hypothesis 1b) as well as my hypothesis that participants would be able to calibrate their performance across trials (Hypothesis 4.2a). It does not support my hypothesis that feedback would have a direct effect on misestimation (Hypotheses 5.1a & b), however.

Hypotheses 6.1, 6.2a, & 6.2b. To better understand the potential regression to the mean effect, I first compared the correlations between subtest scores to the theoretical maximum correlations one would expect to observe on the basis of measurement error alone. Given an average reliability of .67 for each subtest (range .63 to .74), the average maximum correlation one could expect to observe between subtests would be .67 (range .63 to .71), according to the Spearman-Brown prophecy formula. Correlations between subtests averaged .70 (range .66 to .73), which are virtually identical to the predicted range.

Therefore, any subtest-to-subtest performance differences may attributable to unreliability in the measurements.

To further examine this effect, I conducted a repeated-measures ANOVA. Results of this analysis indicated that, overall, participants' scores did not change over trials, $F(3, 1332) = 1.00$, $p = .390$, $\eta_p^2 = .00$. I also conducted a (4: total score quartile x 4: trial) Mixed-ANOVA to examine whether or not change in participants' scores differed by performance level. The results of this analysis indicated that participants' scores differed significantly by quartile (though this effect is essentially meaningless given the nature of the analysis), $F(3, 441) = 22{,}218.80$, $p < .001$, $\eta_p^2 = .98$. More importantly, though, was that there was no significant effect of trial nor an interaction between trial and quartile, Fs < .90, ps > .45. Taken in conjunction with the lack of support for feedback and the strong support for the BTA effect, these findings imply stronger support for the regression to the mean explanation proposed in Hypotheses 6.2 than for the negligible effect proposed in Hypothesis 6.1.

Discussion

The present study expands the current literature on the UUP by both parsing out the effects of different kinds of feedback on different kinds of misestimations while also more closely examining the role of measurement error in explaining the pattern of results seen thus far. Although there are aspects of the UUP phenomenon on which many researchers agree, there are additional aspects on which they disagree. Similar to the sections above, this discussion will be organized in part along that distinction. First, I will discuss the findings concerning and the implications of the hypotheses that the previous researchers largely agree on. Following this, I will discuss the findings concerning the competing hypotheses on the

effect of feedback on misestimation. After this, I will discuss the findings relating to the competing regression to the mean hypotheses.

UUP findings

The first hypothesis concerns overestimation. While previous research has shown that participants reliably overestimate their absolute and relative performance on average (Ehrlinger et al., 2008; Kruger & Dunning, 1999), this was not always the case in this study. Even though participants in the current study did overestimate their absolute performance on average, and this overestimation was greatest when comparing total scores to total absolute estimations, the opposite pattern was found for relative estimations. The approach of this study was to examine misestimation when the average number of questions answered correctly was approximately half (50%) of the total number of questions to avoid the measurement limitations associated with ceiling or floor effects. In the scale development process, I utilized the normed item-level difficulty data provided by Condon and Revelle (2014; see their supplementary online material as well) to construct four subtests that had difficulties as close as possible to 50% correct while also maintaining approximately equal variability in difficulty. The participants in the current sample answered 18 of 40 questions correctly on average, but estimated they had answered approximately 23 of 40 correctly. Much of the previous research relies on measures vulnerable to ceiling effects. This has raised the question of whether overestimation of performance could be in large part an artifact of using "easy" tests (see also Dunning, 2011 for a remark on this potential limitation and Kruger, 1999 for a demonstrations of a "below-average" effect). This study has demonstrated that, even when an arguably more difficult test is used, a very strong tendency to overestimate one's absolute performance can still be elicited.

The same was not found for relative estimations, however. Because relative estimations and scores are measured in percentiles (and thus centered at 50%), average relative estimations below the 50th percentile indicate underestimation. In this sample, when comparing total relative estimations to total relative scores, participants estimated on average that they performed better than 43% of other participants in the sample, statistically significantly different from the true average (50%). These two results taken together present an interesting picture. These results seem to suggest that while participants thought they did a good deal better than they actually did on average, they were still able to identify the test as being a difficult one.

The second hypothesis, that misestimations would be negatively related to performance, more directly addressed the UUP. The results of this study directly support this hypothesis and replicate previous findings on the UUP by demonstrating that the poorer participants performed, the greater their absolute and relative overestimation. These results are interesting because, according to Krueger and Mueller (2002), correcting for the asymmetrical distribution of scores around the median possible correct (which was controlled in this study by constructing a test with an average score of approximately 50% correct) should help to reduce the asymmetry in the magnitude of misestimations of participants who scored above and below the mean. However, this was not the case in the current study. The results of the analyses addressing this hypothesis are not entirely conclusive because they do not take into account the effects of the experimental manipulations present in this study on misestimation and cannot compare the pattern of misestimation of multiple groups that took tests of varying difficulties. Thus, while not definitive, the results of these analyses suggest that the asymmetrical pattern of misestimation may not be due entirely to statistical artifact

seeing how a negative relationship between performance and misestimation was found even when utilizing a test that is both highly reliable and has an average score of approximately 50%. Moreover, employing a repeated-measure design that allowed participants to make multiple estimates of performance reduced the likelihood that observed misestimations were simply a result of chance errors, as is possible when single observation methods are utilized.

While the BTA heuristic and the negative relationship between performance and misestimation are fundamental aspects of the UUP, another focus of the UUP literature has been on differing patterns of misestimation across levels of performance. Specifically, previous research suggests that poor performing participants would overestimate both their absolute and relative scores (Hypothesis 2a) while high performing individuals would underestimate their relative scores but not necessarily their absolute scores (Hypothesis 2b; Ehrlinger et al., 2008). The results of the analyses addressing Hypothesis 2a strongly support the prediction that poor performers would significantly overestimate both their absolute and relative performance. When absolute misestimations are looked at specifically, however, it should be noted that the smallest effect size was found when comparing T1 estimations to total scores, the second smallest effect size was found when comparing T1 estimations to T1 scores, and the largest effect size was found when total estimations were compared to total scores. The large difference in overestimation between the T1-T1 and Total-Total comparisons suggests that absolute estimation increased across trials for poor performers. The same was not true when examining relative overestimation for poor performers as the magnitude of overestimation remained relatively constant. However, this pattern of results might be explained by the increased reliability of the Total scores, relative to the T1 scores. More detailed analyses of the change in misestimation over time will be discussed below.

In comparison to poor performers, high performers exhibited a different pattern of results. When comparing absolute scores and estimates at T1, high performers demonstrated a significant underestimation of their performance similar to what has been observed in previous research on the UUP. However, the difference between absolute estimations and scores disappeared when comparing T1 scores to total absolute estimations. This trend continued and ultimately resulted in a reversal of the pattern of misestimation when total score were compared to total absolute estimations such that, at this point, high performers significantly *overestimated* their performance. Previous research on the UUP almost ubiquitously utilizes single-observation designs that the T1-T1 comparison of absolute scores and estimations here emulates. When a repeated-measure design is used and scores and estimations across all observations are compared, however, it is apparent that the tendency for high performers to underestimate their absolute scores not only reverses, the overestimation across the entire series of tests is greater in magnitude than the initial underestimation. Taken together with the results addressing poor performers' absolute misestimations, these results suggest that instead of an unequal and opposite direction of absolute misestimation between low and high performers (Ehrlinger et al., 2008; Kruger & Dunning, 1999) or even a symmetrical and opposite, but lesser, difference in absolute misestimation (Krueger & Mueller, 2002), a different pattern may be taking place. For both high and low performers, the average absolute estimations seem to be increasing over time relative to those at the first trial of the experiment.

What this may suggest is not an opposite effect of performance on misestimation such that poor performers overestimate their scores and high performers underestimate theirs. Instead, it may suggest that there is a similar positive linear trend in absolute estimations for

low and high performers but that the points at which they initially anchor their estimations differ. If this were the case, research designs employing single observations would fail to detect this trend and might erroneously infer that some set of cognitive biases (those specified originally by the researchers proposing the UUP) or randomly distributed imprecision of ability measures and participant self/other assessments (Krueger & Mueller, 2002) explains the previously observed pattern of results. The data here, however, suggest that while some cognitive process is most likely at work and this pattern of results is not simply a result of poor reliability alone, it is not necessarily caused by the same biases proposed originally to explain it.

To complicate the matter further, the results concerning high performers' relative misestimations follow a different pattern. As can be seen above, high performers greatly underestimated their relative scores and this underestimation remained relatively constant across all three comparison points. Taken together with the results demonstrating poor performers' reliable overestimation of their relative scores, the explanations given both by the proponents of the UUP (Ehrlinger et al., 2008; Kruger & Dunning, 1999) and Krueger and Mueller (2002) are plausible here. The direction of relative misestimation replicates what has been found in previous research on the UUP, supporting in part the explanations provided by these researchers. However, the relative magnitudes of misestimation here do not replicate the results of previous researchers. Instead, high performers actually demonstrated a greater magnitude of relative misestimation than did poor performers. As mentioned earlier, the average percent correct across all four tests was 44.9%, which is slightly under half; this could presumably explain the reversal in the asymmetry in magnitude of misestimation as implied by Krueger and Mueller.

Effects of feedback

It is clear from the results discussed thus far that many of the participants in this study exhibited significant inaccuracy when estimating their performance; however, the main novel question this study aimed to answer regarded the information people need to make accurate self-assessments. Previous research by Ryvkin et al. (2012) proposed this question, and the current study directly expanded upon their investigation. While the original proponents of the UUP suggested that people, especially poor performers, generally lack the requisite skills necessary for accurate self-insight (and thus misestimate their performance; Ehrlinger et al., 2008; Kruger & Dunning, 1999), Ryvkin et al. (2012) found that poor performers not only recognize their poor performance, but that they calibrated their estimations of their performance even without explicit feedback concerning their performance when given enough time. In fact, given that poor performers' misestimations are generally the most inaccurate at first and therefore have the most room to improve, Ryvkin et al. (2012) predicted that poor performers would show the *most* calibration in their estimations over time; a prediction that was supported by the results of their study.

This set of competing hypotheses was tested in two waves in this study. The first utilized the mixed-ANOVAs described above that treated the data analogously to the previous research (by grouping performance into quartiles). In these analyses, the change in both absolute and relative misestimations of participants in different quartiles in the no feedback condition was compared across the four trials. When participants were grouped by their T1 performance it was found that participants in the top and bottom quartiles demonstrated some calibration over time, partially supporting the hypotheses given by Ryvkin et al. (2012). Interestingly, though, this effect disappeared when participants were

grouped by their total score quartiles. This discrepancy could have possibly resulted from the movement of participants in one quartile at T1 to another quartile when total scores were examined. Should this be the case, these results support the regression to the mean hypothesis by emphasizing that participants' scores on a first (i.e., single) observation may not actually be indicative of their true ability. Thus, indicators of misestimation (i.e., the difference of estimated scores and actual scores) based on single observations may be exaggerated, overemphasizing the differences in high and low performers' ability to estimate their own performance. Additionally, the cubic relationship of T1 quartile rank and trial on misestimation (i.e., that the top and bottom quartiles demonstrated more calibration than did the two middle quartiles) also supports the regression to the mean hypothesis.

One limitation of the mixed-ANOVAs used in this study is that they did not examine the effect of different kinds of feedback directly. While it is possible to conduct an ANOVA of this type with three factors instead of just the two used here, doing so would have resulted in a 4-quartile x 4-feedback condition x 4-trial analysis. This analysis would have been comprised of 64 cells and the current sample size of 445 would have had inadequate power to detect the effects of interest. This concern, though, is obviated by the fact that performance was measured as a continuous variable and thus the MLM described above could examine all three factors simultaneously without suffering the same loss of power that utilizing a 4 x 4 x 4 mixed-ANOVA would have encountered.

By examining the data using MLM, I was able to not only measure the change in misestimation over time, but was also able to take into account the effect of feedback while doing so. In contrast to the results of the mixed-ANOVAs, the two resulting MLMs identified significant main effects of performance *and* trial (though, for relative misestimation, the

effect of trial only approached significance). What this indicates is that misestimations do in fact improve over time, but this change may only be observed when the feedback condition is accounted for. Considering the stronger effect of trial on misestimation in MLM analyses than in the Mixed-ANOVA analyses, this would suggest that an interaction might exist between two or more of the predictors or that feedback (with a marginally nonsignificant effect being of absolute feedback on relative misestimation) or that the effect of one of the predictors was being suppressed by another. While significant interactions were detected when more complex models were examined, their addition did not provide a significant benefit to the fit of the model, so they were not be reported or interpreted here.

These results as a whole, while supporting some of the claims of the original proponents of the UUP, seriously challenge others. It is apparent that the design and analyses of previous research have overemphasized (though not completely misconstrued) the lack of individuals' ability to estimate their own performance. Poor performing individuals (as well as high performing ones) can and do improve their estimations of their performance when given the requisite time and sources of information to draw on. Additionally, the asymmetry in misestimation originally identified by the proponents of the UUP does seem to be at least in part a result of the ability measures that have been used and not so much the participants themselves. Curiously, while some analyses implied an effect of feedback, once trial was added as a predictor, feedback appeared to be redundant; therefore, the nature of its relationship to misestimation and how this relationship may depend on an individual's level of performance is still unclear.

One limitation of the analyses of this study was that the feedback that each participant received was unique to that participant; the feedback each participant received was based on

his or her performance such that high performers received different feedback than did low performers. Related to this, it is difficult to make firm conclusions about the effect of feedback on misestimation because even though whether or not participants received feedback was explicitly manipulated, the specific feedback they received was not (it was instead based on their actual performance). Being able to compare the effects of accurate feedback versus inaccurate feedback (e.g., telling participants they did well when they did not) could help to further elucidate this phenomenon by removing the dependency between participants' scores and the feedback they receive. It would also be helpful to vary the difficulty of the test in a single study so that the effects of different test difficulties can be compared directly. Even though the order of the subtests in this study was randomized, all participants ultimately took the exact same 40-item test and therefore the objective difficulty of the test was constant. While this is helpful in many ways, utilizing tests of varying difficulties would allow researchers to study not only how misestimation differs for high and low performers, but whether or not the pattern of misestimation is dependent in part on the difficulty of the test itself.

Ultimately, while this study serves to further elucidate the UUP by demonstrating that participants (especially those who performed poorly) were not entirely blind to their poor performance and that the design of the study and the ability measure used do play some part in explaining the pattern of results seen thus far, it also demonstrates that there is a reliable and systematic pattern of misestimation that may be resulting from some cognitive bias. Additionally, contrary to what was found by Ryvkin et al. (2012), explicit feedback on one's performance seemed to have a minimal effect on misestimation. What this suggests is that most people, even those who perform poorly, can and will make an effort to improve their

estimations of their own performance, given enough time. The information they use to make these calibrations, however, is still unclear, as calibration was seen even in the absence of feedback. Researchers attempting to discover how to mitigate inaccurate self-insight should not assume that there is no hope for the unskilled. However, providing individuals with the information they need to make accurate self-assessments of performance, both relative and absolute, may need to be more of an iterative, contextualized, experiential process such as the classroom setting used in Study 1 by Ryvkin et al. than an objectively explicit one such as the feedback given in the current study.

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, *49,* 1621-1630. doi:10.1037/0022-3514.49.6.1621

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43,* 52-64. doi:10.1016/j.intell.2014.01.004

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson, & M. P. Zanna (Eds.), *Advances in experimental social psychology, Vol 44* (pp. 247-296). San Diego, CA: Academic Press. doi:10.1016/B978-0-12-385522-0.00005-6

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.

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

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39,* 175-191. doi:10.3758/BF03193146

Guenther, C. L., & Alicke, M. D. (2010). Deconstructing the better-than-average effect. *Journal of Personality and Social Psychology*, *99,* 755-770. doi:10.1037/a0020959

Hartwig, M. K., & Dunlosky, J. (2013). 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

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's Mechanical Turk. *Computers in Human Behavior*, *29,* 1749-1754. doi:10.1016/j.chb.2013.02.020

Krajč, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, *29,* 724-738. doi:10.1016/j.joep.2007.12.006

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

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77,* 221-232. doi:10.1037/0022-3514.77.2.221

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

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13,* 279-301. doi:10.1016/0022-1031(77)90049-X

Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? *Journal of Economic Psychology*, *33,* 1012-1031. doi:10.1016/j.joep.2012.06.003

Sheldon, O. J., Dunning, D., & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Reactions to feedback about deficits in emotional intelligence. *Journal of Applied Psychology*, *99,* 125-137. doi:10.1037/a0034138

Table 1

Converging Hypotheses and Associated Statistical Tests

Hypothesis Label	Abridged Hypothesis	Statistical Test
1a	Replicate BTA Heuristic	Paired-sample *t*-test
1b	Replicate UUP	Correlation
2a	Overestimation of poor performers	Paired-sample *t*-test
2b	Underestimation of high performers	Paired-sample *t*-test
3	Misestimation greater for poor performers	One-way ANOVA with planned contrast

Table 2

Competing Hypotheses and Associated Statistical Tests

Hypothesis Label	**Ehrlinger et al. (2008) and Kruger and Dunning (1999)**	Hypothesis Label	**Ryvkin et al. (2012)**	Statistical Test
4.1	Only high performers will gain from indirect feedback	4.2a	Both high and low performers will gain from indirect feedback	Mixed-ANOVA & multiple regression
		4.2b	Poor performers will benefit the most	Mixed-ANOVA & multiple regression
5.1a	Absolute performance feedback will help high and low performers better calibrate all estimations	5.2	(Makes no prediction of varying effects of absolute vs. relative feedback.)	Mixed-ANOVA & multiple regression
5.1b	Relative performance feedback will aid high performers' estimations but exacerbate low performers' relative misestimations			Mixed-ANOVA & multiple regression
	Ehrlinger et al. (2008) and Kruger and Dunning (1999)		**Krueger and Mueller (2002)**	
6.1	Regression to the mean is negligible	6.2a	Participants' score will regress to the mean across trials	Regression & SEM
		6.2b	This effect will be largest for participants with the most extreme scores	Regression & SEM

Table 3

Means (and SDs) of T1 Scores, Estimations, and Misestimations across T1 Quartiles

	Q1 (n = 94)	Q2 (n = 56)	Q3 (n = 164)	Q4 (n = 131)	Total (n = 445)
Abs. Score	1.53 (0.63)	3.00 (0.00)	4.55 (0.50)	7.24 (1.27)	4.51 (2.24)
Est. Abs. Score	3.35 (1.88)	3.70 (1.79)	5.02 (1.81)	6.74 (1.82)	5.01 (2.24)
Abs. Misest.	1.83 (1.97)	0.70 (1.79)	0.46 (1.78)	-0.50 (1.90)	0.50 (2.03)
Rel. Score	10.67 (5.32)	27.52 (0.00)	52.25 (9.19)	85.39 (8.17)	50.11 (28.62)
Est. Rel. Score	34.62 (27.75)	37.63 (24.71)	41.90 (22.77)	56.95 (23.37)	44.26 (25.73)
Rel. Misest.	23.94 (29.20)	10.10 (24.71)	-10.34 (23.62)	-28.44 (22.42)	-5.86 (31.37)

Note. Est. = Estimated, Misest. = Misestimation, Abs. = Absolute, Rel. = Relative, numbers in parentheses are standard deviations.

Table 4

Means (and SDs) of Actual Scores, Estimations, and Misestimations across Total Score Quartiles

	Q1 (n = 100)	Q2 (n = 108)	Q3 (n = 122)	Q4 (n = 115)	Total (n = 445)
Abs. Score	7.86	13.37	19.90	29.01	17.96
	(1.81)	(1.85)	(1.99)	(3.66)	(8.18)
Est. Abs. Score	16.81	19.72	24.92	31.10	23.43
	(6.57)	(4.88)	(5.52)	(6.55)	(7.99)
Abs. Misest.	8.95	6.35	5.02	2.10	5.47
	(6.62)	(4.89)	(5.51)	(5.68)	(6.17)
Rel. Score	11.35	34.71	60.56	87.19	50.11
	(6.39)	(6.92)	(7.85)	(7.46)	(28.88)
Est. Rel. Score	31.56	35.55	43.44	59.72	43.06
	(21.70)	(17.39)	(18.09)	(18.86)	(21.77)
Rel. Misest.	20.21	0.83	-17.12	-27.47	-7.05
	(23.06)	(17.76)	(19.02)	(17.89)	(26.34)

Note. Est. = Estimated, Misest. = Misestimation, Abs. = Absolute, Rel. = Relative, numbers in parentheses are standard deviations.

Table 5

Multilevel Model of Absolute Misestimation Predicted by Participants' Performance, Presence/Absence of Absolute Feedback, Presence/Absence of Relative Feedback, and Trial

Model	***F***	***p***	$\Delta\chi^2$	
1: Score	$F(1,1780) = 509.83$	$p < .001$	--	--
2: Score	$F(1, 1780) = 511.06$	$p < .001$	$\chi^2(4) = 7.28$	$p = .122$
Abs. FB.	$F(1, 1780) = 7.30$	$p = .007$		
3: Score	$F(1, 1780) = 510.03$	$p < .001$	$\chi^2(5) = 1.85$	$p = .869$
Abs. FB.	$F(1, 1780) = 7.58$	$p = .006$		
Rel. FB.	$F(1, 1780) = 1.84$	$p = .175$		
4: Score	$F(1, 1664.39) = 894.89$	$p < .001$	$\chi^2(6) = 563.08$	$p < .001$
Abs. FB.	$F(1, 380.94) = 2.38$	$p = .124$		
Rel. FB.	$F(1, 381.01) = 0.40$	$p = .526$		
Trial	$F(1, 440.95) = 5.57$	$p = .019$		
Final Model	***b* (*SE*$_b$)**	***p***		
Score	$b = -.64$ (.02)	$p < .001$		
Abs. FB.	$b = -.23$ (.15)	$p = .124$		
Rel. FB.	$b = -.09$ (.15)	$p = .526$		
Trial	$b = -.07$ (.03)	$p = .019$		

Note. All *b*'s are unstandardized. Abs = Absolute. Rel = Relative. FB = Feedback.

Table 6

Multilevel Model of Relative Misestimation Predicted by Participants' Performance, Presence/Absence of Absolute Feedback, Presence/Absence of Relative Feedback, and Trial

Model	***F***	***p***	$\Delta\chi^2$	
1: Score	$F(1,1780) = 947.52$	$p < .001$	--	--
2: Score	$F(1, 1780) = 950.4$	$p < .001$	$\chi^2(4) = 7.67$	$p = .104$
Abs. FB.	$F(1, 1780) = 7.68$	$p = .006$		
3: Score	$F(1, 1780) = 950.88$	$p < .001$	$\chi^2(5) = 0.32$	$p = .997$
Abs. FB.	$F(1, 1780) = 7.55$	$p = .006$		
Rel. FB.	$F(1, 1780) = 0.32$	$p = .573$		
4: Score	$F(1, 1520.34) = 1012.85$	$p < .001$	$\chi^2(6) = 453.76$	$p < .001$
Abs. FB.	$F(1, 418.78) = 3.22$	$p = .073$		
Rel. FB.	$F(1, 418.86) = 0.20$	$p = .652$		
Trial	$F(1, 442.57) = 3.84$	$p = .051$		
Final Model	***b* (*SE*$_b$)**	***p***		
Score	$b = -9.41$ (2.30)	$p < .001$		
Abs. FB.	$b = -3.34$ (0.30)	$p = .073$		
Rel. FB.	$b = 0.84$ (1.86)	$p = .652$		
Trial	$b = -0.76$ (0.39)	$p = .051$		

Note. All *b*'s are unstandardized. Abs = Absolute. Rel = Relative. FB = Feedback.

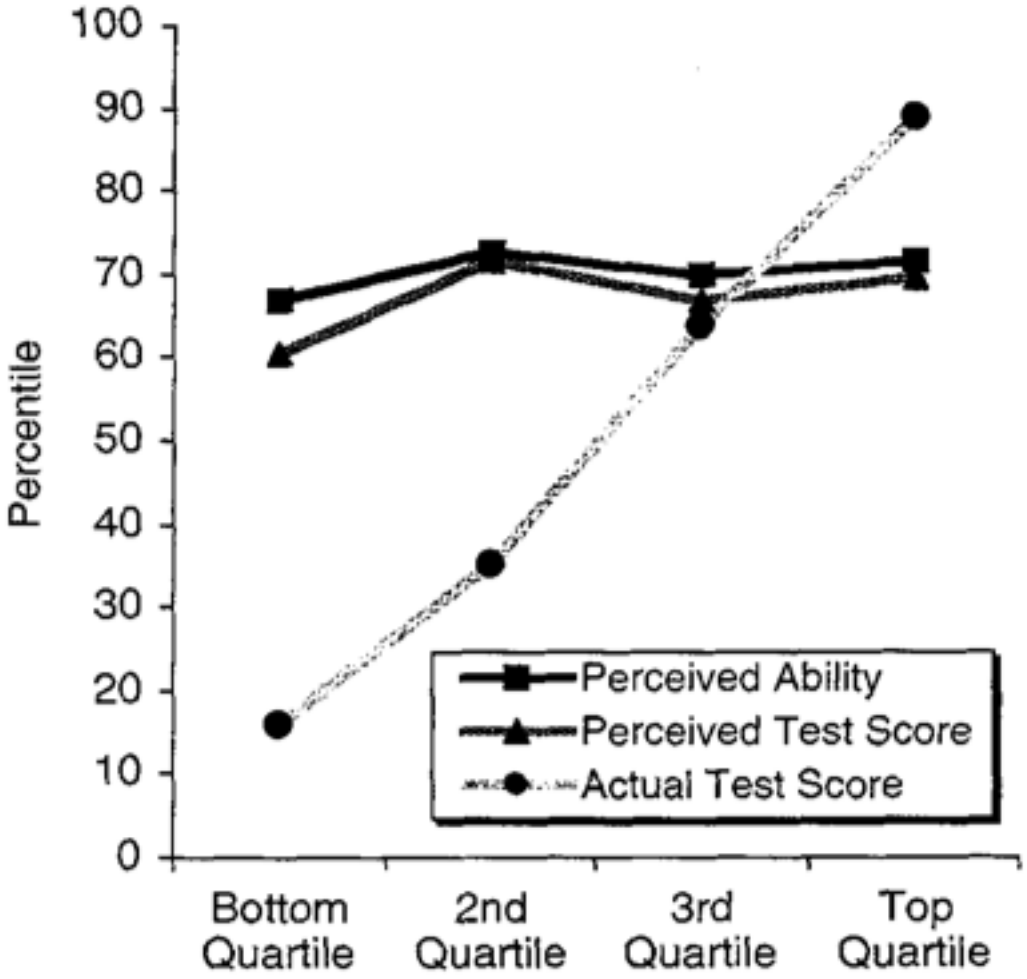


Figure 1. Perceived grammar ability and test performance as a function of test performance. Reproduced from Kruger and Dunning (1999), Figure 3.

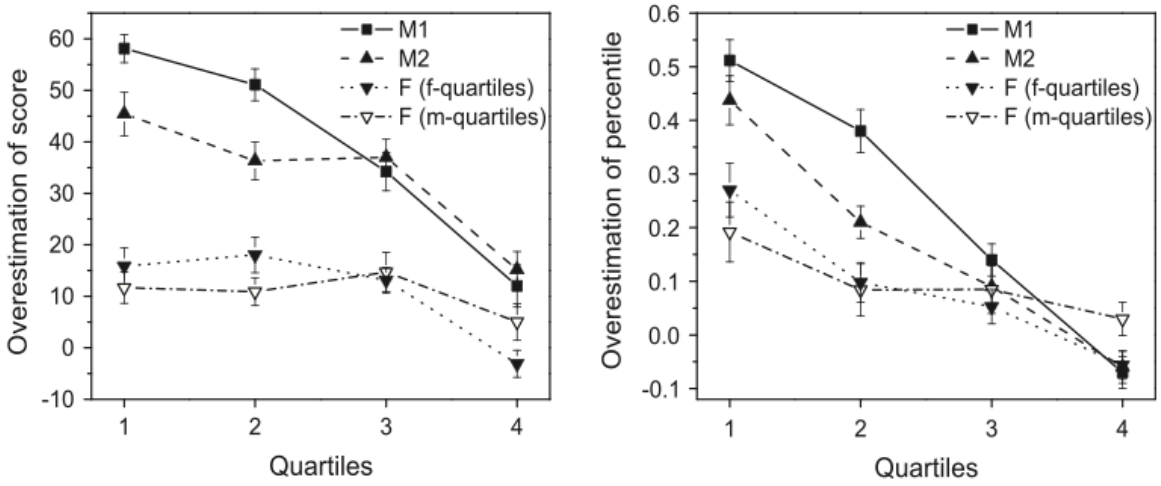


Figure 2. Study 1: Average overestimation of score (left) and percentile (right) by performance quartile in predictions M1 (first midterm prediction), M2 (second midterm prediction) and F (final exam prediction), with error bars. The results for prediction F are shown using the breakdown by quartiles based on final performance (f-quartiles) and midterm performance (m-quartiles). Reproduced from Ryvkin et al. (2012), Figure 1.

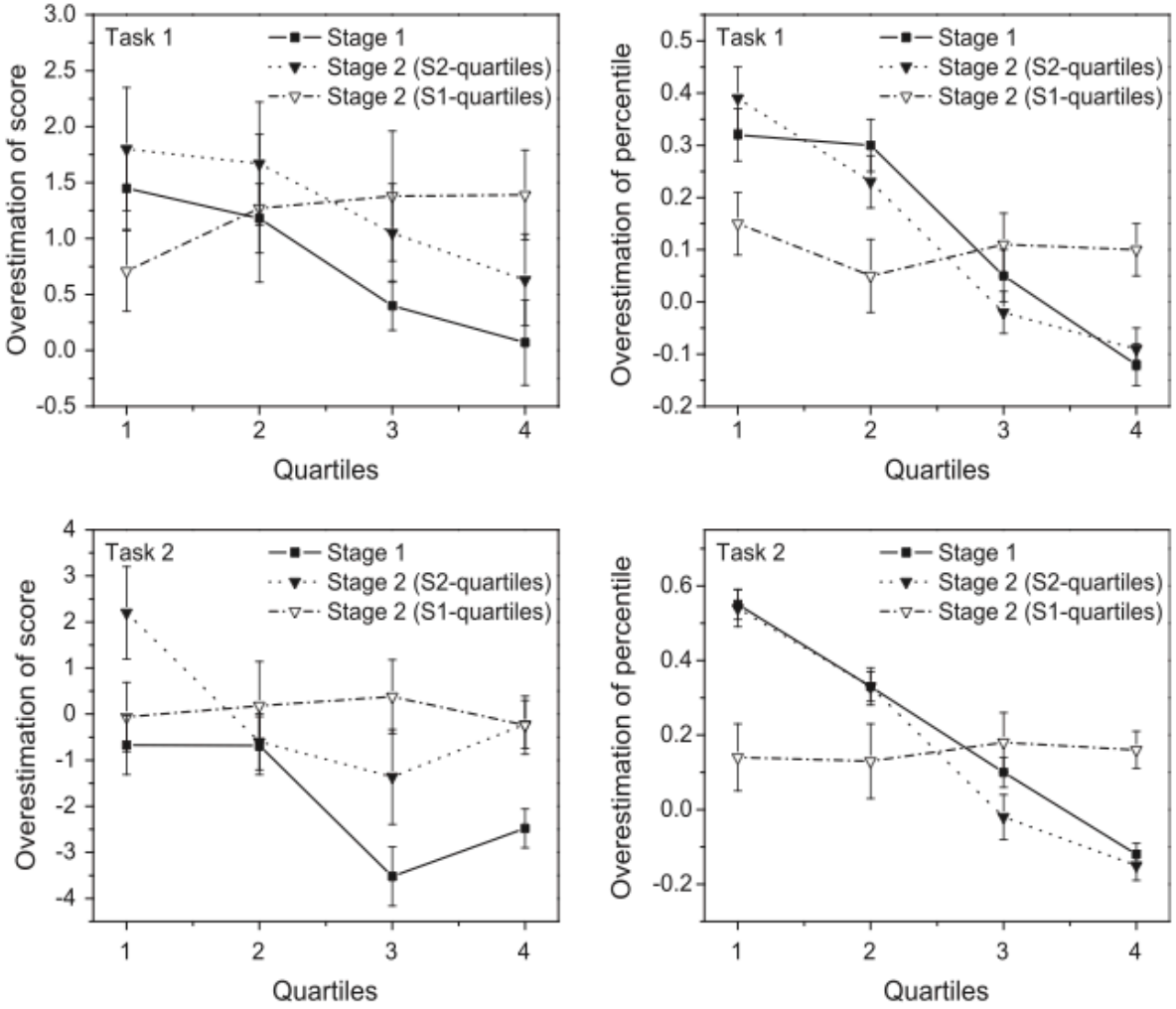


Figure 3. Study 2: Average overestimation of score (left) and percentile (right) by performance quartile in Task 1 (top) and Task 2 (bottom) at Stages 1 and 2, with error bars. The results for Stage 2 are shown using the breakdown by quartiles based on Stage 2 performance (S2-quartiles) and Stage 1 performance (S1-quartiles). This figure demonstrates the effect of indirect feedback on estimation calibration. Reproduced from Ryvkin et al. (2012), Figure 2.

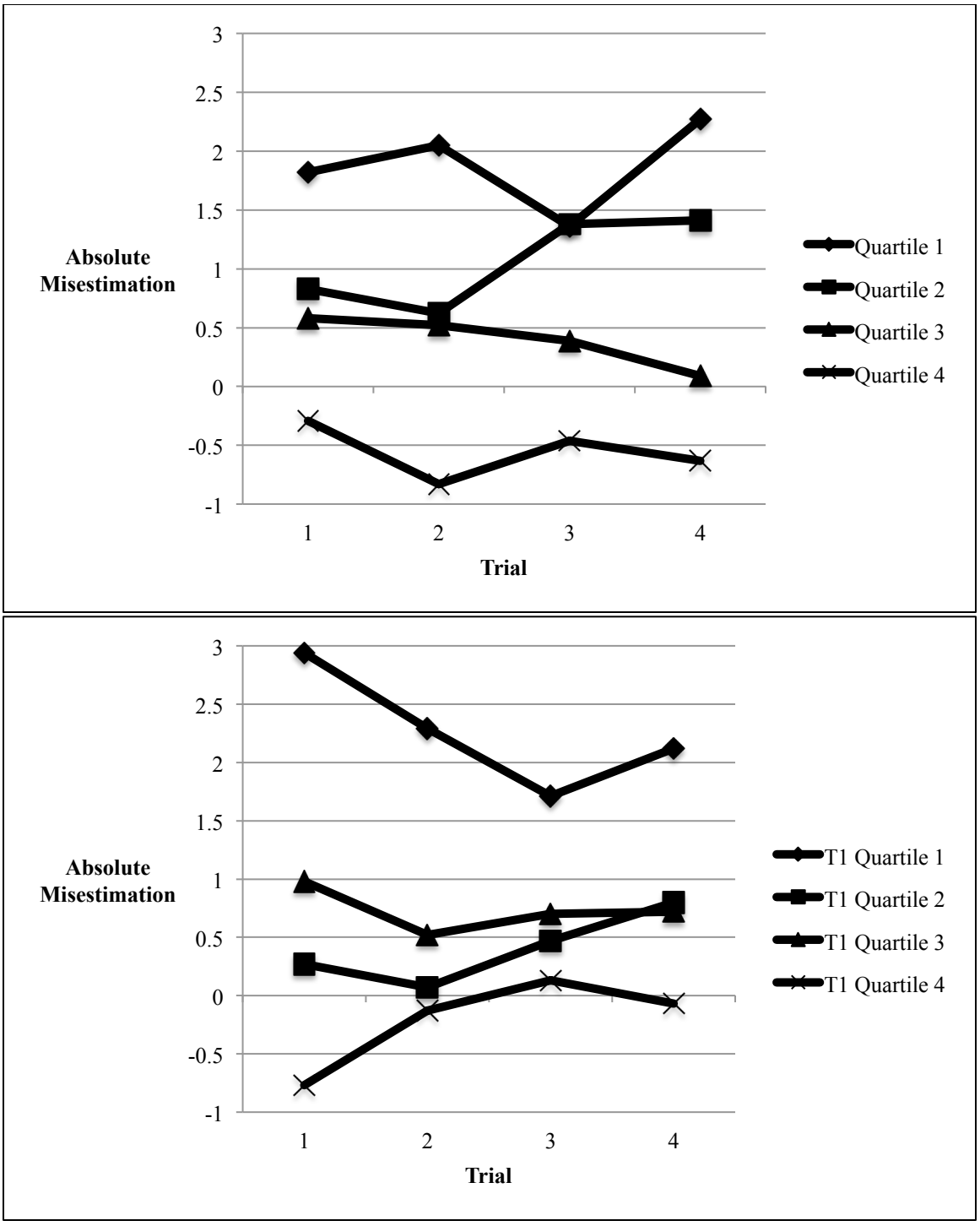


Figure 4. Absolute misestimation of the no feedback condition as a function of total score quartile (top) or T1 quartile (bottom) and trial. The figure compares the change in absolute misestimation over time when participants are grouped by their total score vs. their T1 score.

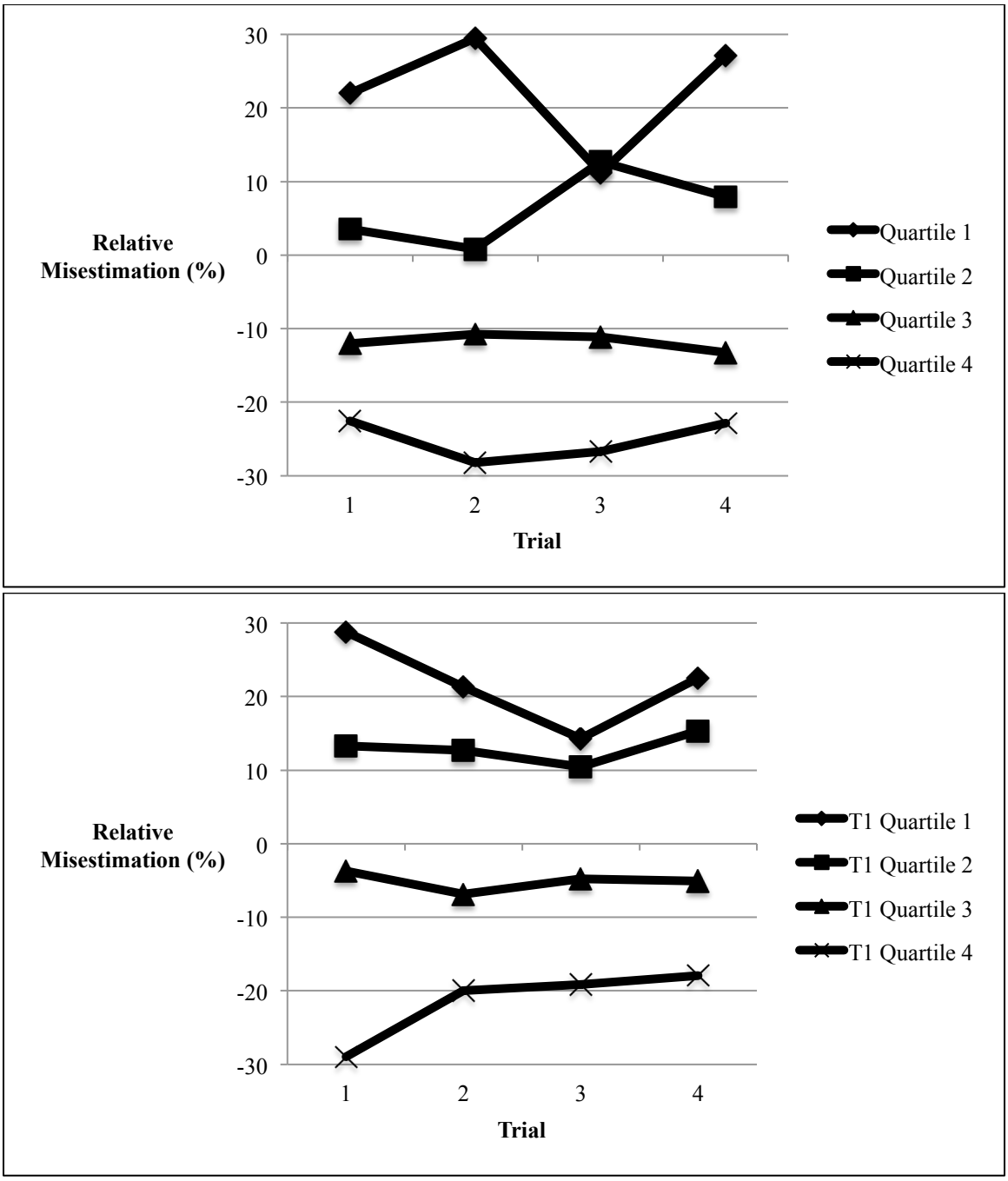


Figure 5. Relative misestimation of the no feedback condition as a function of total score quartile (top) or T1 quartile (bottom) and trial. The figure compares the change in relative misestimation over time when participants are grouped by their total score vs. their T1 score.

Appendix A

IRB Approval

From: Dr. Lisa Curtin, Institutional Review Board Chairperson
Date: 12/08/2014
RE: Notice of IRB Approval by Expedited Review (under 45 CFR 46.110)
Study #: 15-0140

Study Title: Deconstructing the Unskilled-and-Unaware Problem: Examining the Effect of Feedback on Misestimation While Disentangling Cognitive Bias from Statistical Artifact
Submission Type: Initial
Expedited Category: (7) Research on Group Characteristics or Behavior, or Surveys, Interviews, etc.
Approval Date: 12/08/2014
Expiration Date of Approval: 12/07/2015

The Institutional Review Board (IRB) approved this study for the period indicated above. The IRB found that the research procedures meet the expedited category cited above. IRB approval is limited to the activities described in the IRB approved materials, and extends to the performance of the described activities in the sites identified in the IRB application. In accordance with this approval, IRB findings and approval conditions for the conduct of this research are listed below.

Approval Conditions:

Appalachian State University Policies: All individuals engaged in research with human participants are responsible for compliance with the University policies and procedures, and IRB determinations.

Principal Investigator Responsibilities: The PI should review the IRB's list of PI responsibilities. The Principal Investigator (PI), or Faculty Advisor if the PI is a student, is ultimately responsible for ensuring the protection of research participants; conducting sound ethical research that complies with federal regulations, University policy and procedures; and maintaining study records.

Modifications and Addendums: IRB approval must be sought and obtained for any proposed modification or addendum (e.g., a change in procedure, personnel, study location, study instruments) to the IRB approved protocol, and informed consent form before changes may be implemented, unless changes are necessary to eliminate apparent immediate hazards to participants. Changes to eliminate apparent immediate hazards must be reported promptly to the IRB.

Approval Expiration and Continuing Review: The PI is responsible for requesting continuing review in a timely manner and receiving continuing approval for the duration of the research with human participants. Lapses in approval should be avoided to protect the welfare of

enrolled participants. If approval expires, all research activities with human participants must cease.

Prompt Reporting of Events: Unanticipated Problems involving risks to participants or others; serious or continuing noncompliance with IRB requirements and determinations; and suspension or termination of IRB approval by an external entity, must be promptly reported to the IRB.

Closing a study: When research procedures with human subjects are completed, please

complete the Request for Closure of IRB review form and send it to irb@appstate.edu.

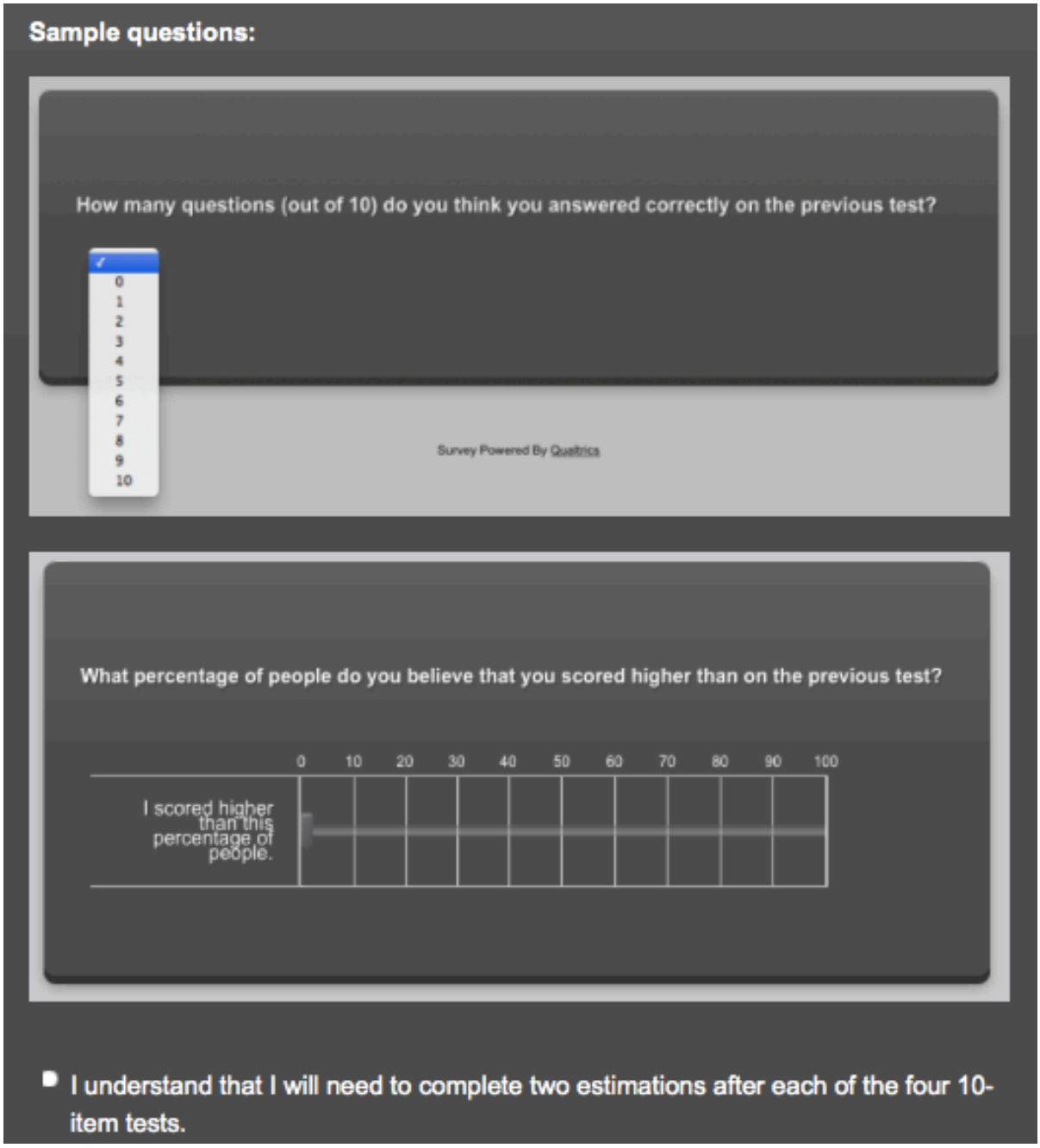
Appendix B

Instructions for Providing Estimations

Instructions. Read Carefully.

In this study you will be asked to complete four 10-item tests. Please read each question carefully and choose the response that is most appropriate. Please try to answer the questions as accurately as possible without spending too much time on any one question. Please note that the response options might change. You will have the option to indicate that you do not know the answer to a question; however, please try to identify the correct answer to the question before selecting this option.

After each test you will be asked to make two estimations about your performance on the test you just completed. The first question will ask you to estimate the number of questions out of 10 that you believe you answered correctly. The second question will ask you to estimate the percentage of people who you think you scored higher than. The question will provide you with a slider bar (see below) with which you can indicate the percentage of people who you believe you scored higher than. For example, if you were to set the slider bar to "80", that would indicate that you believe that you scored higher than 80% of other people who have taken the test that you just took.


Sample questions:
How many questions (out of 10) do you think you answered correctly on the previous test?
0
1
2
3
4
5
6
7
8
9
10
Survey Powered By Qualtrics
What percentage of people do you believe that you scored higher than on the previous test?
0 10 20 30 40 50 60 70 80 90 100
I scored higher than this percentage of people.
I understand that I will need to complete two estimations after each of the four 10-item tests.

Vita

Jeffrey Chase Hood was born in Birmingham, Alabama, the son of Jeff and Linda Hood. He graduated from Christ Church Episcopal High School in 2008 and began his undergraduate studies at Furman University. He received a Bachelor of Arts in Psychology and Philosophy in May 2012. After a year spent with his grandmother, Chase began working towards a Master of Arts degree in General Experimental Psychology at Appalachian State University in the Fall of 2013 and was awarded the degree in August 2015. He plans to begin study at Kansas State University in August of 2015, where he will pursue a Doctorate degree in Experimental Psychology with a focus on Cognitive Psychology, Educational Practices, and Quantitative Methodology.