Published in final edited form as:
Organ Behav Hum Decis Process. 2008 January 1; 105(1): 98–121. doi:10.1016/j.obhdp.2007.05.002.

Why the Unskilled Are Unaware: Further Explorations of (Absent) Self-Insight Among the Incompetent

Joyce Ehrlinger[1], **Kerri Johnson**[2], **Matthew Banner**[3], **David Dunning**[3], and **Justin Kruger**[2]

[1] Florida State University

[2] New York University

[3] Cornell University

Abstract

People are typically overly optimistic when evaluating the quality of their performance on social and intellectual tasks. In particular, poor performers grossly overestimate their performances because their incompetence deprives them of the skills needed to recognize their deficits. Five studies demonstrated that poor performers lack insight into their shortcomings even in real world settings and when given incentives to be accurate. An additional meta-analysis showed that it was lack of insight into their own errors (and not mistaken assessments of their peers) that led to overly optimistic estimates among poor performers. Along the way, these studies ruled out recent alternative accounts that have been proposed to explain why poor performers hold such positive impressions of their performance.

> One of the painful things about our time is that those who feel certainty are stupid, and those with any imagination and understanding are filled with doubt and indecision
>
> - Bertrand Russell (1951)

As Bertrand Russell noted, those most confident in their level of expertise and skill are not necessarily those who should be. Surveys of the psychological literature suggest that perception of skill is often only moderately or modestly correlated with actual level of performance, a pattern found not only in the laboratory but also in the classroom, health clinic, and the workplace (for reviews, see Dunning, 2005; Dunning, Heath, & Suls, 2004; Ehrlinger & Dunning, 2003; Falchikov & Boud, 1989; Harris & Schaubroeck, 1988; Mabe & West, 1982).

Surveys of the literature also suggest that people hold positive beliefs about their competence to a logically impossible degree (for reviews, see Alicke & Govorun, 2005; Dunning, 2005; Dunning, Heath, & Suls, 2004). In one common example of this tendency, several research studies have shown that the average person, when asked, typically claims that he or she is "above average," (Alicke, 1985; Brown, 1986; Dunning, Meyerowitz, & Holzberg, 1989; Weinstein, 1980) which is, of course, statistically impossible. These biased self-evaluations are not only seen in the laboratory, but also arise in important real world settings. In a survey of engineers at one company, for example, 42% thought their work ranked in the top 5% among

---

Correspondence regarding this manuscript should be addressed to Joyce Ehrlinger at the Department of Psychology, Florida State University, Tallahassee, FL 32306-4301. Electronic correspondence can be sent to E-mail: ehrlinger@psy.fsu.edu.

their peers (Zenger, 1992), a fact that could easily impede their motivation to improve. Elderly people tend to believe they are "above average" drivers (Marottoli, & Richardson, 1998), a perception that that is, in reality, associated with being labeled an unsafe driver (Freund, Colgrove, Burke, & McLeod, 2005). Academics are not immune. A survey of college professors revealed that 94% thought they do "above average" work—a figure that defies mathematical plausibility (Cross, 1977).

(Absent) Self-Insight Among the Incompetent

Why are people typically so convinced that they are more capable than they, in fact, are? In recent years, an active and emerging literature has grown to explain errors in self-assessment. One strategy for understanding the sources of error in self-assessment is to identify those individuals who make the most mistaken self-judgments. By examining how these error-prone individuals differ from their more accurate peers, one can identify sources of error in general.

Adopting this approach, Kruger and Dunning (1999) suggested that, across many intellectual and social domains, it is the poorest performers who hold the least accurate assessments of their skill and performances, grossly overestimating how well their performances stack up against those of their peers. For example, students performing in the bottom 25% among their peers on tests of grammar, logical reasoning, and humor tended to think that they are performing above the 60th percentile (Kruger & Dunning, 1999). Further, this pattern has been conceptually replicated among undergraduates completing a classroom exam (Dunning, Johnson, Ehrlinger, & Kruger, 2003), medical students assessing their interviewing skills (Hodges, Regehr, & Martin, 2001) clerks evaluating their performance (Edwards, Kellner, Sistrom, & Magyari, 2003), and medical lab technicians evaluating their on-the-job expertise (Haun, Zeringue, Leach, & Foley, 2000).

Kruger and Dunning (1999) argued that this gross overconfidence occurs because those who lack skill simply are not in a position to accurately recognize the magnitude of their deficits. Their incompetence produces a double curse. First, their lack of skill, by definition, makes it difficult to produce correct responses and, thus, they make many mistakes. Second, this very same lack of skill also deprives them of success at the metacognitive task of *recognizing* when a particular decision is a correct or an incorrect one. For example, to produce a grammatically correct sentence, one must know something about the rules of grammar. But one must also have an adequate knowledge of the rules of grammar in order to *recognize* when a sentence is grammatically correct, whether written my one's self or by another person. Thus, those who lack grammatical expertise are not in a position to accurately judge the quality of their attempts or the attempts of other people. In addition, because people tend to choose the responses they think are most reasonable, people with deficits are likely to believe they are doing quite well even when they are, in reality, doing quite poorly (Kruger & Dunning, 1999). Consistent with this argument, poor performers are significantly worse at distinguishing between correct and incorrect responses than are their more competent peers (for a review, see Dunning, 2005). This is true when judging their own responses (e.g., Chi, Glaser, & Rees, 1982; Keren, 1987; Kruger & Dunning, 1999; Maki, Jonas, & Kallod, 1994; Shaughnessy, 1979; Sinkavich, 1995) as well as those of others (Carney & Harrigan, 2003; Kruger & Dunning, 1999).

To be sure, the incompetent are not alone in their difficulty with accurate self-assessment. These same studies suggest that top performers consistently *underestimate* how superior or distinctive their performances are relative to their peers. In Kruger and Dunning's (1999) studies, the top 25% tended to think that their skills lay in the 70th to 75th percentile, although their performances fell roughly in the 87th percentile. Kruger and Dunning suggested that this underestimation stems from a different source — because top performers find the tests they confront to be easy, they mistakenly assume that their peers find the tests to be equally easy.

As such, their own performances seem unexceptional. Supporting this proposal, Kruger and Dunning found that exposing top performers to how their peers performed on the same task caused them to recognize, in part, just how exceptional their performances were relative to their peers (see Hodges et al., 2001, for similar findings).

Goals of the Present Research

The primary aim of this manuscript is to advance an understanding of why the incompetent, in particular, tend to lack self-insight. Although a growing body of evidence has provided support for the claim that incompetence hinders self-insight (e.g., Dunning et al., 2003; Haun et al., 2000; Hodges et al., 2001), this analysis has been subject to criticism. These critiques argue that the self-assessment errors observed by Kruger and Dunning can be largely reduced to statistical or methodological artifacts rather than to an absence of metacognitive competence among poor performers. Through the present research, we sought to examine whether Kruger and Dunning's (1999) analysis or these competing alternatives better explains overestimation among the bottom performers and underestimation among top performers.

Metacognitive Error or Statistical Artifact?

Two alternative accounts have been published to explain the pattern of over- and underestimation of performance observed by Kruger and Dunning (1999). Central to these critiques is the notion that top and bottom performers actually do not differ in their ability to evaluate the quality of their own performances. Instead, it is argued, people of all skill levels have equal difficulty estimating the quality of their performance — and it is this shared difficulty, coupled with statistical or methodological artifacts, that account for the observed patterns of over- and underestimation.

Regression to the mean account

In one critique, Krueger and Mueller (2002; see also Ackerman, Beier, & Bowen, 2002; Krueger & Funder, 2004) proposed that the patterns of over- and underestimation found by Kruger and Dunning (1999) were not evidence of a relationship between skill level and metacognitive skill. Instead, they argued, the pattern was produced by regression to the mean coupled with the fact that people tend overall to rate themselves as above average (Alicke, 1985; Brown 1986; Krueger, 1998). Because perceptions of performance correlate imperfectly with actual performance, it was nearly inevitable that the self-assessments of bottom performers would regress back to toward an average self-assessment, thus ensuring that their estimates would be higher than the performance they achieved. Similarly, the performance estimates made by top performers would regress back toward the average, ensuring that their true performance would be higher than their corresponding estimates.

According to Krueger and Mueller (2002) this regression to the mean phenomenon arises, in part, because measures used to assess the skill level of participants are statistically unreliable and, thus, fraught with measurement error. This unreliability would ensure a smaller correlation between perceptions and the reality of performance, leading to more regression to the mean and greater levels of over- and underestimation. If one measured and then corrected for that lack of reliability, they argued, a good deal of over- and underestimation would evaporate. In two datasets, Krueger and Mueller did just that — demonstrating that a good deal of the overestimation among bottom performers and underestimation among top performers in their datasets evaporated after unreliability, and thus the impact of measurement error, were corrected (for a reply, see Kruger & Dunning, 2002).

Task difficulty account

Burson, Larrick, and Klayman (2006) provided a similar account for Kruger and Dunning's interpretation of over and underestimation, this focused on comparative performance estimates in which people assessed how well they performed relative to their peers. They agreed with Krueger and Mueller (2002) that comparative judgments of one' performance are difficult for everyone to provide and that bottom and top performers do not differ in their ability to evaluate their performance. Their argument, drawing upon Kruger (1999), noted that above average effect frequently occurs for tasks that people perceive to be easy but that tasks perceived to be difficult can produce *below* average effects. When faced with great difficulty in completing a task, individuals believe that they are performing poorly and, failing to properly account for the degree to which others also experience this difficulty, assess their relative performance as *worse* than average. Burson and colleagues argued that, if everyone produces similar estimates (estimates that are high for tasks perceived to be easy but low for tasks perceived to be difficult) what dictates accuracy is less a matter of greater insight on the part of some participants and more a matter of perceived difficulty. When a test seems easy, everyone will believe they have performed well relative to their peers but only top performers will be accurate, leaving bottom performers overconfident. When the test is construed to be hard, however, everyone will think they have done poorly relative to the peers and bottom performers will be more accurate than their more competent peers. In short, whether top or bottom performers are most inaccurate is an artifact of the perceived difficulty of the task.

Indeed, Burson and colleagues (2006) presented participants with tasks perceived to be difficult in three studies and found support for their assertions. Participants estimated how well they had performed on tasks (e.g., trivia and the "word prospector" tasks) that were designed to appear either particularly easy or difficult. Across these studies, Burson and colleagues found that estimates of performance did not correlate well with actual performance but correlated highly with difficulty condition. After completing an easy task, participants of all skill levels estimated that they had performed well relative to their peers, such that top performers looked relatively accurate and bottom performers were grossly overconfident. However, after completing a difficult task, participants of all skill levels estimated that they had performed quite poorly relative to their peers, making poor performers look quite accurate and top performers vastly underconfident.

Although Burson and colleagues largely focused this critique on comparative and not absolute estimates of performance, they took their results as evidence that the Kruger and Dunning (1999) pattern of over- and underestimation of relative performance was simply a function of using seemingly easy tasks and, as such, did not provide evidence of a relationship between skill level and accuracy in self-assessments.

The Present Investigations

The following studies were designed to address the above critiques and, more generally, provide a better understanding of the relationship between level of skill and accuracy in self-assessment. We have organized the studies described in this manuscript in three sections, each with a separate aim.

Section 1

Section 1 was designed to directly address the claims that apparent over and underestimation among bottom and top performers can be reduced to statistical and methodological artifacts. We did this in several ways. To address the claims made by Krueger and Mueller (2002), we explored the accuracy of self-assessments after correcting for lack of statistical reliability in our datasets. Once correcting for unreliability, would we still see dramatic overestimation on

the part of poor performers and underestimation among top performers (as predicted by Kruger and Dunning, 1999) or would this pattern of misestimation largely vanish (as predicted by Kreuger & Mueller, 2002)?

Second, given the Burson et al. (2006) critique, we thought it critical to explore self-assessment squarely in the real world, with tasks of ecological validity that participants approached with an ecologically representative range of competencies. Burson et al. explicitly chose tasks that would appear to be difficult or easy for their participants. As such, they showed what could happen at the extremes of human performance. In addition, they also chose tasks that participants were likely not to have much experience or familiarity with, such as trivia questions or a word game, which meant that participants faced a notable degree of uncertainty about how they or their peers would perform. Those choices left open the question of what patterns of assessments people would make if they dealt with a type of task they were very likely to face in real life—with which they had some familiarity about how they and their peers would performed. A quick look at overall performance levels attained by participants in Kruger and Dunning (1999) suggests that the patterns of over and under-confidence would look quite different from what Burson et al. proposed. According to Burson et al. (2006), poor performers will grossly overestimate their performance on only tasks that are perceived to be easy. The tasks used in Kruger and Dunning, however, look anything but easy. The average performance participants attained ranged from 66.4% correct (Study 3) to 49.1% correct (Study 4). Bottom performers answered between 48.2% (Study 3) and 3.2% (Study 4) questions correctly in these studies. Yet, even facing these difficult tasks, poor performing participants still grossly overestimated their performance relative to their peers.

Thus, in Part 1, we looked at real world cases in which people approached (often challenging) tasks that they would encounter anyway in their everyday lives, rather than ones managed by experimenters to seem either easy or difficult. In these ecologically valid circumstances, would we tend to find the pattern of self-assessments observed by Kruger and Dunning (1999) or would the pattern look different? We took this direction because we thought it would be critical to explore error in self-assessment on ecologically representative and familiar tasks in real-world settings. In particular, we asked undergraduate students to estimate how well they had performed on course exams and asked members of college debate teams to evaluate their tournament performance. These tasks were chosen because they were ones that individuals approached out of their own volition (as opposed to having the task imposed by an experimenter), they were devised by naturally-occurring agents (e.g., a course instructor) rather than by experimenters, and participants had reasonable amount of prior experience and feedback on the tasks.

In addition, Burson et al (2006) inspired us to explore a wider range of self-assessment measures. Their argument about task difficulty rested largely on the use of comparative measures in which people evaluated their performance relative to their peers. They argued that people would underestimate themselves on difficult tasks and overestimate themselves on easy tasks because of the inherent difficulty of knowing how their peers had done, regardless of the skill level exhibited by the person making the evaluation. But what about estimates that do not involve comparisons with peers? Burson et al. are largely silent on this, but Kruger and Dunning (1999) are not: Peer performers should still grossly overestimate their performance regardless of the type of measure used. Thus, in the following studies, we took care to ask participants to judge their performance on “absolute” evaluations — measures that required no comparison to another person (e.g., how many test questions did they answer correctly) — in addition to comparative judgments. We predicted that poor performers would overestimate their performance on absolute as well as relative measures, with top performers being largely accurate in their assessments.

Section 2

In Section 2, we examined a third plausible alternative explanation of the pattern of over- and underestimation observed by Kruger and Dunning (1999). One could argue that a goal to preserve a positive, if not accurate, view of the self may be particularly strong among those who have performed poorly precisely because these are the individuals who might suffer the most from admitting the reality of their poor performance. Those who score very well, in contrast, would have considerably less motivation to glorify the quality of their performance. Indeed, they may be motivated instead to be overly modest about their achievement.

If this is the case, what appears to be an inability to assess the quality of one's performance on the part of the unskilled might actually be an unwillingness to do so accurately, in that the unskilled prefer to report a rosy view of their performance. Under this analysis, those who are unskilled can and will recognize how poorly they have performed if properly motivated. Thus, in the three studies comprising the second section, we offered incentives to encourage participants to provide accurate self-assessments. If the unskilled are truly unable to evaluate the quality of their performances, their performance estimates should remain inflated even in the face of strong incentives to be accurate.

Section 3

The first two sections of this paper speak primarily to factors that *do not* influence performance estimates, while simply referring back to previous literature to clarify what *does* influence estimates. This focus stems directly from the alternative explanations provided in critiques of Kruger and Dunning (1999). In Section 3, however, we provide a meta-analysis of existing data to look directly at the specific errors leading to overestimation of comparative performance among poor performers and underestimation by top performers. According to Kruger and Dunning (1999), poor performers overestimate their abilities because they do not have the metacognitive skill to know that they themselves are doing poorly. The major problem is one of self-estimation, not estimation of peers. Misestimates of their peers' performance may contribute to their bias, but it is not the primary mechanism that leads to their overestimation. Top performers, on the other hand, may underestimate how well they are doing relative to their peers because they overestimate how well their peers are doing. That is, mistaken assessments of peers become a more substantive portion of why top performers fail to recognize the rarity of their competence.

Our analysis in Section 3 directly explored the influence of these differing sources of error on self-assessments made by top and bottom performers. In doing so, it served as a response to both the Krueger and Mueller (2002) and Burson et al. (2006) who attributed errors in performance evaluation to methodological or statistical artifacts—to overall bias in self-estimates (e.g., were people largely over- or underestimating their performance relative to their peers) as well as measurement error in the assessment of skill. If we could tie patterns of over- and underestimation more closely to the types of specific errors predicted by Kruger and Dunning (1999), we would then provide evidence in support of (or against) their account.

Section 1: Correcting for Reliability in Self-Assessments for Real World Tasks

All too often, social psychological research remains in the laboratory and we are left to infer that the same phenomenon routinely occur in the real world. For this reason, the discipline is often open to the criticism that what we find is limited to particular contrived laboratory situations or to particular demographics (e.g., Mintz, Redd, & Vedlitz, 2006; Sears, 1986) Thus, those few opportunities to measure social psychological phenomenon in the real world are particularly valuable. Real world demonstrations are particularly important in this case because

critiques of Kruger and Dunning (1999) have centered on whether their findings are limited to particular types of tasks (e.g., easy tasks or tasks with unreliable measures; Burson et al., 2006; Krueger & Mueller, 2002). Thus, in this section, we examined the accuracy of self-assessment among top and bottom performers on real world tasks.

We were concerned not just with the generality of our phenomenon across tasks but also across measures. Thus, throughout this paper, we broadened the types of self-assessment measures examined to include absolute measures of performance as well as comparative ones. We, like Kruger and Dunning (1999), asked participants to rate the quality of their performance relative to their peers. However, in addition to this comparative estimate, we asked participants to provide absolute estimates of their performance, such as the number of exam questions answered correctly (Studies 1, 3, 4, and 5), the number of debate matches won, and how a judge had rated their debate performance (Study 2). Would poor performers dramatically overestimate their performance on absolute as well as relative self-assessments?

In addition, the studies in Section 1 allowed for a more ecologically valid test of the regression-to-the-mean critique made by Krueger and Mueller (2002). According to that critique, once error in the measurement of perceived and actual performance is accounted for, bottom performers should not overestimate their abilities any more than do other individuals. Thus, in both studies, we measured and corrected for any unreliability in our tests. This correction should make estimates of performance more accurate—but by how much? If error in self-assessments stems from a psychological rather than a statistical source, the original pattern of over- and underestimation should remain largely intact.

Study 1

In Study 1, college students assessed their performance on a challenging in-class exam immediately after completing it. They judged how well they had done relative to other students in the class and also estimated their raw score — the number of questions answered correctly. We predicted that bottom performers would overestimate their performance regardless of type of measure used. Top performers would underestimate their performance relative to their peers, but would show much less, if any, underestimation on the raw score measure.

Study 1 replicates a study in Dunning et al. (2003), showing that students doing badly on a course exam tended to grossly overestimate their performance whether relative (e.g., percentile) or absolute (e.g., raw score) measures were used. This study also adds one important extension. Using students' performances on a second course exam, we could calculate the test-retest reliability of students' performance in the class. This reliability estimate could be used to correct for measurement error. Then, we could see the extent to which the original pattern of results evaporated once controlling for measurement error in this ecologically valid setting. We predicted that much of the original pattern would remain strong even after correcting for measurement error.

Method

Participants—Participants were 124 out of 238 students enrolled in an intermediate-level large-lecture psychology course psychology course. Students received extra credit toward their course grade for participating.

Procedure—Participants responded to a short questionnaire attached to the end of their first preliminary examination in the course. Participants were asked to provide a percentile rating of their mastery of course material, as well as their specific performance on the course examination. Ratings could range from 1 (the student believed they would have the lowest score out of every hundred students) to 99 (the student would have the best score out of every

hundred students. Participants also estimated their raw exam score (out of 40 possible), as well as the average score students in the class would attain.

The exam consisted of 22 multiple-choice questions and 3 essays worth six points each. Participants completed the questionnaire and handed it in before leaving the exam session. On the questionnaire, participants also gave permission for the researchers to later obtain their score on the test. Five weeks later, during the second preliminary exam session in the course, we followed the identical procedure to collect participants' perceptions of their exam performance, as well as the reality of performance.

Results and Discussion

Although the difficulty of this real life course exam was, of course, not manipulated for experimental purposes, we think it important to note that it was, in fact, a challenging exam. On average, students answered 71.2% of questions correctly, with bottom performers failing the exam (55.5% correct) and top performers earning, on average, a B+ (87% correct).

Despite the difficulty of this exam, as predicted, participants overestimated their performance and their ability level (relative to their performance). Participants thought that their mastery of the course material lay in the 71st percentile, when their performance actually placed them in the 49th, $t(120) = 8.74$, $p < .0001$. Similarly, participants thought that their test performance placed them in the 68th percentile—again, an overestimate, $t(120) = 8.12$, $p < .0001$. Not surprisingly, participants also tended to overestimate their raw score on the test by roughly 3.5 points (estimated score = 32.0; actual score = 28.5), $t(120) = 7.88$, $p < .0001$. Students who completed the survey did not differ significantly in terms of average performance or variance from students who opted not to complete the survey.

We expected that the degree to which participants provided overconfident estimates would depend upon their actual level of performance. To determine whether this was the case in this and all subsequent studies in this paper, we followed the practice outlined in Kruger and Dunning (1999) and split our participants into four groups based on their objective performance. Figure 1 shows both the actual performance achieved by students in each performance quartile and also students' perceptions of both their raw score and percentile performance. As displayed by the difference between perceived and actual performance in Figure 1, those performing in the bottom 25% of the class ($n = 33$) dramatically overestimated their performance. They thought their mastery of course material lay in the 63rd percentile and their test performance at the 61st, even though their objective performance placed them in the 15th percentile, ts > 17, $p < .0001$. Additionally, they thought, on average, that they had posted a raw score of 30.4 on the test, when in fact their average score hovered around 22.2, $t(32) = 11.28$, $p < .0001$.

Top performers—students in the top 25% ($n = 27$)—estimated their performance much more accurately, albeit not perfectly. First, they underestimated the distinctiveness of their mastery and performance relative to their peers. They thought that their mastery of course material lay in roughly the 74th percentile and their test performance in the 73rd, when in fact it lay in the 87th, ts > 4.5, ps < .0001. They also slightly underestimated their raw score, thinking on average that they scored 32.9 points on the test when they in reality had scored 34.8, $t(26) = -2.83$, $p < .01$. Table 1 shows the difference between students' estimates of their scores (both the raw score and percentile) and the score they actually achieved, split by level of performance.

Correcting for Measurement Error—How much of this over- and underestimation was due to lack of reliability, and thus measurement error? To estimate this, we used participants' scores on the second preliminary examination to provide an estimate of test-retest reliability. In terms of percentile rankings, the ranks participants achieved on this first exam correlated .

52 with the ranks they obtained on the second exam. In terms of raw score, performance on the first exam correlated .50 with performance on the second.

Using these reliability estimates, we then recalculated what the regression slope would be if we assumed perfect reliability. The classic formula (Bollen, 1989) for that calculation is:

$$B_{corrected} = B_{observed} / \text{reliability estimate}$$

This typically results in a steeper regression slope than that originally observed. Correcting for the reliability associated with the dependent measure (in this case, participants' performance estimates) does not alter this relationship or enter into the correction of the regression slope (Bollen, 1989). This altered regression slope correction, however, also calls for a revision of the intercept associated with the relevant regression equation. Because any regression slope must pass through the point representing the mean of both the independent and dependent variables (i.e., objective performance, estimated performance, respectively), the corrected intercept can be calculated as:

$$\text{intercept}_{corrected} = \text{average performance estimate} - B_{corrected} \times \text{average objective performance}$$

Figure 2 depicts the results of an analysis in which perceived performance (including, separately, perceived mastery of course material, percentile score and raw score) is regressed on objective performance. It also depicts what the regression analysis looks like after assuming perfect reliability. As seen in the figure, across three different measures of perceived performance, the relationship between perceived and actual performance was stronger once unreliability was corrected for, but this strengthening was minimal. For example, in terms of test performance relative to other students, participants in the bottom quartile typically overestimated their percentile rank by 49 percentile points. After correcting for unreliability, their overestimates are reduced by only roughly 5 points. In terms of raw score, bottom performers overestimated their score by 8.4 points before correction; 7.2 points after. However, as Figure 2 also shows, a good portion of the misestimates among top performers were eliminated when we corrected for unreliability. For example, concerning estimates of raw score, top performers underestimated their score by 1.7 points before correction; but only .2 points afterward.

Summary—In sum, Study 1 replicated many of the findings of Kruger and Dunning (1999), showing that poor performers grossly overestimated their performances in an ecologically valid setting. This was not only true when relative measures were used in which participants gauged their performances against those of their peers, but it was also true on an absolute measure (i.e., estimates of raw score). In addition, this overestimation was found with respect to an ecologically valid task of some importance and familiarity to participants, in which a natural range of competence and incompetence was observed. Also replicating Kruger and Dunning (1999), top performers underestimated their percentile rankings relative to their peers. They also slightly underestimated their absolute performance, but the magnitude of that underestimation did not approach the misjudgments seen among their bottom performing peers. Finally, correcting for measurement error, or rather the unreliability of performance measures, only minimally attenuated the patterns of misestimation. Indeed, even after assuming perfect reliability, bottom performers still dramatically overestimated how well they had done.

Study 2

Study 2 provided a conceptual replication of Study 1. A few times each year, Cornell University holds debate tournaments in which teams from the Northeastern United States and Eastern Canada converge to compete. During preliminary rounds of the tournament, teams compete with each other but are not given any feedback about how well they are doing. We took advantage of this situation to ask participants to estimate how well they were doing, and then compared their perceptions against the reality of their performance.

Thus, this was the perfect forum to gauge whether debater's perceptions of competence match those observed by Kruger and Dunning (1999) in another ecologically valid situation. This situation had an additional advantage in that we did not create the performance metric nor were we involved in the evaluation of participants.

Finally, we corrected for unreliability in one performance estimate, as in Study 1, to see to the degree to which psychological, rather than statistical errors explain a lack of insight among the incompetent.

Method

Participants—Participants came from 54 2-person teams who had convened for a regional debate tournament at Cornell University. Of those, 58 provided an adequate number of self-evaluations (see below) to qualify to be in our sample.

Procedure—The study was conducted during the preliminary rounds of the debate tournament. In each of six preliminary rounds, each team debates another team on a topic chosen by tournament organizers. The two teams are observed by a judge, who determines which team wins and rank orders the 4 individuals in terms of the quality of their performance. The judge also scores the 4 after each round on a 30-point scale running from 1 to 30, although in practice virtually all ratings fall between 22 and 27. Thus, for each participant on each round, we had three measures of objective performance: whether the participants' team had won, his or her rank, and the score the judge had given.

After each round, teams are re-paired. Teams that win on the first round are paired with other winners; losers on each round are paired with losers. Importantly, no participant is told how well he or she, or the team, did until all six preliminary rounds are over.

We asked tournament participants to estimate how well they thought they had done. Specifically, they predicted whether their team had won, what rank they had personally attained, as well as what score the judge had given them. Participants filled out these estimates and identified themselves on short questionnaires after each round, which they then deposited in a box. Of 108 total tournament participants, 58 rated their performance on at least 3 of the 6 rounds, and so were included in our final sample.

After the preliminary rounds are over, how each individual and team did was made publicly available. We examined and transcribed these records to obtain our measures of objective performance. Data from 4 individuals were lost. Thus, we had objective performance measures for 104 debate tournament participants.

Results and Discussion

To make the ranking measure commensurate with the two other measures of objective performance, in which higher scores meant better performance, we reverse-scored the ranking measure so that 4 meant that the individual was rated best in group and 1 as worst in group.

Looking over all measures, we find that tournament participants overrated their performance. On average, they thought they won 75.4% of their matches, whereas they actually won only 46.7% of them, $t(57) = 8.85$, $p < .0001$. They thought they achieved a rank of 2.8, whereas they actually achieved only 2.4, $t(57) = 5.09$, $p < .0001$. They thought on average that judges would give them a score of 25.6, whereas the judge gave only a 25.0, $t(57) = 5.74$, $p < .0001$.

As in Study 1, we separated participants into four groups based on their objective performance, using the sum of the scores judges had given them over the six rounds. As seen in Figure 3, participants performing in the bottom 25% grossly overestimated their performance. They thought they had won nearly 59% of their matches, yet they won only 22% of them, $t(17) = 5.09$, $p < .0001$. They thought they achieved a ranking of 2.4 when their actual rank was 1.9, $t(17) = 4.80$, $p < .0001$. They thought judges would score them as a 24.9 when they actually received scores on average of around 23.8, $t(16) = 5.65$, $p < .0001$.

Top performers (those in the top 25%) did not tend to overestimate their performances as much. To be sure, they did overestimate the percentage of matches they won, perceiving that they won 95% when they actually won only 77%, $t(10) = 3.46$, $p < .01$. However, they did not misestimate, on average, what rank they achieved (3.3 and 3.2 for perceived versus actual, respectively), $t(10) = .52$, *ns*, nor the score judge's gave them (26.5 and 26.4 for perceived versus actual, respectively), $t(10) = .51$, *ns*.

Correcting for Measurement Error—In sum, this study replicated much of the pattern found in Kruger and Dunning (1999). Bottom performers significantly overestimated their performances; top performers gave estimates that, although not perfect, were more closely on the mark. However, how much of the erroneous self-assessment by bottom performers was due to measurement error? To assess this, we focused on the scores judges gave to participants. We focused on this measure because calculating reliability on the other two measures was inappropriate for two reasons. First, how well a person did on these measures depended on the performances of others as well as their own performance. Second, debate tournament procedures ensured that the reliability of these measures would be low. Because winners in one round would be paired with other winners in the next round, this heightened the probability that a winner in a previous round would lose in the subsequent round. The reverse was true for losers, who were more likely to win in subsequent rounds because they were paired with previous round losers. This procedure ensured low reliability for the matches-won measure. Similarly, this same problem applied to the ranking measure. The scores judges gave participants from round to round, however, were not influenced artifactually by performance in previous rounds.

Based on the performance of the 104 participants for whom we had objective performance data, the internal consistency of participants' performance was .85 on the judge's score measure. Figure 4 displays the results of a regression analysis in which perceived performance was regressed on objective performance. It also displays the results of a regression, following the procedures outlined in Study 1, which corrects for any unreliability on the objective performance measure. As seen in the figure, the relationship between perceived and objective performance was stronger after correcting for unreliability, but only slightly so. Bottom performers, according to this analysis, overestimated their score by 2.0 points; this fell to 1.9 points after correction. Top performers originally did not overestimate their score, but do so by .1 points after correction.

Summary—In sum, Study 2 revealed that poor performers grossly overestimated their performance regardless of the performance metric examined, even after accounting for error stemming from measurement unreliability. Top performers did not show a consistent pattern. They overestimated how many matches they would win, but did provided accurate assessments

on the other two performance metrics examined (e.g., rank of performance; judge's score). Again, in an ecologically valid setting, bottom performers remained strongly overconfident.

Section 2: Incentives for Accuracy

The studies comprising Section 2 were designed, in part, to address another aspect of ecological validity. One could argue that participants are not properly motivated to offer accurate self-assessment. People just want to think good things about themselves while denying bad things—a tendency that has been documented quite strongly across several decades of psychological research (Baumeister & Newman, 1994; Dunning, 2001; Kunda, 1990). Beyond this motive, there is the motive to impress the experimenter (Baumeister, 1982). Thus, people might provide inaccurate—and positive—self-assessments in order to look good in front of the person requesting those assessments.

Although the motive to believe positive things and disbelieve negative things about oneself can be strong, it can also be tempered when other goals become more prominent. The desire to earn money or to appear accurate in front of a respected authority figure might, for example, temper or even trump one's desire to believe that one has performed well. Although monetary incentives do not lead to greater accuracy on all tasks and types of judgments, they do often produce greater accuracy in judgments. Such incentives have been shown to improve performance on a variety of intellectual tasks (Atkinson, 1958; Glucksberg, 1962), probability judgments (Grether, 1980; Wright & Anderson, 1989), judgment (Awasthi & Pratt, 1990), prediction (Ashton, 1990; Hogarth, Gibbs, Mckenzie, & Marquis, 1991) and memory tasks (Kahneman & Peavler, 1969; Libby & Lipe, 1992; Salthouse, Rogan, & Prill, 1984) (for a Jenkins, Mitra, Gupta, & Shaw, 1998). Drawing upon a meta-analysis of 74 studies, Camerer and Hogarth (1999) argued that monetary incentives are effective as a means of reducing self-presentational concerns and encouraging additional effort to determine the correct answer. Thus, we hoped to minimize the impact of the motive to self-enhance by activating the motive to make money.

In the following three studies, we investigated the impact of incentives toward accuracy on the validity of respondents' self-assessments. The first two studies examined the effect of monetary incentives on self-assessment and, in particular, on self-assessment by the unskilled. Although monetary incentives can reduce the effects of motivation on judgment, we expected that they would have no noticeable effect on the degree to which participants accurately estimated the quality of their performance. In particular, we expected that this manipulation would not lead to greater calibration among those who had performed the worst. In the third study, we examined the impact of making people accountable for their self-assessments, in that they would have to justify those self-assessments to another person. Again, we predicted that this incentive toward accuracy would have no impact.

Study 3

Beyond exploring the potential impact of monetary incentives, we also sought to examine self-insight within a new population. It might be argued that researchers sometimes stack the deck, so to speak, towards overestimation of one's performance by focusing primarily on college students' assessments of performance on intellectual tasks. Surely college students may be right to be confident in their intellectual abilities and may not correct for the fact that they are asked to compare themselves relative to other college students rather than to the population at large. Thus, we left the lab and attended a Trap and Skeet competition at a nearby gun club to quiz gun owners on their knowledge of firearms. Competitors in that competition completed a test of their knowledge of gun safety and usage and estimated the quality of their performance, either in the absence or presence of a monetary incentive for accurate estimation.

Method

Participants—A total of 46 participants were recruited at a Trap and Skeet competition in exchange for a payment of $5. Most participants reported owning at least one firearm (96%) and having taken a course in firearm safety (89%). They possessed between 6 and 65 years experience with firearms (mean = 34.5 years).

Materials and Procedure—Contestants in a Trap and Skeet competition were invited to participate in exchange for a $5 payment. They were explicitly told that our interest was in how their level of confidence in each answer on a test of gun knowledge and safety matched, on average, with whether they were correct. We explained that we were not interested in how many questions they actually answered correctly. Those who agreed to participate completed a 10-item multiple-choice test of Gun Safety and Knowledge modeled after one published by the National Rifle Association that individuals are required to pass to receive a license for gun ownership. The test included general gun knowledge questions (e.g. identifying the difference between blank and dummy bullets) as well as questions on proper care and safety (e.g. what to do when a cartridge fails to fire immediately). After choosing the best of 4 possible responses for each question, participants indicated the extent to which they were confident in their response by circling a number on a scale ranging from 25% (simply guessing) to 100% confident (positive that the answer was correct).

Before beginning the exam, participants randomly assigned to the "incentive" condition were told that they had to opportunity to receive $10, doubling their payment, if their ratings of confidence in each response averaged within 5% of their actual score on the test. Control participants received no such incentive. Upon completing the test, all participants were asked to estimate how many of the 10 questions they answered correctly and to estimate their percentile rank. We thought that non-students might be more familiar with ranks because low numbers indicate better performance (e.g., "we're number 1"). Thus, we asked participants to imagine that we had randomly chosen 100 people at this event, including them, and that we rank these individuals according to their score on the test of gun safety and knowledge. We then asked participants to estimate what their rank would be from 1 ("My score would be at the very bottom, worse than the other 99 people") to 100 ("My score would be at the very top, better than the other 99 people"). Finally, participants estimated the number of questions that would be answered correctly, on average, by their peers.

Results and Discussion[1]

As in Study 2, in order to maintain a consistency across experiments, we reverse scored estimates of one's rank so that higher numbers mean better performance.

Accuracy of Self-Assessments—Participants dramatically overestimated the quality of their performance on the test of gun safety and knowledge. They believed, on average, that they had correctly answered 2.06 more questions than they actually did, $t(41) = 7.22$, $p < .001$) and they overestimated the likelihood that individual responses were accurate, on average, by 28%, $t(43) = 11.28$, $p < .001$). Although participants believed that their percentile score was, on average, only 6.8% higher than it actually was — a difference that was not significantly different from zero, $t(38) = 1.00$, *ns*).[2]

We again split participants into four groups based on their objective performance on the quiz. We then examined self-insight as measured by accuracy in (1) estimates of the number of questions answered correctly, (2) estimates of one's percentile score, and (3) the level of confidence one reported for each individual response. Participants in the bottom 25th percentile

[1]Not all participants answered every question such that the degrees of freedom varied across analyses.

were dramatically overconfident on all three measures (see Table 1). They offered overconfident estimates of the number of questions answered correctly, $t(7)= 4.08$, $p < .005$, they were far more likely than their more skilled peers to overestimate their percentile score, $t(7) = 5.14$, $p < .001$ and to be overconfident with respect to individual items, $t(7) = 10.40$, $p < .001$. Note that overconfidence might be particularly worrisome in this case. Certainly we should worry about the poor students whose overconfidence might keep them from studying and improving their level of skill. Note, however, these individuals primarily hurt themselves. This is not necessarily true for individuals who perform poorly on a test of gun safety and knowledge but whom own and use guns often enough to participate in Trap and Skeet competition. Recognizing their misconceptions about how to safely use a gun is critical not just for the individual but also for those within firing range.

Top performers provided self-assessments that lay closer to objective performance. They know about guns and they have greater insight into just how much they know, although they did tend to underestimate their performance, such as total number of items gotten right ($t(11) = -2.38$, $p < .05$), how distinctive their performances were relative to their peers ($t(10) = -2.57$, $p < .05$) and the likelihood that each individual item would be correct ($t(12) = 4.13$, $p < .001$).

The influence of incentives on accuracy—To determine whether the presence of a monetary incentive motivated our participants to provide more accurate self-estimates, we performed multiple regressions predicting self-assessments from condition (the presence or absence of a monetary incentive for accuracy), quartile level of competence, and the interaction between incentive condition and quartile.[3] Although there are three possible measures of overconfidence in this study, we offered a monetary incentive to those whose confidence in individual responses matched their overall rate of accuracy. Thus we focused our analysis of the effect of incentive on that measure. As expected, there was no main effect of incentive condition on the accuracy of confidence ratings, $\beta = .17$, $t(40) = 1.05$, *ns*. Further, we found no evidence that poor performers were particularly overconfident merely because they were not properly motivated to admit to their poor performance. As can be seen in Figure 5, there was a marginally significant interaction between level of competence and incentive condition but in the direction opposite what critics might have suggested. Instead, poor performers became *more* overconfident in the presence of a monetary incentive, $\beta = -.31$, $t(40) = -1.74$, $p < .10$.

Although participants were explicitly told that they would earn money only if they accurately evaluated whether they had answered individual questions correctly, one might wonder whether this incentive influenced the accuracy of other self-assessment measures. We performed two additional multiple regressions to determine whether monetary incentives influenced accuracy in estimates of the number of questions answered correctly and estimates of one's score relative to other participants at the Trap and Skeet competition. As with confidence in individual questions, the opportunity to win money did not lead to more accurate

[2]One should interpret results concerning percentile score in this study with some caution. While one cannot know for sure, this accuracy in relative estimates might be less an indication that participants understood how their performance compared with others and more an indication that participants had difficulty understanding how to estimate their rank. College students might be uniquely familiar with the concept of percentiles because it is such a common measure in academia but rarely used elsewhere. In the present study, 4 participants provided no estimate of their rank and others seemed to confuse rank with percentile For example, three participants estimated that they had answered 9 out of 10 questions correctly but had very different conceptions of their rank relative to other participants. Two estimated that this score would earn them very high ranks (1 and 10) but a third estimated that his or her rank would be 94. We cannot know for sure but it seems plausible that this third person was estimating percentile rather than rank. There is some evidence that participants in this study had difficulty understanding our relative measure in that estimates of absolute and relative performance are correlated highly in the other 3 studies (ranging from .51 to .67, all *p*s < .005) but correlated only .27 ($p = .10$) in the present study. We should note that this issue remains confined to this study as participants in all other studies were college students who, we believe, are more familiar with measures of relative performance.

[3]Following the recommendation of Aiken & West (1991) regarding multiple regression with interaction terms, all independent variables in Studies 3–5 were centered before inclusion in the regressions.

assessments of one's relative score ($\beta = .01$, $t(35) = .14$, *ns*) or the number of questions answered correctly ($\beta = -.22$, $t(38) = 1.44$, *ns*). Further, we found additional support for the somewhat counterintuitive finding that monetary incentives made poor performers *more* overconfident, relative to controls. Regression analyses revealed a significant interaction (see Figure 5) between level of competence and incentive condition on estimates of one's percentile score, $\beta = -.37$, $t(35) = -2.22$, $p < .05$, though not on estimates of the number of questions answered correctly, $\beta = -.21$, $t(38) = -1.35$, *ns*.

Summary—Study 3 showed that Trap and Skeet shooters overestimated their performance on a test of gun knowledge and safety even when offered a monetary incentive to be accurate. Further, this study demonstrated that strong overconfidence among the least skilled is not a phenomenon limited to college students evaluating their performance on intellectual tasks. Instead, individuals reflecting upon a hobby in which they possess considerable experience can have rather less than perfect insight into their level of knowledge regarding a central — and critical — feature of that hobby.

Study 4

In Study 4, a replication and extension of Study 3, we sought to provide a particularly strong test of the hypothesis that error in self-assessment cannot be attributed to a motivation to think well of oneself or to present oneself positively to other people. In this study, we returned to the laboratory and to the domain in which Kruger & Dunning (1999) first demonstrated particularly strong overconfidence among those who lack skill – logical ability. A critic might argue that $5 is not a sufficient incentive to motivate individuals to put aside self-presentational concerns and provide an accurate assessment of one's performance. For this reason, we offered participants up to $100 if they were able to accurately determine how well they had performed on the test. For college students who, stereotypically, are uniquely familiar with Ramen noodles, we presumed that $100 would be a strong incentive. To ensure that participants believed they could actually earn this money, the experimenter showed participants a (small) pile of one hundred dollar bills.

Method

Participants—57 undergraduates participated in exchange for extra credit.

Materials and Procedure—All participants completed a 20 item multiple-choice test of Logical Reasoning Ability. As in the previous study, participants indicated how confident they were in each answer by circling a number on a scale anchored by 20% (simply guessing among the 5 multiple-choice answers) and 100% confident (positive that the answer was correct). After completing the test but before estimating the quality of their performance, participants in the "incentive" condition were told that they would receive $100 if they were exactly correct in their prediction of how many of the 20 logical reasoning questions they had answered correctly. They were told that they would win $30 if they correctly estimated within 5% of their actual score on the test.

Participants then completed a closing questionnaire similar to that in the previous study in which they estimated their percentile score on the exam, how many of the 10 questions they answered correctly and how many of the questions would be answered correctly on average by their peers.

Results and Discussion

Accuracy of Self-Assessments—As in the previous study, we compared participants' confidence to their test performances. Participants estimated that they had correctly answered 1.42 test questions more than they actually did, $t(56) = 3.11$, $p < .005$. On average, participants

believed their score to be 15.11 percentile points higher than was the case, $t(56)= 3.92$, $p < .001$. Participants in the bottom 25% provided the most overconfident estimates of the number of questions answered correctly, $t(15) = 4.52$, $p < .001$, and of their percentile score, $t(15) = 9.47$, $p < .001$. Participants in the top quartile were again underconfident regarding their percentile score, $t(8) = -4.6$, $p < .005$ but accurate about the number of questions answered correctly, $t(8) = -.61$, *ns* (see Table 1).

Impact of Monetary Incentives—The primary aim of this study was to offer a very strong incentive to provide accurate estimates. If college students cannot look within themselves and provide an accurate estimate of how well they had performed in order to receive $100, something other than motivation is likely keeping them from accurate self-assessments. To determine whether this was the case, we conducted multiple regressions predicting self-assessments from quartile level of competence, monetary incentive condition, and the interaction between level of competence and incentive condition. As predicted, even offering $100 for accuracy did not lead to more accurate estimates of the number of questions answered correctly ($\beta = -.18$, $t(53) = -1.31$, *ns*) or of one's percentile score ($\beta = -.09$, $t(53) = .50$, *ns*). Indeed, no students were able to accurately tell us exactly how well they had performed and win the $100 prize. Only two students were able to come close enough to win the $25 prize.

But we were interested in whether those who score poorly, in particular, were influenced by the offer of $100 to be accurate. It is this group, after all, who might be particularly motivated to claim that they have performed better than they actually have. However, the incentive did not differentially affect those performing in the bottom quartile compared to the rest of the sample (see Figure 6). There was no significant interaction between quartile level of competence and the incentive condition for estimates of the number of questions answered correctly ($\beta = -.14$, $t(53) = -.14$, *ns*) or for estimates of one's score relative to others ($\beta = -.04$, $t(53) = -.283$, *ns*).

Summary—Study 4 provided further evidence that overconfidence among poor performers is not characterized by an ability to be accurate superceded by the desire to save face or preserve self esteem. The studies in this section suggest that overconfidence is not a result of simply deciding to believe that one has performed well. In the absence of an incentive, it possible that participants are inaccurate because they have not thought carefully about the question or do not wish to admit how poorly they performed. Were this the case, however, surely they would be more accurate when given the opportunity to win $100. Instead, this incentive did not lead to accurate self-assessments even among poor performers who might have been most motivated to exaggerate the quality of their performance and who provide self-assessments that are the least accurate.

Study 5

We explored the effects of a different type of incentive in order to be sure that overconfidence and, in particular, overconfidence among the unskilled, did not stem from inadequate motivation. In Study 5, we moved from monetary incentives to social ones. People reach more considered and careful judgments when they must justify those judgments to other people—that is, when they are held *accountable* for their conclusions. This is obvious in its most extreme examples. Executives who have the success of multi-million dollar deals depending on the accuracy of their judgments are likely to be very careful in way that they might not later that evening, when simply selecting which television show they might most enjoy. Even less extreme forms of accountability, however, have been shown to be very effective in encouraging individuals to increase effort and attention paid to a task and inspiring greater motivation to get the correct answer (Tetlock, 1983; Tetlock & Kim, 1987; Lerner & Tetlock, 1999). In these studies, for example, researchers have elicited greater attention and less overconfident

predictions regarding the personality of other individuals by making participants accountable to the experimenter (Tetlock & Kim, 1987). Similarly, students who anticipate discussing their answers on a general knowledge quiz with a group of their peers show less overconfidence in those answers (Arkes, Christensen, Lai, & Blumer, 1987). In the realm of self-evaluation, recent evidence suggests that making individuals accountable for performance evaluations does indeed lead to less self-enhancing, more accurate estimates of performance (Sedikides, Herbert, Hardin, & Dardis, 2002).

In this study, we examined whether making individuals accountable for performance assessments made them more accurate. We were particularly interested in whether accountability would influence the confidence estimates of those who are the least skilled, in an attempt to learn whether individuals give over optimistic estimates of performance because they are not sufficiently motivated to know how well they have performed.

We expected to replicate the pattern of confidence shown in past research, whereby those who possess the least logical reasoning skill also make the greater errors in estimation of their performance. We also expected that making individuals accountable for their responses would have minimal, if any, effect on this pattern. We predicted that participants who expected to justify their responses in an interview with the primary researcher would be no more or less accurate when estimating their score than would participants who were not exposed to this measure of accountability.

Method

Participants—Participants were 42 undergraduates who participated in exchange for extra credit in undergraduate psychology courses.

Materials and Procedure—Upon arrival in the lab, participants were told that they would be asked to complete a test of logical reasoning ability and evaluate the quality of their performance. Participants randomly assigned to the "accountable" condition were told that supervising professor would interview each participant for 5–10 minutes regarding the rationale for his or her answers. This interview was described to "accountable" participants within verbal instructions and was also mentioned in the consent form along with a request to give consent for the interview to be audiotaped. Control participants received only information about the test and signed a consent form that made no mention of an interview.

The presence or absence of accountability manipulation described above represents the sole condition difference in the experimental session. Participants completed a test of 10 multiple-choice items taken from a Law School Aptitude test preparation guide (Orton, 1993). They indicated the best response for each question and then indicated their level of confidence in that response by circling a number on a scale ranging from 20% (purely guessing) to 100% confident (positive that they were correct). After completing the test, participants estimated how many of the 10 questions they had answered correctly and also made a percentile estimate of their performance relative to other Cornell students participating in the experiment (between 1 and 100).

Results and Discussion

Participants overestimated their percentile score by 20 percentage points on average, $t(33) = 4.10$, $p < .0005$). We did not find evidence that participants overestimated the number of questions answered correctly overall, $t\ (34) = .05$, *ns*, but we did find this overconfidence among those in the bottom 25^{th} percentile (see Table 1). One-way ANOVAs revealed that those who are least skilled provided the most overconfident estimates both of the number of questions answered correctly, $t(10)= 3.21$, $p < .01$, and their percentile score, $t(10)= 3.21$, $p < .01$.

Participants in the top percentile were underconfident with respect to the number of questions answered correctly, $t(7) = -5.02$, $p < .005$, and how they scored relative to their peers, $t(10) = 3.21$, $p < .01$.

Making Participants Accountable—The primary goal of this study was to determine whether facing the prospect of having to justify one's self-assessments to an authority figure would reduce the draw of self-enhancement motives and result in greater accuracy. In order to address this issue, we performed multiple regressions predicting assessments of one's percentile score and, separately, estimates of one's raw score from actual performance on the exam (split into 4 groups), presence or absence of the accountability manipulation, and the interaction between the accountability manipulation and actual performance. This analysis did reveal a marginally significant effect of the accountability manipulation ($\beta = 1.33$, $t(30) = 1.81$, $p < .10$) on assessment of performance but, as with the monetary incentive, in the direction *opposite* what a critic might argue (see Figure 7). Participants were marginally *more* confident in ratings of the number of questions they answered correctly, $\beta = .38$, $5(31) = 2.49$, $p < .01$, and in estimates of one's percentile score, $\beta = .28$, $t(33) = 1.82$, $p < .10$, when they expected to have to justify those estimates than when they had no such expectation.

Although accountability did not improve the accuracy of performance evaluations overall, one might expect the manipulation to be most effective for those who possess the least skill. It is this group, after all, who might have the greatest motive to self-enhance. There was, however, no support for this conclusion. Our accountability manipulation did not interact with skill level as measured by estimates of the number of questions answered correctly ($\beta = -.19$, $t(33) = -1.23$, *ns*). There was an interaction between the manipulation and level of skill for estimates of one's relative score ($\beta = -.38$, $t(33) = -2.36$, $p < .05$) but not in the direction one might think. Echoing Study 3, if not with intuition, making individual accountable poor performers *more* rather than less overconfident.

Summary—Study 5 provides evidence that making individuals accountable for their self-assessments, a strong social incentive, did not lead to greater accuracy. Coupled with Studies 3 and 4, this study demonstrates how truly difficult it is to determine one well one has performed, particularly for those who lack skill and, as a result, lack the ability to distinguish a strong from a weak performance. Even when participants try very hard to accurately assess the quality of their performances, in the presence of strong social and monetary incentives, they are unable to gain any additional insight into their level of knowledge or skill.

Sections 1 and 2 provide evidence that poor performers fail to recognize the inferior quality of their performance even in naturalistic settings — in the classroom, within a debate, and in the gun club. This pattern of overestimation appears consistently across multiple tasks and subject populations, even after controlling for unreliability in the test used. As such, these studies speak against the alternative accounts proposed by Krueger and Mueller (2002) and Burson et al. (2006).

Section 3: Sources of Inaccuracy in Performance Estimates

In this final section, we turn attention to the psychological mechanisms behind poor performers' inability to recognize the deficient nature of their own performance relative to others and top performers to identify the exceptional nature of their own. When poor performers thought they were outperforming a majority of their peers, why were they so mistaken? When top performers failed to recognize how superior their performance was relative to others, why did they fail to recognize this fact?

Kruger and Dunning (1999) proposed that different errors underlie the mistaken evaluations made by poor and top performers. For bottom performers, the problem is one of metacognition and self-perception. They argued that poor performers provided inaccurate assessments of how well they have performed relative to their peers because their lack of skill leaves them unable to recognize just how poorly *they themselves* are doing. That is, their estimates of how well they are doing along absolute performance measures (i.e., how many items they were getting right on a test) are just too high. Misperceptions about the quality of their own performance lead to mistaken estimates along relative measures (e.g., percentile estimates of performance relative to peers). That is, bottom performers believe they are providing a large number of correct answers, and so think they must be doing well relative to their peers. To be sure, bottom performers may also misjudge how well other people perform on average, but the degree to which they overestimate their own raw score nearly guarantees that estimates of their percentile score will be wrong.

For top performers, the underlying mechanism producing biased relative estimates would be different (Kruger & Dunning, 1999). Top performers possess more accurate impressions of how they themselves are doing on a test—but they misjudge their peers' performances. Because they do well on tests, they presumed that others do well, too. That leads them to be relatively accurate on their estimates along absolute measures of performance (e.g., how many items they answered corrected). However, they overestimate how well their peers do along the same absolute performance measures. As a result, top performers underestimate how well they are doing relative to their peers.

Thus, according to Kruger and Dunning (1999), poor performers provide inaccurate percentile estimates primarily because they are wrong about their own performance; top performers provide inaccurate estimates because they are wrong about other people. This account makes predictions that can be addressed statistically. Different predictions can be made about bottom and top performers based on whether they "became more knowledgeable" (through statistical correction) about their own performance versus that of their peers. Giving bottom performers knowledge of their own performance along absolute dimensions should go a long way toward correcting their mistaken views of how they are doing relative to their peers. Giving top performers similar information about their own performance might lead them to correct their relative estimates (for people are rarely perfect in assessing their performance), but the improvement should not be as dramatic as it is for bottom performers. Correcting misperceptions of peers should lead top performers to more accurate relative estimates because these misperceptions lead them to be underconfident. However, the same knowledge for bottom performers should not have as obvious an impact on their accuracy of their relative judgments.

We tested these predictions in Section 3 by performing a statistical "what if" exercise. In the first part of this exercise, we looked at data we had from several studies in which we had asked participants to estimate both their raw score on a test as well as the raw score of the average person taking the test. For each study, we calculated via regression analysis how much weight participants gave to those two raw score estimates when estimating how well they performed relative to others. Armed with the resulting regression equation, we then began the what-if phase of the analysis, borrowing a technique called *counterfactual regression analysis*. This technique, commonly found in sociological and economic research (Winship & Morgan, 2000), is used to answer such questions as whether a teenager would have gained in IQ had he or she stayed in school one more year (Winship & Korenman, 1997), or whether parochial schools are superior to public ones (Morgan, 2001).

In this analysis, we asked what each participants' percentile ranking would have been had we replaced their raw score estimates with their actual raw score, given the weight participants gave self and other scores when making relative estimates? What if, instead, we replaced their

estimate of the average raw score obtained by their peers with the actual average? By examining the degree to which assessments become more accurate through these statistical corrections, we can learn about the degree to which misperceptions of the self vs. misperceptions of others lead relative estimates astray.

Method

Studies Included in the Meta-Analysis—Studies were included in the meta-analysis if they did not involve any experimental manipulation and if they included the three estimates needed to conduct the analysis. To be included, participants in these studies must have provided percentile estimates of the degree to which participants possessed the relevant ability (e.g., logic) being tested relative to their peers, as well as percentile estimates of their test performance. We also required that they estimated their raw score (the number of questions answered correctly) as well as the raw score of the "average" student in the study. The 4 studies that met these criteria were Kruger & Dunning (1999, Study 2), in which 45 participants completed a 20-item test on logical reasoning ability; Kruger and Dunning (Study 3), in which 84 participants completed a 20-item grammar test; Kruger and Dunning (Study 4), in which 140 participants confronted a 10-item logic test; and Study 1 from this manuscript, in which 122 participants completed a mid-term exam in a large-lecture psychology course. Details about the studies from Kruger and Dunning can be found in the original article.

Results and Discussion

Analyses focused on three separate issues. First, how accurate were estimates of one's own raw score and of the raw score achieved on average by participants? Second, to what degree did people rely upon perceptions of their own and of the average raw score when estimating how they performed, and the degree to which they possessed the relevant skill, compared with other participants. Finally, how accurate would percentile estimates be if participants knew, rather than having to guess, their exact raw score or, separately, exactly how participants scored on average.

Errors in Estimates of Own and Others' Raw Score—As expected, participants in the bottom and top performance quartile in each study were inaccurate both with respect to their own raw score and that of others. Across the four studies, bottom performers thought they had answered roughly 66% of the items correctly, on average, whereas they answered to only 33% of the items correctly, $Z = 8.12$, $p < .0001$ (See Table 2). Top performers also misestimated their performances, but much less dramatically, underestimating their objective performances by roughly 6% across the studies, $Z = -3.34$, $p < .0001$. Perceptions of the average score were also inaccurate, with both top ($Z = 6.33$, $p < .0001$) and bottom performers ($Z = 4.78$, $p < .0001$) overestimating how well the average other would do. Consistent with Kruger and Dunning's (1999) claim that top performers mistakenly believe that others find the test very easy, they overestimated the average score of their peers significantly more than did their bottom performing peers (mean overestimation = 13% and 8%, respectively), $Z = 2.31$, $p < .05$.

Weight Given to Own and Others' Score in Percentile Estimates—In order to determine the degree to which participants attended to perceptions of their own and the average score, we first converted these scores to a common metric. The number of test items differed across studies so each estimate was converted into estimated percent correct. Next, for each study, we performed two multiple regressions in which we predicted participants' percentile estimates of their general level of ability and, separately, their test performance from estimates of their own raw score and of the average score attained by their peers. We noted the unstandardized regression weights for estimates of one's own and of the average score from each regression as well as the relevant constants. As can be seen in Table 3, participants gave

greater weight to their own raw score but also attended to perceptions of the average other's score when estimating both their level of ability (b = .84, Z = 10.21, p < .001 for one's own score and b = −.42, Z = 5.66, p < .001 for the average score) and their test performance (collapsing across studies, b = 1.00, Z = 13.58, p < .001 for one's own score and b = −.56, Z = 7.55, p < .001 for the average score) relative to other participants.

Impact of Correcting for Errors in Own and Other Raw Score Estimates—Our final analysis provides a window into the source of inaccurate beliefs regarding one's performance relative to peers. How would correcting notions of one's own and, separately, of one's peers raw score impact estimates of percentile score? To conduct this analysis, we created a multiple regression equation for each study using the unstandardized regression weights shown in Table 3 and their accompanying constants. Within each study, we then characterized each individual participant (i)'s observed percentile estimate as:

$$\text{Measured Percentile Estimate}_i = \text{Constant} + B_1 \times \text{Estimate of Own Raw Score}_i + B_2 \times \text{Estimate of Average Other Raw Score}_i + \text{Error}_i$$

We have collected participants' estimates of their percentile scores, their own raw scores and their estimate of the average raw score. From the resulting regressions, we have attained B_1 and B_2 (shown in Table 3) and corresponding constants. Thus, the only unknown value is the component of each participants' percentile estimate—namely, the statistical error—not captured by their raw score estimates for the self and the average. We can use the above equation to solve for the error value for each participant. For example, imagine a participant who estimated that she had answered 80% of the questions correctly, that this score placed her in the 75th percentile in terms of test performance and that her peers answered 60% of the questions correctly on average. We could solve for this participant's error term using:

$$\text{error} = 75 - \text{constant} - (B_1 * 80) - (B_2 * 60)$$

(replacing the constant, B_1 and B_2 with the relevant values from the study in which she participated). This will leave us with two equations for each participant, equations that perfectly describe that participant's estimates of percentile scores as a function of that person's estimate of their own score, estimate of the average person's score, and several other known values (the slopes, constant, and error term). By correcting raw score estimates within these equations, we can learn how erroneous perceptions of one's own raw score and of the average raw score lead these percentile estimates astray.

First, we were interested in the degree to which mispredicting one's own raw score contributes to misperceptions of one's performance relative to others. To do this, we simply replaced estimates of one's raw score with the actual raw score in the equations for each person. Imagine that our example participant from the last paragraph actually answered 50% of the questions correctly. We would replace her estimated raw score (80) with this actual score (50) in order to determine how she might have estimated her percentile score, had she known that she had answered only 50% of the questions correctly. Thus, the percentile estimate corrected for her own raw score would be

$$\text{percentile estimate}_{(\text{corrected for self})} = \text{constant} + (B_1 * \underline{\mathbf{50}}) + (B_2 * 60) + \text{error}$$

We followed this procedure for each participant in order to compute percentile estimates of both their level of ability and of their test performance relative to peers they were corrected for misperceptions of their own score.

We then conducted a similar analysis in order to determine how percentile estimates were affected by misperceptions of how participants perform on average. To do this, we returned to the equations for each individual, leaving their estimated raw score alone but replacing their estimates of the raw score achieved on average with the actual value in that study. If the average participant in our sample participant's experiment answered 70% of the questions correctly her correct percentile score would be:

$$\text{percentile estimate}_{(\text{corrected for other})} = \text{constant} + (B_1 * 80) + (B_2 * \underline{\mathbf{70}}) + \text{error}$$

From these computations, we could see how much percentile estimates would be corrected if participants had accurate conceptions of their own raw score and of the raw score of others, independently. The clearest way to explore these corrections is to examine bottom and top performers separately.

Bottom Performers—Table 4 depicts the estimates of how bottom performers would have assessed their percentile score if they had known their own raw scores and, separately, if they had known the actual average score. As seen in the table, on average, bottom performers would have provided much more accurate percentile estimates of ability and performance had they known their actual raw score. In the actual studies, bottom performers tended to rate their ability and performance at roughly the 60th percentile. Correcting these estimates for errors about the self suggests that bottom performers, on average, would have rated their ability at the 37th percentile and their test performance at the 28th had they known their true raw score—estimates that would have been significantly more accurate, both *Z*s > 8. It is worth noting that these corrected estimates are still quite overconfident. Certainly there are factors that introduce error into self-assessments made by poor performers as well as their more skilled peers and that are not accounted for in the present analysis (for review, see Dunning, 2005). Still, simply correcting for the estimates that poor performers make about their own score brings estimates of their percentile score far closer to their reality.

Curiously, correcting for errors about the average other would have made bottom performers *less*, not more, accurate in their percentile estimates. For ability ratings, percentile estimates would have risen approximately 6 percentile points if bottom performers were given this information—a revision going in the wrong direction; for test performance ratings, the increase would have been almost 5 percentile points, both *Z*s > 5. This increased error, it appears, comes from the fact that bottom performers tended to overestimate how well their peers on average performed (as noted above), although not as much as did their top performing counterparts. Correcting this overestimation in estimates of others had the effect of exacerbating overestimation of how well the self did relative to others.

Top Performers—Table 5 depicts the results of a similar counterfactual regression analysis for top performers. In contrast to the results for bottom performers, top performers misestimate their percentile score, in part, because they overestimate how well others have performed. Across the 4 studies, top performers tended to rate both their ability and their performance in roughly the 74th percentile when their performance was actually in the 88th, on average. Correcting for errors in self-estimates reduced this discrepancy by approximately 4 percentile points for ability and over 5 points for test performance, *Z*s > 3. However, correcting for erroneous impressions of the average persons' performance would also have produced significant reductions in error. Percentile estimates of ability would have risen by over 5 percentile points; estimates of test performance would have increased by nearly 7 points, *Z*s > 6.

Summary—In sum, a comparison of bottom and top performers via a counterfactual regression analysis suggested that a different pattern of errors underlie the misestimates that each provides when estimating their relative performance. For bottom performers, correcting for errors in their estimates of their own raw score would bring them much closer to an accurate impression of their actual percentile ranking. Correcting for errors about others would make their percentile estimates less accurate, not more. Thus, poor performers are overconfident in estimates of how well they performed relative to others because they have little insight into the quality of their own performance. Their estimates are flawed because of misconceptions about their own performance, rather than misconceptions about the performances of others.

For top performers, the pattern is different. Counterfactual regression analyses suggest that top performers' mistakenly modest relative estimates were produced by erroneous impressions of both their own objective performance and that of their peers. Correcting for either of these misconceptions resulted in more accurate percentile estimates. It is not surprising that there would be at least some error in even top performer's perceptions of their own score and that this error would predict error in relative estimates. However, more interesting, is that top performers offer particularly overoptimistic estimates of their peers' objective performance on the test and that this overoptimism produces undue modesty in their relative estimates.

General Discussion

As Bertrand Russell so sagely noted in the quotation that opens this manuscript, the confidence people hold is often not matched by their actual achievement. Understanding why confidence and competence so rarely match has been an enduring interest in cognitive, social, personality, organizational, and clinical psychology (for reviews, see Dunning, 2005; Dunning, Heath, & Suls, 2004; Lichtenstein, Fischhoff, & Phillips, 1982).

In this manuscript, we examined the relationship between self-insight and level of competence. In all, we considered three explanations for the dramatic overconfidence seen among the unskilled — (a) that it is merely a statistical or methodological artifact, (b) that it stems from insufficient motivation to be accurate and (c) that it stems from a true inability to distinguish weak from strong performance. The studies described herein are most consistent with Kruger and Dunning's (1999) explanation, that a lack of skill leaves individuals both performing poorly and unable to recognize their poor performances.

We found that overestimation among poor performers emerged across a variety of tasks and in real world settings (Section 1). In Study 1, for example, students performing poorly on a class exam reported that they had outperformed a majority of their peers. In Study 2, poor performers in a college debate tournament overestimated the number of matches they were winning by 167%. We also found that overestimation among the unskilled did not depend on the measure used. Burson et al. (2006) argued that what appeared to be lesser self-insight among the incompetent had more to do with task difficulty coupled with the use of relative measures. If, instead, Kruger and Dunning (1999) were right that a lack of skill creates an inability to evaluate one's performance, poor performers should make erroneous estimates of absolute as well as relative performance and on difficult as well as easy tasks. We found support for these conclusions. Within the 5 studies in this manuscript, poor performers offered overconfident assessments of their absolute performance (e.g., raw score on test; judge's rating on debate performance) as well as ones of relative performance on a range of challenging real world tasks.

Further, this pattern of overestimation cannot be attributed to a mere statistical artifact, as suggested by Krueger & Mueller (2002), based on notions of statistical reliability and measurement error. We estimated the level of reliability our performance measures possessed

in two different ways (test-retest reliability in Study 1 and internal consistency in Study 2). After correcting for imperfections in reliability in both studies, we found that the magnitude of misestimates by bottom and top performance were reduced slightly, but that the misestimates that each group provided were still left largely intact.

We also provided evidence against the possibility that overestimation among poor performers is a product of insufficient motivation to provide accurate assessments. Poor performers overestimate their performances even when given strong incentives for accuracy (Section 2). In Study 3, giving gun owners a $5 incentive to accurately judge how well they completed an exam on firearm use and safety did not improve even the dramatically overconfident estimates made by poor performers. Offering up to $100 in Study 4 for an accurate estimate of performance on a logic test did not to prompt participants, including poor performers, toward accuracy. In Study 5, we replaced a monetary incentive with a social one—having to justify one's self-assessment to another person—and, again, found no improvement in the accuracy with which people judged their performances.

Along the way, the studies comprising Sections 1 and 2 also replicated a pattern of *under*estimation among top performers. Whenever top performers were asked to rate their skill and accomplishment along a percentile scale—thus comparing their performance in general against those of their peers—top performers tended to underestimate their performances (Studies 1, 3, 4, and 5). However, when other measures of assessment were examined, the picture proved to be more mixed. In Study 1, top performers slightly underestimated their raw score on an exam, a pattern replicated in Studies 3, 4, and 5. But in Study 2, top performers in a debate tournament did not consistently underrate their performances. This lack of consistency across measures, we argue, is tied to the type of measure used, and is consistent with the analysis of Kruger and Dunning (1999). When assessing the quality of their performance in comparison to peers, top performers should underestimate their performances because they overestimate how well their peers are doing (Hodges et al., 2001; Kruger and Dunning, 1999, Study 3). However, on absolute measures that merely require an assessment of self-performance, top performers should be largely accurate and show no directional bias toward over- or underestimation. The mixed results we obtained on absolute measures across the studies are consistent with this inference.

In Section 3, we directly examined Kruger and Dunning's (1999) claim that error in relative estimates made by the unskilled stems from an inability to evaluate the quality of their own performance while error top performers' estimates is also related to a misperception of how well others perform. Counterfactual regression analyses (e.g., Winship & Morgan, 1999) revealed that poor performers would have provided much less optimistic, more accurate percentile scores had known just how low their raw scores had been. Correcting for their misperceptions of how others performed, however, did not improve accuracy in their self-assessments. These analyses suggested that estimates made by top performers, in contrast, are led astray by misperceptions of others. Correcting their slight underestimates of their raw score performance would, of course, have led to more accurate estimates of relative performance. However, correcting misperceptions regarding their peers' performance would have produced an equally large improvement in accuracy.

Concluding Remarks

Taken together, these findings reaffirm the notion that poor performers show little insight into the depth of their deficiencies relative to their peers. They tend to think they are doing just fine relative to their peers when, in fact, they are at the bottom of the performance distribution. By now, this phenomenon has been demonstrated even for everyday tasks, about which individuals have likely received substantial feedback regarding their level of knowledge and skill. College

students have, through out their education, received feedback on their grammatical and logical skill, the domains in which poor metacognitive ability among the unskilled was first demonstrated (Kruger & Dunning's, 1999). Similarly, medical lab technicians do not recognize when they are performing poorly on a test of the skills they use in the lab every day (Haun et al., 2000). In this manuscript, we asked college students to assess how well they had done on a course exam and experienced debaters whether they were winning matches. Yet, in each of these familiar circumstances, poor performing participants did not seem to know how poorly they were doing.

Part of why the dramatic overestimation demonstrated by poor performers is so fascinating is precisely because they show dramatic overconfidence on tasks about which they have likely received substantial feedback in the past. While this issue is beyond the scope of the present manuscript, we remain fascinated by the question of why it is that poor performers do not give accurate performance evaluations on familiar tasks. It seems that poor performers do not learn from feedback suggesting a need to improve. Hacker, Bol, Horgan, & Rankow (2000) provided direct evidence for this failure to learn from feedback when they tracked students during a semester-long class. As time went on, good students became more accurate in predicting how they would do on future exams. The poorest performers did not—showing no recognition, despite clear and repeated feedback, that they were doing badly. As a consequence, they continued to provide overly optimistic predictions about how well they would do in future tests. We hope that future research might shed light on the motivational and cognitive contributors to this failure to update predictions in the light of negative feedback on past performances.

If one cannot rely on life experience to teach people about their deficits, how are people to gain self-insight? While this seems a difficult task, there are clues in the psychological literature that suggest strategies for gaining self-insight. If a lack of skill leads to an inability to evaluate the quality of one's performances, one means of improving metacognitive ability—and thus self-insight — is to improve one's level of skill. Kruger and Dunning (1999) found that training students in logic did, indeed, improve their ability to distinguish correct from incorrect answers and, concurrently, improved the quality of their performances. We might than encourage greater self-insight just by encouraging learning.

Surely we cannot expect individuals to gain some level of competence in all areas just so that they may better understand their strengths and weaknesses. However, it is quite possible to encourage a mindset that leads to greater excitement about learning and, by extension greater self-insight. Dweck and colleagues find that encouraging beliefs in the malleability of traits leads to a host of behaviors that might contribute to more accurate perceptions of one's abilities (for review, see Dweck, 1999). This approach might lead to more accurate self-assessment for the same reason that Kruger and Dunning's (1999) training in logic was effective — by improving people's level of skill. School children who are taught that intelligence is malleable get more excited about learning, become more motivated in the classroom and achieve better grades (Blackwell, Trzesniewski, & Dweck, In Press). Thus, teaching individuals that intelligence is malleable might lead to more accurate self-assessments because this measure leads to an improvement of knowledge and skill that, in and of itself, promotes greater self-insight.

In addition, teaching individuals that traits and, in particular, intelligence is malleable also leads to a greater openness to challenging new tasks (Dweck & Leggett, 1988; Hong, Lin, Wan, Dweck, & Chiu, 1999). Experience with a variety of tasks is likely to provide individuals with extensive feedback from which they may garner information about their abilities. Perhaps not surprisingly, then, recent research reveals that individuals who hold a view that intelligence is malleable make far more accurate assessments of the quality of their performance than do those

who believe intelligence to be fixed (Ehrlinger & Dweck, 2007). Often those with a malleable view of intelligence are not at all overconfident on tasks that inspire dramatic overconfidence in those with a fixed view of the trait. Further, teaching individuals about the malleability of intelligence results in less overconfident assessments of performance (Ehrlinger & Dweck, 2007).

Thus, teachers might help students to better identify what are their strengths and where they need to improve just by imparting knowledge and also by teaching an incremental view of intelligence. These lines of research are exciting in that these among the first strategies identified to help individuals gain greater self-insight however it is also time intensive and considerably easier to implement with students than with adults outside of educational contexts. Further research might explore more general means of improving insight into one's level of skill and one's character. These are crucial questions to address in future research.

Acknowledgments

We thank Alba Cabral, Leah Doane, Alex Emmot, Donny Thometz, Kevin Van Aelst, and Nathalie Vizueta for assisting in the collection of data. We would also like to thank members of the Dunning lab and, in particular, Nicholas Epley, for many helpful suggestions. This research was supported financially by National Institute of Mental Health Grant RO1 56072, awarded to Dunning.

References

Ackerman PL, Beier ME, Bowen KR. What we really know about our abilities and our knowledge. Personality and Individual Differences 2002;33:587–605.

Aiken, LS.; West, SG. Multiple regression: Testing and interpreting interactions. Thousand Oaks, CA: Sage; 1991.

Alicke MD. Global self-evaluation as determined by the desirability and controllability of trait adjectives. Journal of Personality and Social Psychology 1985;49:1621–1630.

Alicke, MD.; Govorun, O. The Better-Than-Average Effect. In: Alicke, MD.; Dunning, DA.; Krueger, JI., editors. The Self in Social Judgment. Studies in self and identity. 2005. p. 85-106.

Ames DR, Kammrath LK. Mind-reading and metacognition: Narcissism, not actual competence, predicts self-estimated ability. Journal of Nonverbal Behavior 2004;28:187–209.

Arkes HR, Christensen C, Lai C, Blumer C. Two methods of reducing overconfidence. Organizational Behavior and Human Decision Processes 1987;39:133–144.

Ashton RH. Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback and justification. Journal of Accounting Research 1990;28:148–180.

Atkinson, JW. Towards experimental analysis of human motivation in terms of motivesss, expectancies, and incentives. In: Atkinson, J., editor. Motives in Fantasy, Action and Society. New York: Van Nostrand; 1958.

Awasthi V, Pratt J. The effects of monetary incentives on effort and decision performance: The role of cognitive characteristics. The Accounting Review 1990;65:797–811.

Baumeister RF. A self-presentational view of social phenomena. Psychological Bulletin 1982;91:3–26.

Baumeister RF, Newman LS. Self-regulation of cognitive inference and decision processes. Personality and Social Psychology Bulletin 1994;20:3–19.

Blackwell LA, Trzesniewski KH, Dweck CS. Theories of intelligence and achievement across the junior high school transition: A longitudinal study and an intervention. Child Development 2007;78:246–263. [PubMed: 17328703]

Bollen, KA. Structural equations with latent variables. New York: Wiley & Sons; 1989.

Brown JD. Evaluations of self and others: Self-enhancement biases in social judgments. Social Cognition 1986;4:353–376.

Burson KA, Larrick RP, Klayman J. Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. Journal of Personality and Social Psychology 2006;90:60–77. [PubMed: 16448310]

Camerer C, Hogarth R. The effects of financial incentives in experiments: A review and capital-labor production framework. Journal of Risk and Uncertainty 1999;19:7–42.

Carney DR, Harrigan JA. It takes one to know one: Interpersonal sensitivity is related to accurate assessments of others' interpersonal sensitivity. Emotion 2003;3:194–200. [PubMed: 12899418]

Chi, MTH.; Glaser, R.; Rees, E. Expertise in problem-solving. In: Sternberg, R., editor. Advances in the psychology of human intelligence. Vol. 1. Hillsdale, NJ: Erlbaum; 1982. p. 17-76.

Cross P. Not can but *will* college teaching be improved? New Directions for Higher Education 1977;17:1–15.

DePaulo BM, Charlton K, Cooper H, Lindsay JJ, Muhlenbruck L. The accuracy-confidence correlation in the detection of deception. Personality and Social Psychology Review 1997;1:346–357. [PubMed: 15661668]

Dunning, D. On the motives underlying social cognition. In: Schwarz, N.; Tesser, A., editors. Blackwell handbook of social psychology: Volume 1: Intraindividual processes. New York: Blackwell; 2001. p. 348-374.

Dunning, D. Self-insight: Roadblocks and detours on the path to knowing thyself. New York: Psychology Press; 2005.

Dunning D, Heath D, Suls JM. Flawed self-assessment: Implications for health, education, and the workplace. Psychological Science in the Public Interest 2004;5:69–106.

Dunning D, Johnson K, Ehrlinger J, Kruger J. Why people fail to recognize their own competence. Current Directions in Psychological Science 2003;12:83–87.

Dunning D, Meyerowitz JA, Holzberg AD. Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. Journal of Personality and Social Psychology 1989;57:1082–1090.

Dweck, CS. Self Theories: Their Role in Motivation, Personality, and Development. Philadelphia, PA: Psychology Press/Taylor and Francis; 1999.

Dweck CS, Leggett EL. A social-cognitive approach to motivation and personality. Psychological Review 1988;95:256–273.

Edwards RK, Kellner KR, Sistron CL, Magyari EJ. Medical student self-assessment of performance on an obstetrics and gynecology clerkship. American Journal of Obstetrics and Gynecology 2003;188:1078–1082. [PubMed: 12712114]

Ehrlinger J, Dunning D. How chronic self-views influence (and potentially mislead) estimates of performance. Journal of Personality and Social Psychology 2003;84:5–17. [PubMed: 12518967]

Ehrlinger, J.; Dweck, CS. If I don't see it, it must not exist: How preferential attention allocation contributes to overconfidence. Florida State University; 2007. Manuscript in Preparation

Fagot BI, O'Brien M. Activity level in young children: Cross-age stability, situational influences, correlates with temperament, and the perception of problem behaviors. Merrill Palmer Quarterly 1994;40:378–398.

Falchikov JN, Boud D. Student self-assessment in higher education: A meta-analysis. Review of Educational Research 1989;59:395–430.

Freund B, Colgrove LA, Burke BL, McLeod R. Self-rated driving performance among elderly drivers referred for driving evaluation. Accident: Analysis and Prevention 2005;37:613–618. [PubMed: 15949451]

Glucksberg S. The influence of strength and drive on functional fixedness and perceptual recognition. Journal of Experimental Psychology 1962;63:36–41. [PubMed: 13899303]

Grether DM. Bayes rule as a descriptive model: The representativeness heuristic. Quarterly Journal of Economics 1980;95:537–557.

Hacker DJ, Bol L, Horgan DD, Rakow EA. Test prediction and performance in a classroom context. Journal of Educational Psychology 2000;92:160–170.

Harris MM, Schaubroeck J. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology 1988;41:43–62.

Haun DE, Zeringue A, Leach A, Fole A. Assessing the competence of specimen-processing personnel. Laboratory Medicine 2000;31:633–637.

Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. Academic Medicine 2001;76:S87–S89. [PubMed: 11597883]

Hogarth RM, Gibbs BJ, McKenzie CRM, Marquis MA. Learning from feedback: Exactingness & incentives. Journal of Experimental Psychology: Learning, Memory and Cognition 1991;17:734–752.

Hong Y, Chiu C, Lin DM-S, Wan W, Dweck CS. Implicit theories, attributions, and coping: A meaning system approach. Journal of Personality and Social Psychology 1999;77:588–599.

Jenkins GD Jr, Mitra A, Gupta N, Shaw JD. Are financial incentives related to the performance? A meta-analytic review of empirical research. Journal of Applied Psychology 1998;83:777–787.

Kahneman D, Peavler WS. Incentive effects and papillary changes in association learning. Journal of Experimental Psychology 1969;79:312–318. [PubMed: 5785645]

Keren GB. Facing uncertainty in the game of bridge: A calibration study. Organizational Behavior and Human Decision Processes 1987;139:98–114.

Krueger J. Enhancement bias in descriptions of self and others. Personality and Social Psychology Bulletin 1998;24:505–516.

Krueger JI, Funder DC. Toward a balanced social psychology: Causses, consequences, and cures for the problem-seeking approach to social behavior and cognition. Behavioral and Brain Sciences 2004;27:313–327. [PubMed: 15736870]

Krueger J, Mueller RA. Unskilled, unaware, or both? The contribution of social-perceptual skills and statistical regression to self-enhancement biases. Journal of Personality and Social Psychology 2002;82:180–188. [PubMed: 11831408]

Kruger J, Dunning D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology 1999;77:1121–1134. [PubMed: 10626367]

Kruger J, Dunning D. Unskilled and unaware — but why? A reply to Krueger and Mueller. Journal of Personality and Social Psychology 2002;82:189–192. [PubMed: 11831409]

Kunda Z. The case for motivated reasoning. Psychological Bulletin 1990;108:480–498. [PubMed: 2270237]

Lerner JS, Tetlock PE. Accounting for the effects of accountability. Psychological Bulletin 1999;125:255–275. [PubMed: 10087938]

Libby R, Lipe MG. Incentives, effort and the cognitive processes involved in accounting-related judgments. Journal of Accounting Research 1992;30:249–273.

Lichtenstein, S.; Fischhoff, B.; Phillips, L. Calibration of probabilities: The state of the art to 1980. In: Kahneman, D.; Slovic, P.; Tversky, A., editors. Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press; 1982.

Mabe PA III, West SG. Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology 1982;67:280–296.

Maki RH, Jonas D, Kallod M. The relationship between comprehension and metacomprehension ability. Psychonomic Bulletin and Review 1994;1:126–129.

Marottoli RA, Richardson ED. Confidence in, and self-rating of, driving ability among older drivers. Accident Analysis and Prevention 1998;30:331–336. [PubMed: 9663292]

Mintz A, Redd SB, Vedlitz A. Can we generalize from student experiments to the real world in political science, military affairs, and international relations? Journal of Conflict Resolution 2006;50:757–776.

Morgan SL. Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. Sociology of Education 2001;74:341–374.

Moreland R, Miller J, Laucka F. Academic achievement and self-evaluations of academic performance. Journal of Educational Psychology 1981;73:335–344.

Orton, PZ. Cliffs Law School Admission Test Preparation Guide. Lincoln, NE: Cliffs Notes Incorporated; 1993.

Riggio RE, Widaman KF, Friedman HS. Actual and perceived emotional sending and personalit y correlates. Journal of Nonverbal Behavior 1985;9:69–83.

Salthouse TA, Rogan JD, Prill KA. Division of attention: Age differences on a visually presented memory task. Memory and cognition 1984;12:613–620.

Sears DO. College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. Journal of Personality and Social Psychology 1986;51:515–530.

Sedikides C, Berbst KC, Hardin DP, Dardis GJ. Accountability as a deterrent to self-enhancement: The search for mechanisms. Journal of Personality and Social Psychology 2002;83:592–605. [PubMed: 12219856]

Shaughnessy JJ. Confidence judgment accuracy as a predictor of test performance. Journal of Research in Personality 1979;13:505–514.

Sinkavich FJ. Performance and metamemory: Do students know what they don't know? Instructional Psychology 1995;22:77–87.

Tetlock PE. Accountability and complexity of thought. Journal of Personality & Social Psychology 1983;45:74–83.

Tetlock PE, Kim JI. Accountability and judgment processes in a personality prediction task. Journal of Personality & Social Psychology 1987;52:700–709. [PubMed: 3572733]

Weinstein ND. Unrealistic optimism about future life events. Journal of Personality and Social Psychology 1980;58:806–820.

Winship, C.; Korenman, S. Does staying in school make you smarter? The effect of education on IQ. In: Devlin, B.; Fienberg, SE.; Resnick, D.; Roeder, K., editors. Intelligence and success: Is it all in the genes? Scientists respond to the bell curve. New York: Springer-Verlag; 1997. p. 215-234.

Winship C, Morgan SL. The estimation of causal effects from observational data. Annual Review of Sociology 2000;25:659–706.

Wright WF, Anderson U. Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. Organizational Behavior and Human Decision Processes 1989;44:68–82.

Zenger TR. Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. Administrative Science Quarterly 1992;37:198–219.

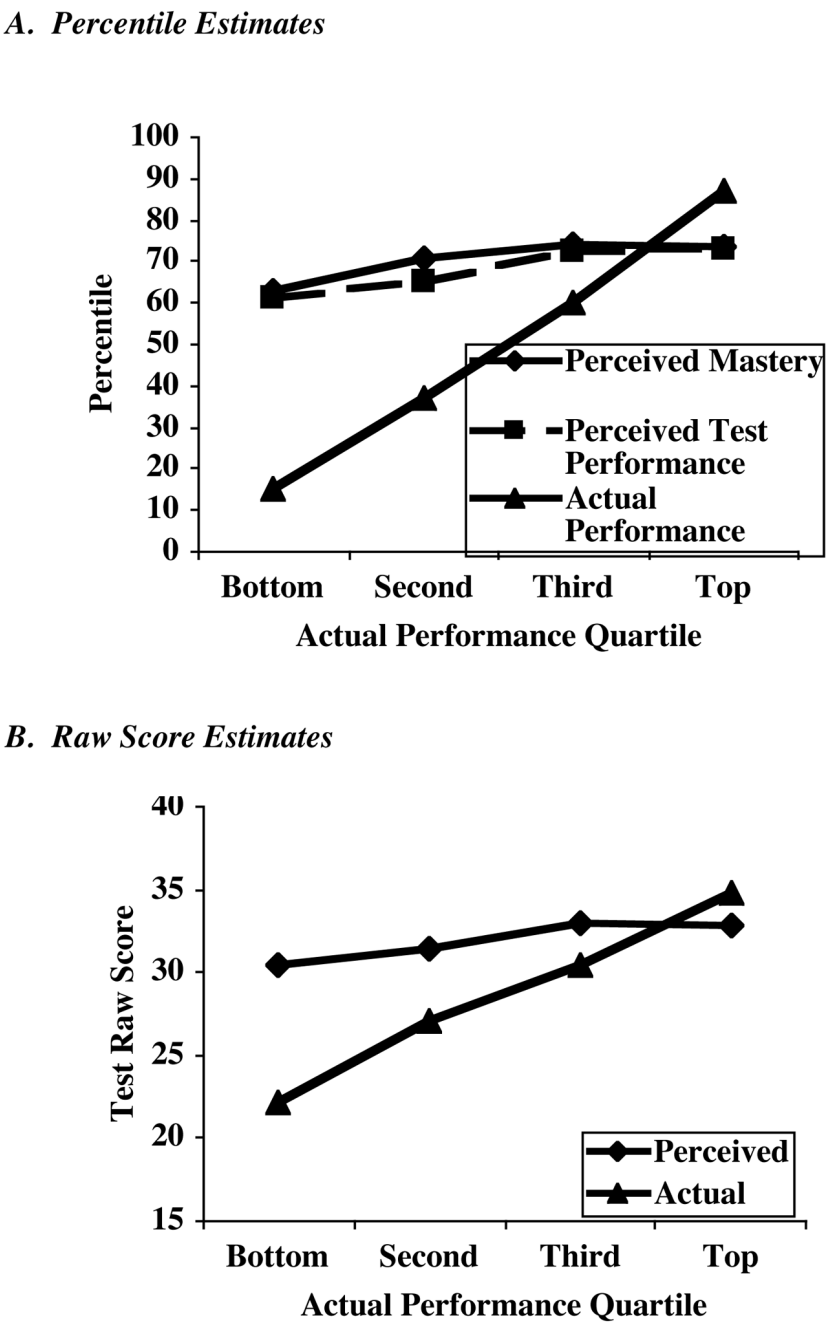
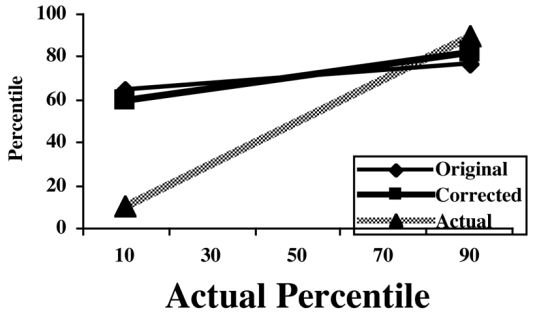
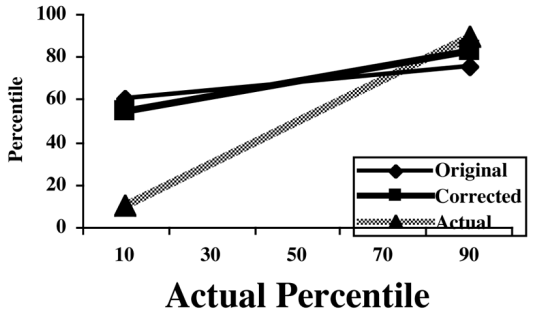


Figure 1.
Students' Estimated and Actual Score on a Psychology Exam, as a Function of Their Actual Exam Performance Quartile, Shown Separately For Estimates of Percentile and Raw Scores (Study 1).

A. Percentile Mastery Estimates



B. Percentile Test Performance Estimates



C. Raw Score Estimates

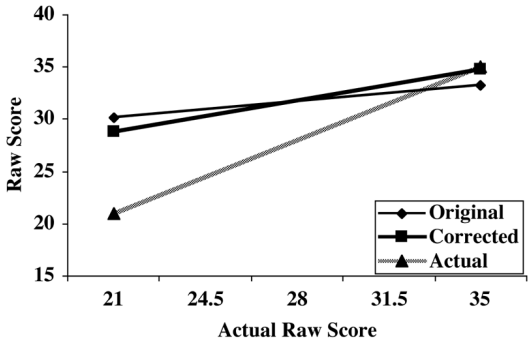


Figure 2.

Comparison of Students' Estimated and Actual Exam Performance Before (Original) and After (Corrected) Correcting for Measurement Unreliability. This Relationship is Displayed Separately For Students' Estimated Mastery of Course Material, Percentile Performance Relative to Other Classmates, and Raw Exam Score. (Study 1).

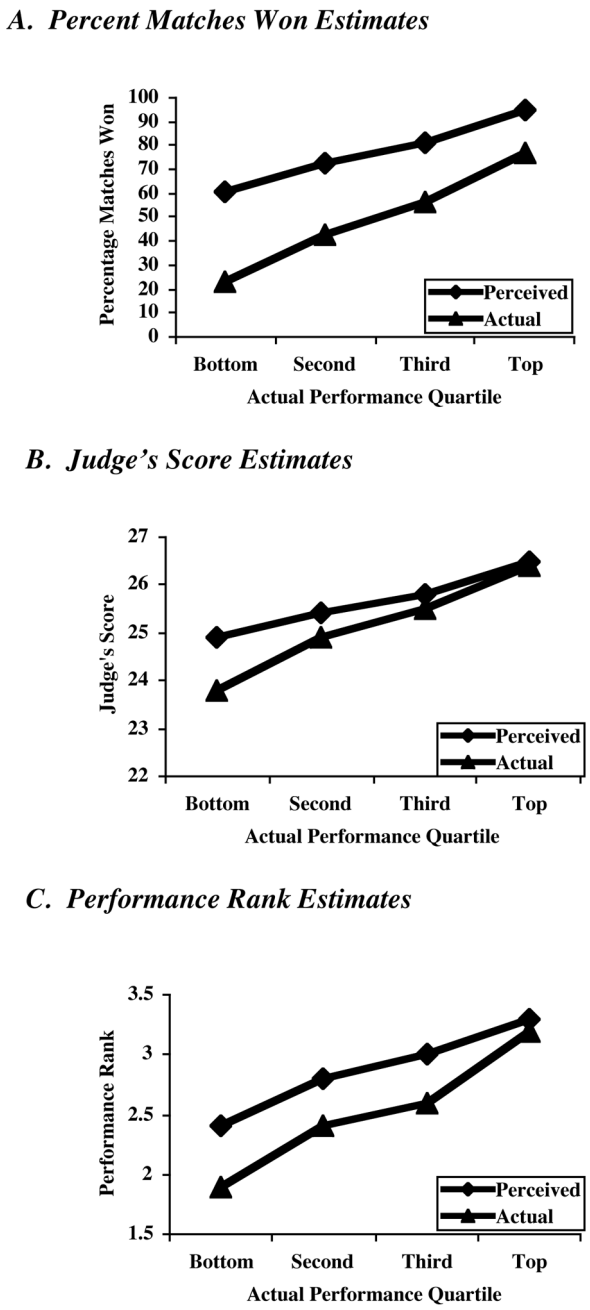


Figure 3.
Perceived Versus Actual Percentage Matches Won, Judge's Scores, and Performance Rank as a Function of Actual Performance (Study 2).

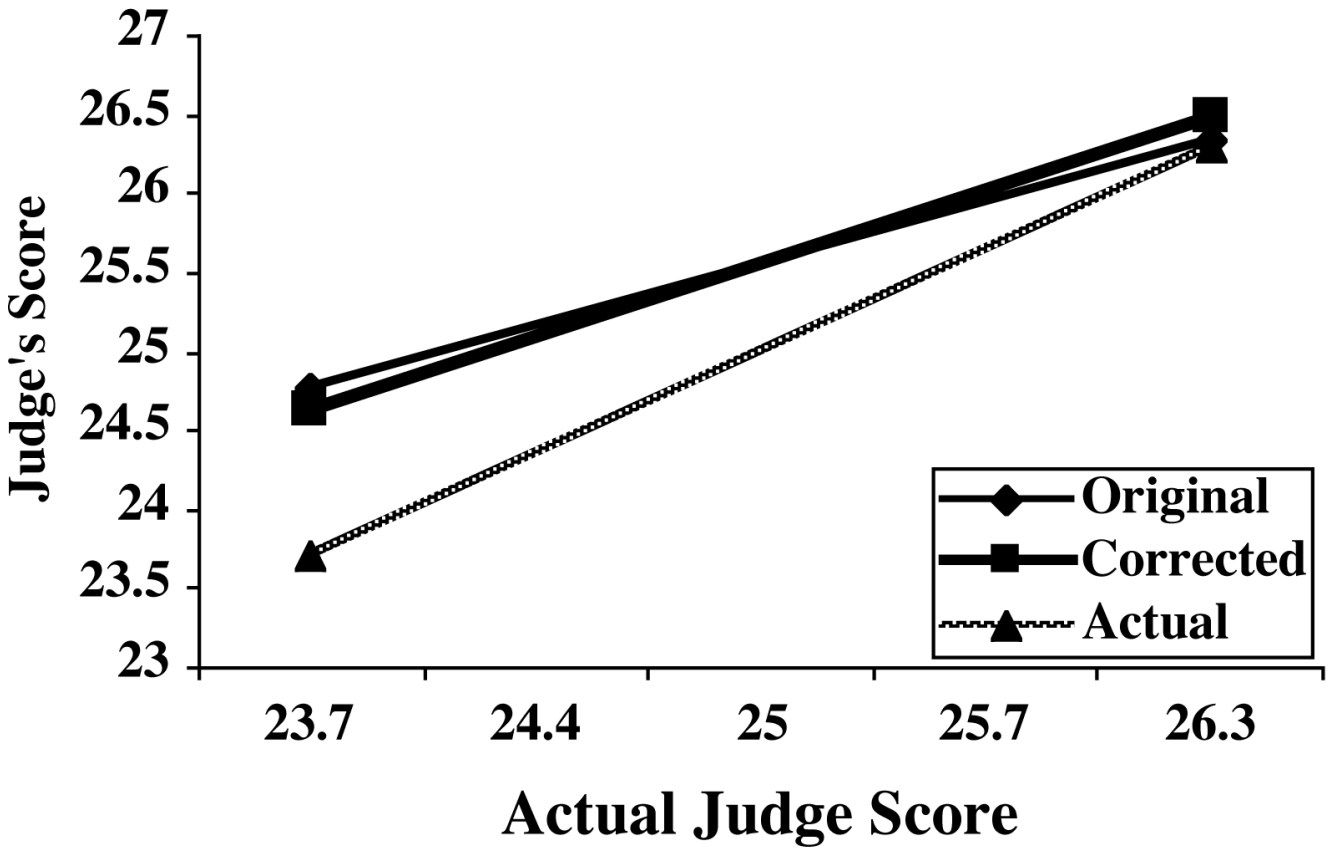
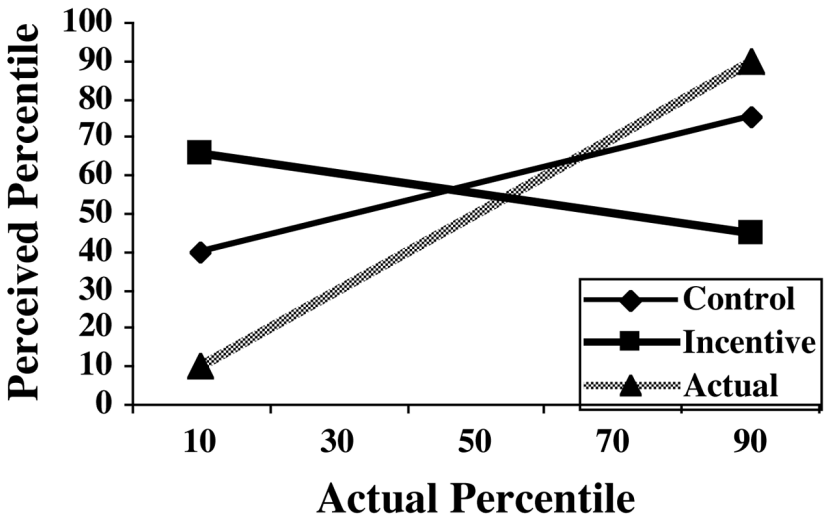


Figure 4.
Estimates of Judge's Score as a Function of Actual Score, Before (Original) and After (Corrected Correcting for Measurement Unreliability (Study 2).

A. Percentile Estimates



B. Raw Score Estimates

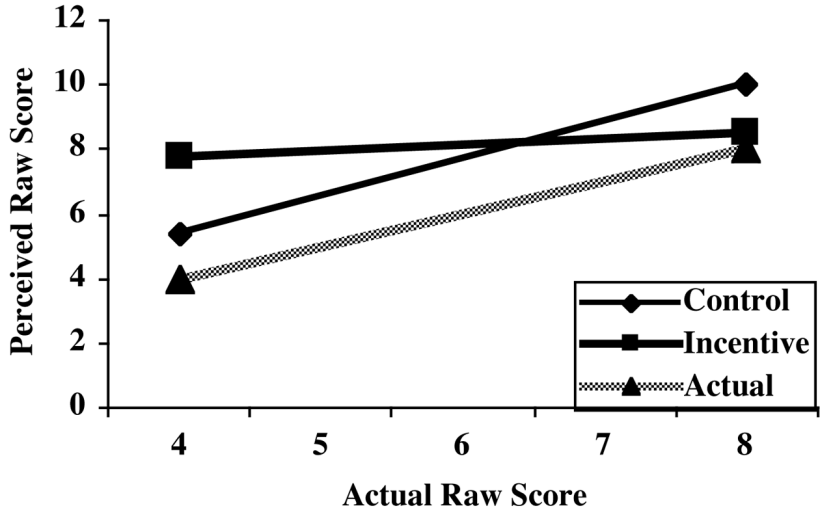
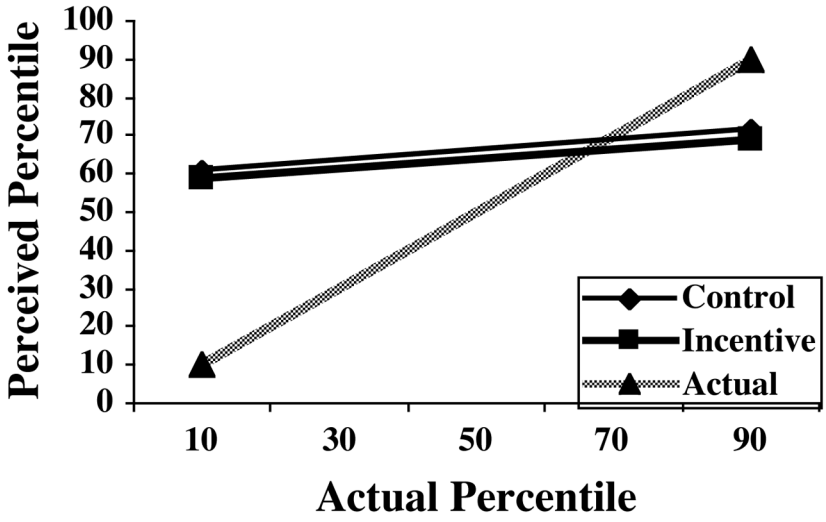


Figure 5.
Perceived Versus Actual Percentile and Raw Score as a Function of Actual Performance and Incentive Condition (Study 3).

A. Percentile Estimates



B. Raw Score Estimates

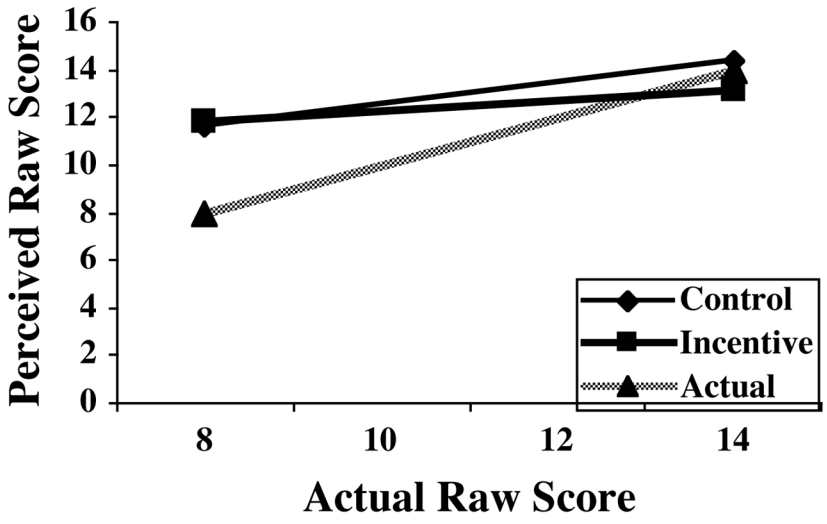


Figure 6.
Perceived Versus Actual Percentile and Raw Score a Function of Actual Performance and Incentive Condition (Study 4).

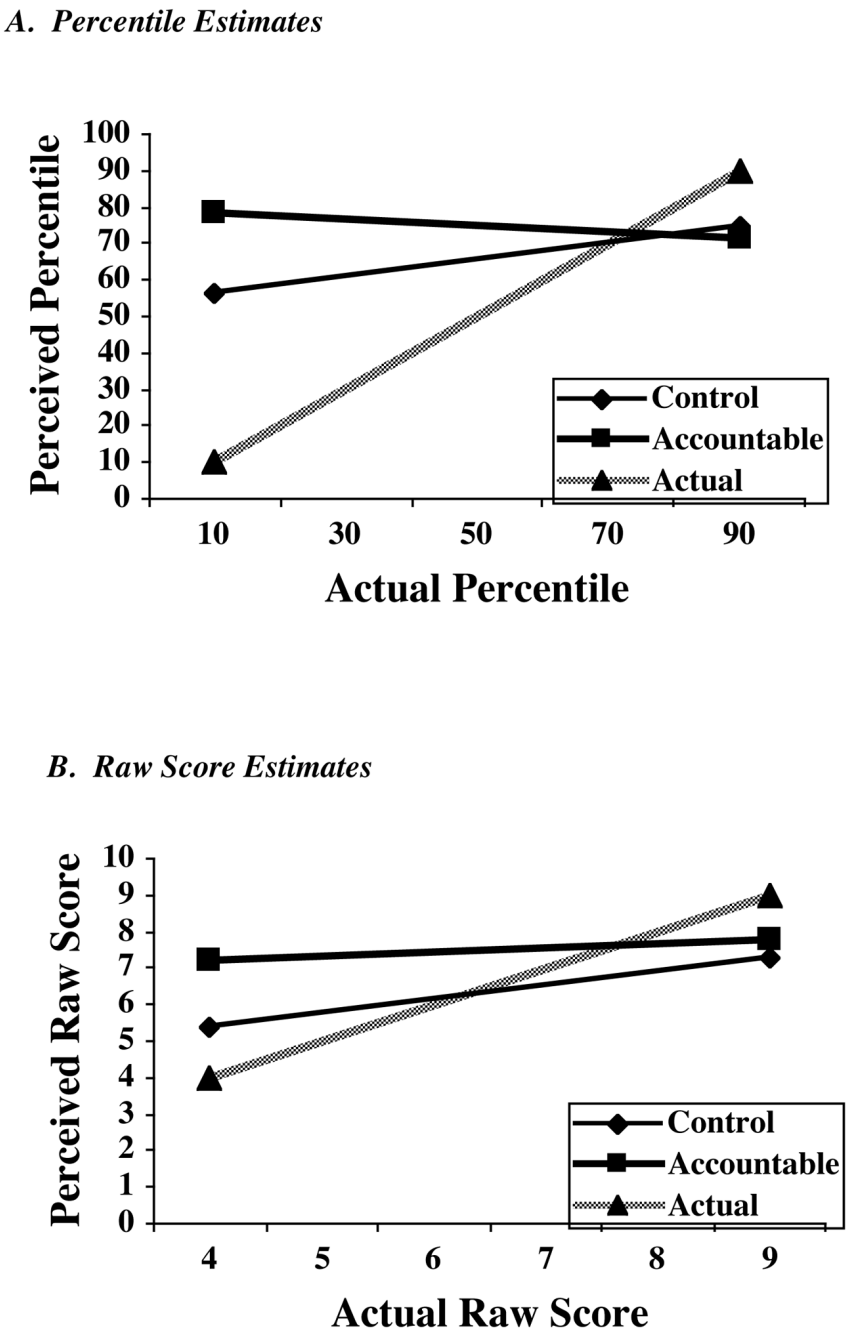


Figure 7.
Perceived Versus Actual Percentile and Raw Score as a Function of Actual Performance and Accountability Condition (Study 5).

Table 1

The Difference between Students' Perceived and Actual Performance Across Studies, as a Function of How Well They Actually Performed (Bottom Quartile, Top Quartile, Or in the Middle 50%), Presented Separately For Estimates of Their Percentile and Raw Scores.

Performance Percentile Estimates

Group	Bottom 25%			Middle 50%			Top 25%		
Study	**Perceived**	**Actual**	**Difference**	**Perceived**	**Actual**	**Difference**	**Perceived**	**Actual**	**Difference**
Study 1	61.2%	14.9%	46.3%***	69.4%	49.8%	19.6%***	73.3%	87.0%	−13.7%***
Study 3	64.0%	11.4%	52.6%**	1.6%	45.4%	6.2%	61.5%	86.9%	−25.4%*
Study 4	58.8%	14.0%	44.8%***	67.3%	56.1%	11.2%**	68.6%	92.1%	−23.5**
Study 5	64.4%	17.1%	47.3%***	73.3%	56.0%	17.3%***	71.4%	89.4%	−18.0%*
Overall	62.1%	14.4%	47.8%	65.4%	51.8%	13.6%	68.7%	88.9%	−20.2%

Raw Score Estimates (as a Percentage of Total Possible)

Group	Bottom 25%			Middle 50%			Top 25%		
Study	**Perceived**	**Actual**	**Difference**	**Perceived**	**Actual**	**Difference**	**Perceived**	**Actual**	**Difference**
Study 1	76.0%	55.3%	20.7%***	76.0%	55.3%	20.1%***	82.3%	87.0%	−4.7%*
Study 3	37.5%	7%	30.5%**	78.2%	56.4%	30.5%	73.3%	83.7%	−10.4%*
Study 4	60.0%	41.3%	18.7%***	64.7%	60.6%	18.7%	71.7%	74.4%	−2.7%
Study 5	66.4%	50.0%	16.4%*	71.6%	74.4%	16.4%	75.0%	91.3%	−16.3**
Overall	60.0%	38.4%	21.6%	72.6%	61.7%	21.4%	75.6%	84.1%	−8.5%

Note: Participants did not provide comparable percentile and raw score estimates in Study 2.

* $p < .05$,

** $p < .005$,

*** $p < .0001$

Table 2

Bottom and Top Performers' Average Actual Score, Estimated Raw Score and Estimates of the Average Raw Score Achieved By Other Participants.

	Bottom Performer Estimates			Top Performer Estimates			
Study	**Self**	**Other**	**Actual**	**Self**	**Other**	**Actual**	**Average Performance**
Logic I	70.9	72.7	48.2	70.0	63.7	79.6	64.7
Grammar	64.7	68.8	45.9	84.7	77.2	82.1	66.4
Logic II	55.3	62.7	3.2	88.9	73.9	100.0	49.1
Exam	76.0	77.6	55.5	82.3	78.8	87.0	71.2
Overall	65.7	69.9	33.3	83.1	74.4	89.0	61.2

Note: Raw Scores are expressed as a percentage of total score possible in each study.

Table 3

Unstandardized Regression Weights Indicating The Strength of the Relationship Between Estimates of One's Own Raw Score for Self and The Average Raw Score Achieved By Others, Shown for Percentile Rankings of Both Ability Level and Test Performance.

	Ability: *B* for		Performance: *B* for	
Study	**Self**	**Average Other**	**Self**	**Average Other**
Logic Study I	.56	−.17	.98	−.39
Grammar Study	.83	−.39	.96	−.58
Logic Study II	.85	−.33	.79	−.43
Exam Study	.93	−.63	1.26	−.77
Overall	.84	−.42	1.00	−.56

Table 4

Bottom Performers' Percentile Estimates Both In Their Original Form and Corrected for Errors in Estimates of the Self or the Average Score Achieved By Others.

	Ability Estimates Corrected for			Performance Estimates Corrected for			
Study	**Self**	**Other**	**Uncorrected**	**Self**	**Other**	**Uncorrected**	**Actual**
Logic I	54.9	69.0	67.6	39.9	66.0	62.3	12.2
Grammar	51.1	67.7	66.8	42.3	61.9	60.5	10.1
Logic II	19.6	63.8	54.9	11.5	59.0	52.8	13.2
Exam	44.2	67.3	62.2	35.5	66.2	61.2	14.9
Overall	37.3	66.2	60.8	28.1	62.7	58.0	13.1

Table 5

Top Performers' Percentile Estimates Both In Their Original Form and Corrected for Errors in Estimates of the Self or the Average Score Achieved By Others.

	Ability Estimates Corrected for			Performance Estimates Corrected for			
Study	**Self**	**Other**	**Uncorrected**	**Self**	**Other**	**Uncorrected**	**Actual**
Logic I	79.6	74.1	74.2	78.2	68.3	68.8	85.6
Grammar	69.4	75.8	71.6	67.0	75.8	69.5	88.7
Logic II	83.5	88.6	76.0	88.0	90.3	79.2	90.0
Exam	78.0	78.5	73.7	79.3	79.2	73.3	87.0
Overall	78.1	79.2	74.0	79.3	80.4	73.7	88.1