IN PRESS AT *JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY*

The Hobgoblin of Consistency:

Algorithmic Judgment Strategies Underlie Inflated Self-Assessments of Performance

Elanor F. Williams

University of California, San Diego

David Dunning

Cornell University

Justin Kruger

New York University

Authors' Note

Elanor F. Williams, Rady School of Management, University of California, San Diego; David Dunning, Department of Psychology, Cornell University; Justin Kruger, Stern School of Business, New York University. This research was financially supported by National Science Foundation Grant 0745806 awarded to Dunning. We thank Fiona Lee for her help in data collection.

Correspondence concerning this article should be directly to either Elanor Williams at Rady School of Management, Otterson Hall, 9500 Gilman Drive, University of California, San Diego, La Jolla, CA 92093-0553, email: ewilliams@ucsd.edu, or to David Dunning at Department of Psychology, Uris Hall, Cornell University, Ithaca, NY 14853, email: dad6@cornell.edu.

Abstract

People often hold inflated views of their performance on intellectual tasks, with poor performers exhibiting the most inflation. What leads to such excessive confidence? We suggest that the more people approach such tasks in a "rational" (i.e., consistent, algorithmic) manner, relative to those who use more variable or *ad hoc* approaches, the more confident they become, irrespective of whether they are reaching correct judgments. In six studies, participants completed tests involving logical reasoning, intuitive physics, or financial investment. Those more consistent in their approach to the task rated their performances more positively, including those consistently pursuing the wrong rule. Indeed, completely consistent but wrong participants thought almost as highly of their performance as did completely consistent and correct participants. Participants were largely aware of the rules they followed, and became more confident in their performance when induced to be more systematic in their approach, no matter how inadequate that performance was. In part, the link between decision consistency and (over)confidence was mediated by a neglect of alternative solutions as participants followed a more uniform approach to a task.

Keywords: self-evaluation, metacognition, performance evaluation, self-enhancement, overconfidence

The 1929 Rose Bowl football game was scoreless midway through the second quarter at the moment that Georgia Tech's Jack Thomason fumbled the ball near the 30-yard line. Roy Riegels, playing center for the opposing University of California, Berkeley, team, then managed to do almost everything right. He scooped up the ball, quickly scanned the chaos of players in front of him, and then ran for the daylight of the goal line to try to score a touchdown. There was only one flaw. In the middle of the maelstrom, Riegels had been turned around and was now running for the wrong goal line. His teammate Benny Lom ran nearly 70 yards to catch Riegels, stopping him just before he crossed the goal line, but by then it was too late. The Georgia Tech team tackled Riegels on his own 1-yard line as he tried to turn around, and then blocked a punt on the very next play to set up the 2-point safety that ultimately provided the winning margin in the game.

We bring up this incident involving Mr. Riegels because we propose that people in everyday life may often face situations much like his. They have some everyday intellectual puzzle to solve, and some circumstance or force turns them to sprint energetically and confidently toward a wrong solution. What circumstances steer people toward confident error? What factors lead people to run unwittingly toward mistaken solutions?

We propose that people often bring to mind some rule or algorithm, one they follow in a systematic way, when they face a particular class of everyday puzzles. Frequently, that algorithm will be right and, thus, using it consistently is an accurate and useful cue to competence. However, at other times, that rule carries a glitch or omission that steers people, instead, toward a wrong answer. It is likely, for example, that Mr. Riegels had presumably prepared for this exact Rose Bowl situation many times on the practice field, following a simple rule: upon gaining possession of a fumbled ball, if there is an open goal line ahead, go for it.

Unfortunately for Mr. Riegels, that rule required an important addendum: make sure first that the goal line ahead is the right one.

The problem for the everyday decision-maker is that there is often no Benny Lom to inform them when the rule they follow leads them in the wrong direction. As a consequence, they remain unaware that their algorithmic solution is incorrect every time they faithfully follow it to a wrong conclusion. What, however, of their confidence? Will having a rule to cite lead people to be mistakenly confident in wrong conclusions? Relative to those who do not follow any hard and fast rule, will those faithfully following some formula express confidence that may not be accompanied by much in the way of demonstrable competence?

The Dunning-Kruger Effect

We bring up the issue of flawed algorithmic thinking because we have frequently observed poor performers, making mistake after mistake, assess their conclusions almost as favorably as do highly competent peers reaching much more accurate judgments. This phenomenon has come to be known as the Dunning-Kruger effect (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999). Incompetent performers display little insight into just how poorly they perform. On average, the bottom 25% of performers on some task, such as logical reasoning, grammar, or taking a classroom exam, think their performance lies above the 60th percentile, overestimating their raw score by as much as 50% (for a review, see Dunning, 2011).

Where assessed, this effect has emerged not only in the laboratory, but also in real world settings. Chess players who are less skilled overestimate how well they will do in an upcoming chess tournament more than their more skilled counterparts (Park & Santos-Pinto, 2010), as do people playing in a weekly bridge club (Simons, in press). Medical students headed for bad

grades (e.g., D+) in their obstetrics and gynecology clerkships overestimate the grade they will get by two full steps (e.g., B+), whereas their more competent peers provide better predictions of future grades. The same overestimation arises among participants at a Trap and Skeet competition assessing their knowledge about firearm usage and safety (Ehrlinger et al., 2008), and among groups as varied as debate teams (Ehrlinger et al., 2008), drivers (Mynttinsen et al., 2009), pharmacy students (Austin, Gregory, & Galli, 2008), and lab technicians (Haun, Zerinque, Leach, & Foley, 2000).

Why and when do poor performers overestimate their achievement? This overestimation by poor performers does not seem to be due to statistical artifact, such as regression to the mean (Krueger & Mueller, 2002). Nor does it seem to be due to participants not caring about accurate self-assessments. Correcting for regression artifacts accounts for very little of the effect (Ehrlinger et al., 2008; Kruger & Dunning, 2002). Providing financial incentives of up to $100 does nothing to enhance participants' accuracy. Nor does making people accountable to another person for their assessments (Ehrlinger et al., 2008).

Could the consistent use of inappropriate algorithms underlie, at least in part, the Dunning-Kruger effect, leading to confidence despite lack of competence? In this manuscript, we investigated the potential impact of such algorithms for self-evaluation of performance. We asserted that the *consistency* or *systematicity* with which people approach a class of problems positively predicts how confident they are in their performance, irrespective of the actual quality of that performance. In contrast, people who approach each problem in a variable or *ad hoc* way would be less confident in their decisions. If this is the case, then we may have a partial explanation for why and when people in general tend to hold overly favorable views of their choices (Dunning, 2005; Dunning, Heath, & Suls, 2004), as well as specifically why so many

poor performers prove to be quite self-assured about their performance despite demonstrating a lackluster level of competence (Dunning, 2011; Dunning et al., 2003; Kruger & Dunning, 1999).

The Role of Rational Errors

Of course, there are many times when being more consistent will also lead to greater accuracy. Often, rule-based judgment is a valid indicator of accuracy. For example, comparisons of clinical and actuarial judgments reveal that experts often identify appropriate markers of accurate judgment but then fail to apply them consistently, and are thus outperformed by statistical models that slavishly apply those judgment rules (e.g., Dawes, 1979; Dawes, Faust, & Meehl, 1989). In addition, work on multiple cue probability learning (MCPL; e.g., Hammond, Hursch, & Todd, 1964; Tucker, 1964), based on Brunswick's classic lens model of judgment (1955, 1956), has traditionally held that more consistent use of cues leads to greater accuracy.

Accurate judgments, however, require not only that judges consistently rely on cues, but also that those cues are valid (e.g., Brunswik, 1956). We asked what happens when people follow a consistent approach to judgment using invalid cues or rules. People are known, at times, to consistently rely on cues that are not probative of truth. In domains as far-reaching as assessing one's personality based on one's possessions (Gosling, Ko, Mannarelli, & Morris, 2002), predicting the outcome of sporting events (Hall, Ariss, & Todorov, 2007), or detecting lies from truths (e.g., Zuckerman, Koestner, & Driver, 1981), people depend on uninformative cues and disregard predictive ones when making judgments. Lie detectors, for instance, may err in believing that any one liar has unique "tells," as Dawes and colleagues warn us (e.g., Dawes et al., 1989), but they may also be led astray by detecting traits and behaviors that they believe all liars share but in actuality do not.

In a phrase, we proposed that it is "rational" errors, more than haphazard ones, that

underlie the undue confidence that poor performers imbue in their decisions. By rational, we refer to errors that are produced by the systematic application of a misguided rule rather than by a more haphazard and inconsistent process (Ben Zeev, 1995, 1998; Ben Zeev & Star, 2001).[1] These are algorithms that contain corruptions or "bugs" that steer people away from providing the right answer. For example, many of the errors made by elementary school math students turn out to be not random but systematic in nature (Ashlock, 1976; Brown & VanLehn, 1980; VanLehn, 1983). A student may look at the equation *63 – 29 = ?* and decide that the correct answer is 46. Apparently, such a student knows that one of the numerals lined up in the "tens" and the "ones" columns must be subtracted from the other, but fails to recognize that it is always the numeral associated with the second number (i.e., 29). Instead, the student consistently assumes it must be the smaller numeral, and so subtracts the 2 from the 6 and the 3 from the 9, systematically and unknowingly arriving at the wrong answer (Ben Zeev, 1995).

The notions that comprise rational errors come from many sources. Students may learn a procedure on a constrained set of examples with solutions that are consistent not only with the correct one but also plausible but erroneous alternatives (Wason, 1960). Or, people may introduce misguided metaphors into their problem solving. For example, novice physics students often bring their everyday experience with substances and fluids into their understanding about the physics of light, heat, and electricity. This leads them to make systematic errors in inference, such as believing that light travels farther at night than during the day because ambient illumination during the day applies more friction to any individual beam of light (Reiner, Slotta, Chi, & Resnick, 2000).

Why might consistent use of erroneous rules lead to confidence? One such plausible mechanism, which we test, is that having a rational rule to follow, whether right or wrong, may

preclude people from considering alternative conclusions. With a rule already in mind, people need merely apply the rule to the problem and not concern themselves with alternative approaches to a problem—alternatives that may lead to different conclusions. Not considered, these alternatives have little or no chance to inject uncertainty or doubt into a person's preferred conclusion. Past work in judgment and decision has shown that inducing people to consider alternative decision options makes them less confident in the one they ultimately choose (Arkes, Faust, Guilmette, & Hart, 1988; Koehler, 1994; Koriat, Lichtenstein, & Fischhoff, 1980; Lord, Lepper, & Preston, 1984; McKenzie, 1998). Herein, we ask if people using rules consistently report thinking about fewer alternatives, whereas their more *ad hoc* counterparts report thinking about a greater number of alternatives. We then examine whether the neglect of alternative approaches underlies a positive view of one's performance in self-evaluation.

Relation to Past Literature

The notion that consistency in decision processes is linked to confidence resonates with past research. Kruglanski's (1980, 1989) lay epistemics, for example, suggests that if people generate inference rules that point uniformly to a specific conclusion when testing a hypothesis (e.g., if she is smiling, she must be happy), they will be confident. If different rules come to mind that indicate different conclusions (e.g., or, a smile might mean she is just covering up embarrassment), confidence wanes.

Two more recent research statements suggest a consistency-confidence link—even when consistency points to a wrong answer. First, in his self-consistency model of subjective confidence, Koriat (2012) suggests that people are confident in their answers to the extent that inputs into possible answers (his terminology refers to inputs as "representations") converge to one possible answer over the alternatives. What predicts those instances in which one answer is

heavily favored by inputs? According to Koriat, it is the consistency one finds in people's answers (Koriat, 2008). To the extent that an individual gives the same answer to a question over repeated questioning, one can assume from that consistency that the individual's inputs heavily favor the answer given. In addition, to the extent that different people give the same answer, one can assume from this cross-person consistency that the available inputs favor the answer everyone gave over alternatives. Importantly, consistency is the key, not accuracy. People will be confident in answers associated with consistency across time or people regardless of whether those answers are wrong or right.

Although the current work shares some similarity with Koriat's (2012) model, it differs from that model in substantive ways. First, Koriat applies his model specifically to answering general knowledge questions (e.g., Who played Dorothy in the movie *The Wizard of Oz?*), whereas we focus more on problem-solving tasks. It is likely that trying to answer questions of fact differs substantively from problem solving, and confidence may rest on different meta-cognitive features and experiences.

More importantly, we focus on a different type of consistency from that explored by Koriat (2012). Whereas he focused on consistency across time and people, we focus on consistency across problems. Do people assume a set of problems fall in a single "equivalence class" that calls for the application of same rule to reach a solution? In examining this type of consistency, instead of asking whether multiple inputs tend to consistently favor one solution over another, we ask instead whether people rely mostly on one input, a rational rule or algorithm, or whether they switch from input to input as they move from problem to problem. As such, in terms of the types of task in focus, the type of consistency being studied, and the psychological process that may lead to that consistency, our framework differs from Koriat.

Second, our work is also reminiscent of Tetlock's (2005) work on political expertise, in which he asked nearly 300 experts to make 28,000 predictions about future world events. Drawing from the famous distinction made by philosopher Isaiah Berlin (1953), Tetlock divided his experts into *hedgehogs*, those who approached their predictions with an overarching grand theory, versus *foxes*, who tailored their approach to each prediction. He found that hedgehogs were more overconfident in their predictions than were foxes. Tetlock anticipates our hypotheses; individuals who consistently applied an overarching rule or theory were the most overconfident, yet again the situation we test differs importantly from the one he placed under scrutiny. He examined the complex and ill-defined task of predicting the future, in which there may very well be no overarching algorithm that provides the best answer. Being consistent or rational may in and of itself be a mistake. Indeed, in his work, he found that foxes, who dismiss the value of overarching theories, were more accurate in their predictions than their hedgehog counterparts.

Here, however, we examine cases in which there is a correct algorithm to follow. The best performers will necessarily be hedgehogs consistently applying the right algorithm, rather than foxes using different approaches for each problem. The question then changes to how much more confident will these "correct" hedgehogs be relative to those applying an incorrect set of rules? We predict that there might be a difference in confidence, but that it will hardly anticipate just how different the performance is between "correct" and "incorrect" hedgehogs.

Overview of Studies

Data from six studies were examined to test whether decision consistency underlies inappropriate confidence among performers in general, as well as misguided confidence among incompetent performers specifically. In four of the studies, we asked participants to tackle

Wason selection tasks (Wason, 1966). In this classic logical reasoning task, participants are shown four cards with information on both sides and are given a logical rule to test. For example, participants may be shown cards that they are told have a number on one side and a letter on the other, such as "A," "B," "4," and "7". They are then presented with a logical rule, such as "If a card has a vowel on one side, then it has an even number on the other side," and are asked which card or cards need to be turned over to make sure the set of cards faithfully follows the stated rule. Participants can turn over from one to four cards, but are instructed to turn over the minimal number necessary.

In Wason tasks, the general form of the logical rule being tested is "If P, then Q," and the information shown on each card corresponds to either P, Q, not-P, or not-Q. The correct answer is always to turn over are the P and not-Q cards (the *A* card and the *7* card in the example above). These cards must be turned over because they are the only ones that can produce a violation of the logical rule in question. However, a long lineage of research on the Wason task has shown that many people choose to turn only the P card over, or the P and Q cards (Wason, 1977). Thus, the Wason selection task presented us with a logical reasoning task in which the correct and the common but incorrect algorithms that people tend to use are well-known.

In two additional studies, we examined other everyday problems for which the types of errors people commonly make are also well-known. One was an intuitive physics task. Participants examined a series of curled tubes and were asked what trajectory a ball or bullet would take once it exited the tube. The correct answer is that the ball or bullet continues in a straight path, but a minority of people mistakenly believe that the ball or bullet's trajectory will preserve the curve of the tube, causing it to curl after it leaves the tube (McCloskey, 1983; McCloskey, Caramazza, & Green, 1980). How confident are people when they apply this faulty

notion consistently, relative to those who provide the right conclusion consistently?

The second task, a financial one, required participants to estimate the amount of money they would gain by investing in bonds. Participants were presented with a scenario in which they invested $100 in a bond, and were asked how much money they would receive if they kept the bond until maturity (e.g., 5% interest compounded annually that matures in 20 years). Past work has shown that many people miss the fact that their money will grow exponentially, and instead assume that it will grow in a steady linear rate (e.g., so that, in the example above, participants would have $200 at maturity rather than $265) (Benzion, Granot, & Yagil, 1979; Christandl & Fetchenhauer, 2009; Wagenaar & Timmers, 1979). By examining those who made this error, we could see if people who consistently assumed linear growth would be more confident in their responses than those answering more variably.

Thus, across the six studies, we examined the following hypotheses:

Replicating the Dunning-Kruger effect. First, we predicted that participants would generally overrate their performance on tasks we presented them, and that this overestimation would show a Dunning-Kruger pattern. Specifically, high performers would slightly underestimate their performance, whereas poor performers would greatly overestimate theirs. However, we investigated this effect in a slightly different way from Kruger and Dunning's (1999) original investigation. Instead of splitting participants into quartiles based on their objective performance, we ran regression analyses investigating the continuous relationship between perceived and objective performance.

We moved to a regression analysis because it allowed us explore a prediction made about the Dunning-Kruger effect. Namely, the relationship between perceived and objective performance should show signs of being curvilinear. It should be relatively tight for better

performers, but it should weaken and flatten as performances become poorer and the link between actual and perceived performance decouples (Burson, Larrick, & Klayman, 2006). Thus, we ran regression analyses with a quadratic component, as well as a linear one, to test for curvilinearity in performance evaluation.

Examining the role of consistency in confidence. The tasks we chose allowed us to assess the degree to which participants approached individual problems on an *ad hoc*, case-by-case, basis or more in an overarching, "rational" way. We predicted that participants would reach more positive self-evaluations of their performance, irrespective of actual performance, to the extent that they approached the task in a systematic way.

We obviously expected participants following a correct rule to do quite well on our tasks and to deserve any positive self-evaluations they provided. However, we also predicted that participants would become more confident in their performance to the extent they faithfully followed a wrong rule even as their performance actually worsened. For them, confidence would increase even as performance deteriorated.

Focusing on poor performers. Of key interest was examining the role that consistency played among poor performers. Would the consistent commission of rational errors explain the exorbitant and unwarranted confidence that poor performers sometimes imbue in their answers? To explore this, we examined three separate issues. The first was whether judgmental consistency would be related to positive perceptions of performance within the group of poor performers: Those following a systematic rule were expected to be more optimistic about their performance, even when that rule leads them to systematic error. The second issue was whether, among poor performers, consistency is a vice rather than a virtue. That is, among poor performers, would consistency be associated with *worse*, not better, performance? The third

issue focused on those who showed complete consistency in their wrong answers. To what extent would they show confidence similar to that expressed by completely consistent but right participants? Would confidence differ across both groups or would it be largely the same?

Determining if rule-following is conscious. We expected people to be more confident when faithfully applying rules to the Wason task, but would they be aware of it? Would they be accurate in recognizing or describing which rule they were applying? We did not want to assume they were (see Nisbett & Wilson, 1977). Thus, in Study 4, we asked participants explicitly whether there was a systematic rule they were following and, if so, to identify it. By doing so, we could better describe the phenomenology of inappropriate confidence due to rational errors. We could see whether participants approaching the Wason task in an algorithmic way actually were aware of it, as well as whether they were aware of the specific rule they were following. Is the pursuit of rational errors a conscious phenomenon or is it a nonconscious one?

Manipulating consistency. Study 5 contained an experimental manipulation aimed at influencing the consistency with which participants approached the Wason task. This allowed us to see if judgmental consistency played a causal role in inappropriate confidence in one's own performance. In inducing more consistency, would we inspire positive self-assessments that were not justified by enhanced performance?

Asking why consistency leads to unwarranted confidence. Finally, Study 6 examined a potential mediator linking judgmental consistency to enhanced self-evaluation. Specifically, we examined the degree to which participants considered alternative solutions to the ones they ultimately decided upon. Did consistent participants rate their solutions as more accurate because they were also more likely to neglect alternative solutions, relative to participates who reached different *ad hoc* solutions for each individual problem?

Method

Methods and results of the six studies are described together due to their similarity. Study 1 is a re-analysis of data from Kruger and Dunning (1999, Study 4). Studies 2 - 6 are novel to this manuscript.

Participants

In Study 1, participants were 140 Cornell University undergraduates who earned extra credit toward their course grade for participating. The participants in Studies 2 through 6 were recruited in the United States via Mechanical Turk, a crowdsourcing platform administered by Amazon.com, and were paid $0.50 for their participation. One-hundred-two participants completed each of Studies 2 and 3. In Study 4, 122 individuals participated, 137 individuals participated in Study 5, and Study 6 included 80 Mechanical Turk participants. Save where explicitly stated in the text below, we did not exclude any participants from our analyses.

Common Procedure for Studies 1, 4, 5, and 6

In four of the six studies, participants were told they were taking part in a study on logical reasoning. They completed a quiz with 10 logical reasoning problems, all of which were based on the Wason selection task (Wason, 1966). The items ranged from those that were abstract in content (i.e., the A, B, 4, 7 problem mentioned earlier) to those that were more concrete or familiar (i.e., a classic problem involving checking ID in a bar for a 16-year-old, a 21-year-old, a beer drinker, and a tea drinker). Except for Study 1, participants for each item indicated how likely it was that they had reached the correct response, estimating the percentage likelihood that they were right. This served as an indicator of their confidence in each individual response.

After completing all items, participants in all studies rated their performance in three additional ways. They rated their overall logical reasoning ability in percentile terms, indicating

the percentage of Cornell students (Study 1) or Mechanical Turk workers (Studies 4 - 6) they outperformed in terms of logical ability. They also rated their specific test performance along the same percentile scale. Finally, they indicated how many out of the 10 items they had answered correctly.

Study 2: Intuitive Physics

Study 2 served as a conceptual replication of Study 1, with only the task switched. Instead of a logical reasoning task, participants were shown a series of curved tubes and asked what trajectory either a ball or a bullet would take once it exited the tube. In all, participants answered questions about 6 different displays. For each, they could state that the object would (1) fly in a straight trajectory after leaving the tube, (2) continue to curve in a manner similar to the tube's curve, an inward curving path, (3) curl against the curve of the tube, that is a outward curving path, or (4) travel a straight path somewhat against the curve of the tube. After each item, as a measure of confidence, participants rated the chance that their answer was correct.

After providing answers to each of the six tube displays, participants rated their spatial ability in "tasks like these" on a percentile scale relative to "others participating in this experiment". They also rated their specific task performance in percentile terms, and estimated how many of the six items they had gotten right.

Study 3: Finance Task

Study 3 involved a different task. Participants were presented with a scenario in which they invested $100 in a bond, and were asked how much money they would receive if they kept the bond until its maturity. Each of ten scenarios varied the interest rate of the bond and its maturity period (e.g., 5% interest compounded annually that matures in 20 years). Participants were given four choices about how much money they would make. One choice merely stated the

original amount raised to its stated interest level (e.g., $105). Another raised the original amount to an average of the maturity period and the interest rate (e.g., $113). One raised the original amount linearly, multiplying the interest rate by the number of years until maturity and adding it to the original amount (e.g., $200). One response was the correct answer, which accounted for the fact that the amount returned would grow exponentially (e.g., $265). These response options were based on typical responses people give to questions about exponential growth (e.g., Christandl & Fetchenhauer, 2009, Study 4).

For each response, participants rated the chance that their judgment was correct. At the end of the quiz, participants rated their financial reasoning ability and performance on the quiz in percentile terms. They also estimated how many of the 10 items they had answered correctly.

Study 4: Reporting Explicit Judgmental Strategies

Study 4 returned to the Wason selection task. After answering the 10 items but before evaluating their overall performance, participants reported whether they had followed a specific strategy in approaching the test. They were told that the test had a structure, in which some specific logical rule corresponding to the general format "If card shows P, then it must have a Q on the back" had been stated, and then options showing P, not-P, Q, and not-Q had been displayed. Participants were asked to answer *yes* or *no* to the question: "To respond to the problems above accurately, was there an overarching rule you had to follow about which cards got turned around?" If participants indicated "yes," they were asked which specific rule they had followed: turning over the P card only, the P and Q cards, the P and not-P cards, or the P and not-Q cards (the correct response), or some other strategy, which they described in their own words.

Study 5: Manipulating Judgmental Consistency

Study 5 also included procedures to manipulate the level of consistency with which they

approached the Wason task. For roughly half of participants (n = 71), participants examined 4 practice items before they began the quiz proper. After completing these items, participants were asked, yes or no, whether there was any overarching rule they should follow in addressing these types of items, using the same question used in Study 4. Note that this question was designed not to suggest that there was or was not a rule to follow. However, we felt that merely asking the question might lead some participants to conclude that the Wason task required a rule-based response. If participants answered "yes" to the question, they were asked what the specific rule was, using the same response options described in Study 4. They then continued on to respond to the 10 problems that comprising the quiz itself.

Participants in the control condition (n = 66) completed a similar task, except they confronted 4 practice questions on intuitive physics like those used in Study 2 (McCloskey et al., 1980). They were similarly asked if there were a specific rule to answer those questions.

Study 6: Assessing the Consideration of Alternative Solutions

In Study 6, after providing their final answer for each item but before they recorded their confidence in that answer, participants answered two additional queries. First, they were asked if they had "seriously considered" selecting any of the cards they ultimately decided not to turn over. Second, they were asked if they had seriously considered not turning over any of the cards they had selected. For both questions, participants indicated the individual cards they had seriously considered in their alternative solutions.

Results

In the following, we report the results of analyses after combining the results across studies where possible. Statistics associated with individual studies or measures are depicted in the tables and figures noted.

Overall, participants' self-perceptions of their performance were unrealistically positive. As seen in Table 1, participants in each study rated their ability to be at least in the 64.5th percentile and their test score to rank in at least the 60.6th percentile—self-ratings that deviated from the true average (the 50th percentile) quite significantly, all ts > 5.5 for ability and 3.6 for test score, respectively, ps < .0001. Participants also overestimated their raw score from 16% to 55% items, all ts > 4.8, ps < .0001. In terms of confidence, the average confidence participants imbued in their individual responses exceeded their actual accuracy rates by at least 20% across all studies, ts > 7.1, ps < .0001.

To examine the relationship between the perception of performance and its reality, we first examined the four studies using the Wason selection task (Studies 1, 4, 5, and 6).

Replicating the Dunning-Kruger Effect

The results above indicate that the degree of unrealistic self-views was substantial in all six studies, but who exactly showed the most overestimation? Replicating the Dunning-Kruger pattern, poor performers supplied much of the overestimation. Figures 1 (percentile measures) and 2 (absolute measures) depict the relationship between self-ratings and objective performance for the studies involving the Wason selection task. Specifically, the figures depict the results of regression analyses in which each self-assessment measure is regressed on objective performance. For each regression, both linear and quadratic components were included.

Across all studies, a general and replicable picture emerges of the relationship between self-perception and the reality of performance. As evident in Table 2, the relationship between self-evaluations and objective performance tended not to be completely linear. Summed across all studies (see Table 2 for the results associated with individual studies), there is a curvilinear bend to the relationship, in that the quadratic coefficient is significant when combined across

studies for all self-evaluation measures, from $z = 2.56$, $p < .01$, for average item confidence to $z = 6.31$, $p < .0001$, for estimates of raw score. For high performers, as seen in Figures 1 and 2, there is a tighter relationship between the perception of performance and the reality, with ratings of performance actually falling a little below the reality of that performance. That is, top performers are slightly underconfident in how well they are doing, replicating the original pattern found in Kruger and Dunning (1999). The relationship between self-perception and reality of performance, however, flattens as one looks at poor performers, and the typical self-evaluations reported by low performers tend to be much more positive than their objective performance should allow.

The Role of Consistency in Self-Evaluation

Across all studies, judgmental consistency was related to both confidence in self-perception and accuracy in objective performance. To assess each participant's level of consistency in the studies involving the Wason selection task, we performed an ANOVA for each individual participant, looking to see how much of the variance of their responses was explained by the category (e.g., P, not-P) of the response options they had available. More specifically, across the 10 Wason selection items, participants had 40 opportunities to turn cards over. We coded a turned-over card as 1 and a card left alone as 0. Then, across all 40 cards, we calculated the total variance each participant displayed (i.e., 40 x p[1-p]) where p is the proportion of cards a participant turned over). We then calculated the between-category variance based on the four categories (P, not-P, Q, not-Q) the cards fell into (i.e., the sum of 10 x $[p_{category} - p_{overall}]^2$ for the four categories). We then divided this between-category variance by the total variance to calculate the proportion of variance in each participant's responses explained by category, or R^2.

An R^2 of 0 means that the card category explained none of the person's response variance; it is as though a participant were choosing among cards randomly and showed no judgmental consistency in their decisions. An R^2 of 1.0 indicated that a participant's responses were governed completely by the card's category membership. This circumstance reflects complete judgmental consistency. For example, a participant who unfailingly chooses P and Q for all 10 items would display an R^2 of 1.0. An R^2 between 0 and 1.0 reflects a more intermediate reliance on card category to inform a participant's response, that is, an intermediate level of judgmental consistency.

Measured in this way, consistency was positively correlated with the favorability of self-evaluation, when combined across all studies, and for all measures (see Table 3, for the results of each individual study). This is true when consistency and self-evaluation measures are simply correlated, all zs > 7.51, $p < .0001$, for individual self-evaluation measures. It is also true after controlling for the quality of objective performance, all zs > 6.50, $p < .0001$, for each self-evaluation measure. Indeed, the correlation between consistency and self-evaluations proved to be higher in 14 of 15 comparisons across the four studies involving the Wason task than it was between objective performance and self-evaluations. Thus, although perceptions of performance tracked actual performance somewhat, those perceptions presumably tracked decision consistency even more.

Consistency was also related to objective performance, but in a complex way (see Table 4 for individual study results). Across all studies and measures involving the Wason task, the more consistent participants were in their approach to the task, the higher their objective performance, zs = 2.72 and 5.98, $p < .01$, for percentile and raw score, respectively. This stands to reason: As a logical reasoning task, all 10 items required the same solution, and so to be

accurate required being more consistent.

However, across all studies involving the Wason, once a quadratic term was introduced into the equation, that linear effect across studies dissipated to nonsignificance, *z*s = 1.23 and -1.26 for percentile and raw score measures, respectively (Table 4 depicts individual study results). Instead, a significant quadratic relationship between objective performance and consistency strongly emerged, *z*s = 11.53 and 9.95, *p*s < .0001, for percentile and raw score measures of performance, respectively. As revealed in Figure 5, high consistency was associated both with performance that was high *and* with that which was low. Lesser consistency was associated with performance that fell more in the middle. In short, high consistency did not indicate good performance as much as it indicated extreme performance, both good and bad.

Poor Performers

We next turned our attention specifically to poor performers and the complex relationship between consistency and accuracy. For high performers, following a consistent rule was associated with high performance, as it should be in a logical reasoning task. If we constrain ourselves across studies to only those scoring 7 or higher on the quiz, we find a strong positive correlation between consistency and objective performance, with the average *r* across studies being .88 and .90 for performance measured in percentiles and absolute raw score, respectively, both *z*s > 14.0, *p*s < .0001. However, for performers at the other end, the direction of the correlation ran in the opposite direction. The more consistent participants are, the lower their scores. Among those scoring 3 or less, those wrong-way correlations were significant across studies, average *r* = -.31 and -.33 for percentile and absolute measures of performance, respectively, both *z*s > 5.6, *p* < .0001.

Consistency also explained high confidence among the poorest performers. For each

study, we identified the bottom 25% of performers as best we could. For Study 1, that comprised individuals who scored 0 or 1 on the quiz (n = 35). For the last three studies, that group comprised those who scored 0 (ns = 47, 84, and 33 for Studies 4 through 6, respectively). We then correlated consistency among these bottom performers with all self-perception measures. As seen in Table 5 (which depicts results from each individual study), for every measure, consistency was positively correlated with self-perception of performance, all zs > 4.8, p < .0001.

How high did consistency drive favorable self-evaluations among poor performers? In each study, there were participants who approached each Wason problem in exactly the same way (i.e., their R^2 was 1.0) and got each item right (n = 52 across all studies). There were also participants who approached each problem in the exact same way but got each item wrong (n = 69). How much did the self-perceptions of performance differ across these two groups? The answer appears to be not much at all. Table 6 shows the average self-ratings provided by the "all-wrong" and "all-right" groups in each individual study. With the exception of Study 6, there is no measure that significantly differentiated the groups within a study, even on a composite measure of self-perception.

It is only after meta-analytically combining the self-ratings across studies that one begins to see a difference, with the all-right group seeing itself as more proficient than the all-wrong group on three self-evaluations measures, zs > 2.2, ps < .05, save ratings of ability in percentile terms, which is only marginally significant, z = 1.69, p < .10. In short, although the all-wrong group rated its performance on average slightly worse than the all-right group, their ratings fail to reflect the reality of their performance. As seen in Figure 4, which combines self-ratings for the all-right and all-wrong groups, the all-wrong group rates their performances, depending on the measure, about 4% to 10% lower than the all-right group, even though the real difference

between the groups is the space between 100% performance and 0% performance.

Physics and Finance Tasks

Analyses centered on the physics (Study 2) and finance task (Study 3) replicated, in general, the results obtained above. Table 2, for example, reveals that the physics (Study 2) and financial tasks (Study 3) tended to produce the same curvilinear relationship between objective performance and self-evaluations. Of the eight separate self-evaluation measures used across the two studies, six displayed a significant quadratic relationship between objective and perceived performance. Figure 5, for the physics task, and Figure 6, for the financial task, shows that top performers tended to underrate their performances somewhat, with the correlation between objective and perceived performance being strong, a relationship that weakened as objective performance became worse, meaning that bottom performers overrated their performance substantially. This pattern replicated the basic Dunning-Kruger effect.

As in the Wason task, judgmental consistency contributed much to perceptions of good performance, irrespective of actual performance. For both tasks, participants' responses on individual items could fall into four different categories. For the physics task, for example, participants could respond that the object would follow a straight, inward curvilinear, outward curvilinear, or outward straight path. To assess the extent to which those categories consistently explained variance in participants' responses, we again calculated an R^2 statistic, taking the ratio of the variance that is attributable to response category to the total variance. If participants consistently chose the same category of response across all items (e.g., they always choose the straight path)— their R^2 was 1.0. However, if participants chose more evenly across the response categories, the R^2 would be smaller, with a lower bound of 0.0.

Overall, consistency computed this way was more tightly related to self-evaluations of

performance than was actual performance, as seen in Table 3. For the physics task, neither consistency nor objective performance was significantly related to self-evaluations along the percentile measures. But for absolute measures, consistency was more strongly related to self-evaluations than objective performance was. For the financial task, both consistency and objective performance was significantly correlated with self-evaluation measures, but for each measure the relationship between consistency and self-evaluations was stronger than the ones involving objective performance.

Consistency was also related to objective performance. In both the physics and the financial task, consistency was positively and linearly correlated with objective performance at least to a marginal degree (see Table 4). However, adding a quadratic term to the analysis revealed that the relationship between consistency and objective performance was strongly curvilinear. As Figure 7 shows, participants performing very well or very poorly tended to be more consistent in their judgments than did those performing more in the middle of the scale.

As with the Wason task, judgmental consistency predicted both good performance among top performers but bad performance among bottom performers. For top performers (i.e., those scoring 4 or more on the physics task; scoring 7 or more on the finance task), we see strong positive correlations between consistency and performance. For the physics task, the correlations are both greater than .99 for percentile and absolute measures of performance, both ps < .0001. For finance, the correlations are .98 and .996, respectively, both ps < .0001. Among bottom performers (i.e., those scoring 2 or less on the physics task and 3 or less on the finance one), the correlation between consistency and performance runs the other way. For physics, the correlations are -.64 and -.61, ps < .0001, for percentile and absolute measures of performance. For finance, the correlations are -.55 and -.58, ps < .0001, respectively. That is,

poor performers who were consistent in their judgments tended to do worse than those who were more variable in the responses they chose.

As in the Wason task, consistency also explained who had positive views of performance among poor performers. In both the physics and finance studies, we identified roughly the worst 25% of performers for each task. For physics, that meant participants who got 2 or fewer right on the task (n = 36); for finance, that meant those who got all items wrong (n = 31). For physics, the composite combining all self-evaluation measures was positively but not significantly correlated with consistency (see Table 5). However, raw score estimates and individual item confidence were significantly correlated with it. For the finance task, all individual measures and the composite were significantly correlated with judgmental consistency. The more consistent people were in their judgment, the more confident they were in their responses. In general, this pattern of responses across both studies fall in line with the results of the Wason studies: Among the poorest performers across all 6 studies, the correlation between consistency and self-view of performance was positive across all 23 tests, and significantly so for 19.

Finally, one last comparison also largely replicated what we found with the Wason task. If we take people who are completely consistent in their responses for the physics task, with 19 getting all items right (i.e., they said the ball's trajectory would be straight) and 3 getting them all wrong (claiming that the ball would continue traveling the tube's inward curve), we find no significant difference in how positive either group is in their performance (see Table 6)—although this is consistent with our prediction, it is also not a surprise, given how few participants fall into each group. For the finance task, if we make the same comparison (with 24 answering correctly and 20 answering consistently but erroneously that money would accumulate linearly), we find that the group that is always correct is more positive about its

performance than its always-wrong counterparts on only one measure (raw score estimates) and is only marginally significantly more positive on the composite measure combining all self-evaluations, $t(42) = 1.83$, $p < .08$. In short, participants who were 100% right in their judgments proved not to be that much more confident in their performance than those who were 100% wrong in their answers.

Conscious Use of Rules

Study 4 was designed to see if the link between consistency and self-perception was a conscious one. That is, did people knowingly follow rules that made them confident in their solutions even if those rules were misguided? Or, did the consistency-confidence link arise without the individual's awareness?

Data from Study 4 suggests that people were mostly aware of whether or not they were following a consistent strategy. First, 59% of participants stated they were following a specific rule in their solutions to the problems. Further, those who stated they were following a rule showed significantly more consistency in their solutions than did those who denied following a rule (see Table 7). Not surprisingly, those who stated they followed a rule rated their performance more positively on all self-perception measures (except for average confidence) relative to those who stated they did not. These more positive self-perceptions arose even though, at least in one measure (raw test score), rule-following participants objectively did worse than their less rule-bound peers.

If rule-bound participants were not actually performing better, what led them to think they were? A mediational analysis showed that it was the consistency of their approach that linked citing a rule to heightened (and misguided) self-perceptions. Figure 8 depicts a mediational analysis linking explicit endorsement of a rule to more positive self-perceptions. As

seen in the figure, citing a rule in Study 4 correlated significantly with consistency (i.e., R^2) and with the composite measure of self-perceptions. When this explicit endorsement is controlled for, consistency still correlates with the favorability of self-perceptions. However, when consistency is controlled for, endorsing a rule is no longer correlated with self-perceptions. A Sobel test reveals that the link between endorsing a rule and self-perceptions is significantly weakened when consistency is controlled for, $z = 3.15$, $p < .005$. In short, those who explicitly cited a rule were, in fact, using a more consistent approach in their decision-making, one that led them to levels of confidence not necessarily matched by objective performance.

The Fate of Correct Versus Incorrect Rules

Other analyses affirmed the notion that people knew what they were doing as they completed the Wason task. Recall that there is a right solution to the Wason task (i.e., turn over the P and not-Q cards). We decided to test the extent to which participants followed that specific rule. We weighted the P and not-Q cards at +1, and the other two cards as -1, thus creating a linear contrast for each participant denoting how much of the total variance of their card-turning could be explained by this specific strategy.

We then did the same for the two most common mistaken strategies people tend to adopt for the Wason task (Wason, 1977). One, the matching strategy, is marked by participants turning over the P and Q cards. To detect this pattern among our participants, we weighted those two cards at +1 and the others at -1, and then examined the extent to which this pattern of responses explained the variance of the participants' choices. The other common erroneous pattern is to turn over the P card only. To detect this pattern, we weighted it +3 and the other cards each -1, and ran for each participant a linear contrast examining how much this pattern explained variance in participants' choices.

We next examined the specific rules that participants said they followed. Of our Study 4 participants, 26 explicitly stated that they were following the correct rule (i.e., turning over the P and not-Q cards), 28 stated they followed a matching (P and Q) rule and 8 reported pursuing the P-only rule. We then subjected the results of each linear contrast described above to a one-way ANOVA, looking to see if the rule participants said they adopted actually explained more of the variances of their choices than it did the choices of participants who cited other rules. As seen in Table 8, those rules did just that for each of the three groups. For example, of participants citing the correct rule, the contrast detecting the use of that rule explained, on average, 61% of the variance of their choices. For those citing the matching rule, the variance explained by the correct rule was only 5%. For those citing the P-only rule, 24%. A linear contrast testing the difference in variance explained between the correct group and the other two groups was significant ($p < .0001$).

The fact that participants largely understood the rules they were following allowed us to construct alternative tests of the notion that following rules led to more confidence, even when it led to more error rather than accuracy. For the three groups described above, we noted the degree to which the specific rule they cited (e.g., the matching rule for the matching group) explained variance in their decisions. In addition, we collapsed the two erroneous groups (matching and P-only) together. We then looked to see if consistency in using the cited rule led to confidence in one's performance. Thus, we regressed the composite measure of participants' self-perception onto their group status (correct versus erroneous), consistency (the variance explained by the specific rule), and the interaction in-between. This regression, depicted in Figure 9, showed that the correct group tended to rate its performance higher than the erroneous group, $\beta = .28$, $p < .04$, but that both groups rated their performances higher to the extent they

followed their cited rule in a consistent manner, $\beta = .28$, $p < .03$. There was no interaction.

A final analysis demonstrated that consistency in decision-making led to confidence regardless of whether it led to better performance. In this analysis, we compared the consistency, self-perceptions, and actual performance of the group of participants stating they were following the correct rule, the group stating they were following a rule (i.e., matching or P-only) that is actually incorrect, and the group stating they were following no rule. As seen in Table 9, both rule-following groups shared displayed more judgmental consistency than the group that denied it was following a rule. On all self-perception measures, a linear contrast weighting the rule-following groups +1 and the non-following group -2 revealed that the rule-following groups had consistently more positive views of their performances than did the non-following group. The correct-rule group rated themselves higher on all 5 self-evaluation measures. The wrong-rule group did the same on 3 of 5 measures including the composite measure, despite the fact that objective performance was lower for the wrong-rule group than it was for the no-rule group.

Causal Role of Consistency in Self-Evaluation

The analyses above all implicate judgmental consistency in heightened confidence in performance, even when that confidence is not warranted. However, these analyses fail to settle the question about whether consistency plays a causal role in confidence. Study 5 was designed, in part, to address this issue. Some participants examined a few practice Wason items and were asked if there was a specific rule to follow in approaching these problems. A full 76% in this condition concluded that there was.

There are a number of interesting implications that arise from Study 5 (see Table 10). Participants in the treatment group made their decisions with more consistency than did those in the control group, as measured by R^2, Ms = .77 and .60, $t(135) = 3.30$, $p < .002$, for the treatment

and control groups, respectively. Importantly, they also rated their performance more positively on all self-perception measures, even though, according to an objective criterion (the number of items they solved correctly), they were doing worse than the control group.

A mediational analysis confirmed that it was the heightened consistency in the treatment group that led them to be more bullish about their performance. The assignment of participants into the treatment or the control group was significantly related to how consistent their decisions were and how favorably they rated their performance, according to the composite self-perception measure (see Figure 10). Consistency was also significantly related to self-perceptions of performance, even after controlling for a participants' experimental condition. The link between experimental condition and self-perceptions, however, was broken after controlling for consistency, Sobel $z = 2.96$, $p < .005$. In short, the manipulation used in Study 5, which gave participants a chance to adopt a rule if they wished to, had a direct impact on judgmental consistency, which led to a rise in how highly participants evaluated their performance.

Neglect of Alternatives

Did participants who were using rational rules displaying more confidence because they were neglecting alternative solutions to the problems they faced? For each participant in Study 6, we counted the total number of cards across the 10 Wason problems that the participant reported seriously considering handling in a different way. Data from 5 participants were omitted because responses on the questions regarding alternative solutions indicated that they did not understand the instructions about what to report. The resulting distribution of number of cards considered was quite positively skewed (skew = 1.14). Thus, for each participant, we took the natural log of (n + 1), where n is the number of cards considered. We added 1 because it meant that a participant who turned over zero cards would have a definable transform (i.e., something

greater than 0).

This measure of the consideration of alternatives was negatively correlated with actual performance (see Table 11). That is, better performing participants reported considering fewer alternatives than poorer performing participants. Because of this, we controlled for actual performance before correlating this consideration of alternatives measure with our consistency and self-evaluation measures. For percentile measures of self-evaluation, we controlled for the percentile of actual performance. For estimates of raw score and average confidence, we controlled for the participants' actual raw score. For the composite self-evaluation measure and consistency, we controlled for a composite measure of actual performance (in which we standardized the raw and percentile score of participants and then averaged the two). As seen in Table 11, consideration of alternatives was significantly correlated with all relevant measures. To the extent that participants reported considering a greater number of alternatives, they were significantly less consistent in their approach and more negative in their self-evaluations.

This suggested that the (lack of) consideration of alternatives might emerge as a mediator of the link between consistency and self-evaluations. To test for mediation more formally, we followed the basic steps for assessing mediation (Baron & Kenny, 1986). As is evident in the top panel of Figure 11, consideration of alternatives successfully passes the tests for mediation. First, consistency is positively correlated with favorable self-evaluations, as indexed by the composite self-evaluation measure, and also negatively with consideration of alternatives. Controlling for the consistency measure, consideration of alternatives is significantly and negatively correlated with the favorability of self-evaluations, and the original relationship between consistency and self-evaluation is significantly weakened, Sobel $z = 3.10$, $p < .005$, although the relationship between consistency and self-evaluation remained significant—

indicating that neglect of alternatives served as a partial but not full mediator of the consistency/self-evaluation link. All of these analyses were done after controlling for actual performance (as assessed by an overall composite measure).

As a second test of mediation, we also assessed the link between explicitly endorsing an overall strategy and the favorability of self-evaluation. Replicating Study 2, and as seen in the bottom panel of Figure 11, those who explicitly endorsed a strategy reported more favorable self-evaluations, assessed by the composite measure. Endorsing a strategy was also negatively correlated with considering alternative solutions. Controlling for explicit endorsement of a strategy, consideration of alternatives was still correlated negatively with self-evaluations, and the initial relationship between strategy endorsement and self-evaluations was significantly weakened, Sobel $z = 2.73$, $p < .01$. However, the link between consistency and self-evaluation remained marginally significant, indicating partial but not full mediation. Again, all these analyses were conducted after controlling for actual performance.

General Discussion

At the beginning of this manuscript, we asked why the relationship between confidence and accuracy is so tenuous. How people can be so confident in decisions and judgments that are, in fact, uniformly erroneous? Data from six studies suggested that people are confident in their conclusions to the extent that those conclusions are reached via a systematic algorithm or procedure that is used faithfully across a class of decision problems. In each of the six studies, participants who approached problems in a consistent, rule-based way were more confident of their answers than were those who answered in a more variable or *ad hoc* fashion.

This increase in confidence emerged when consistency was related to accuracy, but it also arose when participants used the wrong algorithm and ran toward error. This pattern

ultimately revealed that consistency for some participants led to worse performance. Paradoxically, among poor performers, decision consistency was associated with increased confidence in their performance. As such, the data suggest a mechanism underlying the Dunning-Kruger effect, or the lack of recognition of poor performance among those providing it (Kruger & Dunning, 1999). If poor performers follow a rule that leads to error while simultaneously and ironically leading to confidence, then we can explain why and who among poor performers tends to miss recognizing just how deficient their performance is. Their reliance on a rational rule provides a "veil" that hides from them the reality of their failure.

Further study revealed that people were largely conscious of the rules they were following. In Study 4, participants who stated they followed the right rule in solving Wason problems (choosing the P and not-Q cards) tended to choose the correct cards. Those stating they were followed popular rules that were wrong tended to choose the cards they said they were choosing. In addition, following a stated rule did lead to decision consistency, which in turn led to confidence, whether or not it was well-placed.

Further study also showed that decision consistency played a causal role in producing confidence, whether appropriate or not. In Study 5, when participants were asked whether there was an overarching rule governing solutions in the Wason task, most decided that there was. They then were more consistent in their responses to the task and more confident—and they were more confident to the extent that they adopted a rule-based approach to their solutions.

Taken together, these studies provide coherent evidence about the role of systematic errors in self-evaluations of performance. That said, we should note that although our findings differ from the conclusions of classic work in psychology, we do not wish to have our work read as a refutation of that work. That classic work from the Brunswickian (1955, 1956) tradition and

the comparison between clinical and statistical thinking (Dawes, 1979; Dawes et al., 1989) suggests that decision consistency, in general, leads to accurate judgment. We submit that this general conclusion is likely correct, but we hasten to add an important asterisk. At times, when paired with inapt rules for judgment, decision consistency may lead to the opposite of accuracy, coupled with decided and inappropriate confidence. As such, one must be on guard to monitor which camp one might be in.

Underlying Mechanisms

Finally, Study 6 successfully implicated the neglect of alternatives as a plausible mechanism underlying the link between decision consistency and positive self-evaluations. Participants who followed rules faithfully tended not to consider alternative solutions, relative to their peers who varied the solutions they arrived at, and this neglect of alternatives was significantly correlated with increased confidence in one's performance. In short, neglecting alternative solutions mediated the relationship between decision consistency and favorable self-views of performance.

This mediation, however, was only partial, suggesting that there may be other links, yet to be tested, that link decision consistency with confidence. As such, studying this link is valuable grist for future research. We can think of two different mechanisms, potentially related, that might inspire the link. Both suggest that the impact of consistency on confidence might also run through more experiential routes.

First, using a decision rule may change the "look and feel" of a decision, making one's final choice seem more plausible. For example, using a singular rule might make decisions more fluent. The decision may be made more speedily and without conflict or effort—and the experience of this fluency that would likely leave people more confident in their choices

(Benjamin & Bjork, 1996; Kelley & Lindsay, 1993; Koriat, 2012). Second, using a rule repeatedly may make feel it more familiar, and research has shown that familiarity with the terms in a decision breeds positivity and confidence in one's responses (Arkes, Boehm, & Xu, 1991; Schwartz & Metcalfe, 1992). For example, repeating declarative knowledge (i.e., facts) makes people view it as more familiar (Boehm, 1994) and, thus, more likely to be true, whether or not it is actually true (Bacon, 1979; Gigerenzer, 1984). Perhaps what is true of declarative knowledge also proves true for procedural knowledge. Repeatedly bringing a rule to mind may make it feel familiar, and thus valid.

Being Misinformed Versus Uninformed

Taken together, the six studies provide one answer for the emergence of the Dunning-Kruger effect (Dunning, 2011; Kruger & Dunning, 1999), in which people making a surfeit of mistakes seem to have little awareness of just how many mistakes they are making. Dunning and colleagues have argued why people should have little insight into the fact that their performances have been deficient (see Dunning, 2011), but have had less to say specifically about what cues people use to reach impressions that they are doing so well. The data across these six studies suggest that people think they are doing well when they are following a consistent rule. Regrettably, following a flawed rule seems to produce almost as much confidence as following the right one. This means confident individuals might be correct, but they also might exactly be sufferers of the Dunning-Kruger effect. Rule-based confidence is no guarantee of self-insight into performance.

With this in mind, it may be profitable for future research to discuss further the important distinction between being *uninformed* and *misinformed*. People tend to think of people with low knowledge as being uninformed, believing that people with low expertise have few facts and

ideas in mind about how to approach relevant tasks or issues. Their heads are empty voids about the topics they know little about. To be sure, there are clearly topics where this notion of the uninformed holds true (Caputo & Dunning, 2005). People do not know much about Herbrand universes, maxillofacial surgery, or astronautical engineering.

Recent research across several fields, however, increasingly suggests that people in many areas of life are not so much uninformed as they are misinformed. When they know little about a topic, their head is no empty void. Instead, it actually contains a good deal of material it can use to answer questions put to it—the material just tends to be false, misleading, unsubstantiated, or irrelevant. In our experiments, many but not all participants certainly had what they considered to be a bit of knowledge about how to approach Wason selection tasks. Those who felt they "knew" the one true rule to address these problems were more confident in their answers than those without this knowledge, and who instead approached each individual question as unique.

The distinction between the uninformed and the misinformed is an important issue in many areas of life. Political scientists, for example, have become increasingly aware that voters tend to hold confident but misguided beliefs about social conditions and political policy (Bartels, 1996; Kuklinski, Quirk, Jerit, Schwieder, & Rich, 2000; Oppenheimer & Edwards, 2012). Such misinformed views matter, in that they change voter's political preferences and leave people resistant to more accurate information about history, social conditions, and policy proposals (Kuklinski et al., 2000; Nyhan & Riefler, 2010).

In psychology, other evidence suggests that non-experts are often better characterized as misinformed rather than uninformed. People without an education in science have been shown to have systematic and strong ideas about the physical world works. Work on intuitive physics (McClosky, 1983; Hespos, & vanMarle, 2012; Proffitt & Gilden, 1989), optics (Croucher,

Bertamini, & Hecht, 2002), and biology (Inagaki & Hatano, 2006) all show that so-called unknowledgeable people often have ready-made hunches that may seem reasonable but that are instead incorrect. The hunches may lie in full-blown theories (McCloskey, 1983) or they may instead be just fragments of notions that are stitched together when relevant (diSessa, 1988), but these intuitions can lead people to make systematic mistakes when they try to predict the behavior of physical objects in the world. These misguided hunches also make teaching more correct models of the physical world more difficult to achieve (Carey, 2000; Keil, 2011), and may never be fully erased (Shtulman & Valcarcel, 2012).

Concluding Remarks

In sum, although our data suggest that a wise consistency can be the fount of accurate judgment, but they also suggest that a foolish consistency, as Ralph Waldo Emerson once observed, may serve as a hobgoblin for little minds. Across our studies, it was also a foolish consistency that gave metaphorical swelled heads to some of those little minds as they evaluated the quality of their work.

As for Roy Riegels, after being convinced by his coach to play the second half of the Rose Bowl game, he blocked a Georgia Tech punt and was named captain of his football team the following season. He served honorably in the United Stated Army Air Force during World War II and then started his own agricultural chemicals firm. He even coached football at the high school and college level. In 1993, he died at the age of 84, survived by four children, and was inducted five years later into the University of California, Berkeley, Hall of Fame. His story is often taken as a lesson that, even after humiliating error, time goes on to provide many chances for redemption, success, and a life well lived.

References

Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology, 27*, 576–605.

Arkes, H., R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology, 73*, 305-307.

Ashlock, R. B. (1976). *Error patterns in computation*. Columbus, OH: Bell & Howell.

Austin, Z., Gregory, P. A. M., & Galli, M. (2008). "I just don't know what I'm supposed to know": Evaluating self-assessment skills of international pharmacy graduates in Canada. *Administrative Pharmacy, 4*, 115-124.

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 241-252.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bartels, L. (1996). Uninformed votes: Information effects in Presidential elections. *American Journal of Political Science, 40*, 194-230.

Ben-Zeev, T. (1995). The nature and origin of rational errors in arithmetic thinking: Induction from examples and prior knowledge. *Cognitive Science, 19*, 341-376.

Ben Zeev, T. (1998). Rational errors and the mathematical mind. *Review of General Psychology, 2*, 366-383.

Ben Zeev, T., & Star, J. R. (2001). Spurious correlations in mathematical thinking. *Cognitive and Instruction, 19*, 253-275.

Benjamin, A. S. & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M.

Reder (Ed.), *Implicit memory and metacognition* (pp. 309-338). Mahwah, NJ: Erlbaum.

Benzion, U., Granot, A., & Yagil, J. (1992). The valuation of the exponential function and implications for derived interest rates. *Economics Letters, 38*, 299–303.

Berlin, I. (2953). *The hedgehog and the fox*. London: Weidenfeld & Nicolson.

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin, 20*, 285-293.

Brown, J. S., & VanLehn, D. (1980). Repair theory: A generative theory of bugs in procdural skills. *Cognitive Science, 4*, 379-426.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193-217.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

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

Caputo, D., & Dunning, D. (2005). What you don't know: The role played by errors of omission in imperfect self-assessments. *Journal of Experimental Social Psychology, 41*, 488-505.

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Development, 21*, 13-19.

Christandl, F., & Fetchenhauer, D. (2009). How laypeople and experts misperceive the effect of economic growth. *Journal of Economic Psychology, 30*, 381-392.

Croucher, C. J., Bertamini, M., & Hecht, H. (2002). Naïve optics: Understanding the geometry of mirror reflections. *Journal of Experimental Psychology: Human Perception and*

Performance, 28, 546-562.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Contructivism in the computer age* (pp. 49-70). Hillsdale, NJ: Erlbaum.

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

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (vol. 44, pp. 247-296). New York: Elsevier.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69-106.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12,* 83-86.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware? Further explorations of (lack of) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105,* 98-121.

Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology, 97*, 185-195.

Gosling, S.D., Ko, S.J., Mannarelli, T., & Morris, M.E. (2002). A room with a cue: Judgments of personality based on offices and bedrooms. *Journal of Personality and Social*

Psychology, 82, 379-398.

Hall, C.C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes, 103*, 277-290.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*, 438-456.

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

Hespos, S. J., & vanMarle, K. (2012). Physics for infants: characterizing the origins of knowledge about objects, substances, and number. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3,* 19-27.

Inagaki, K., & Hatano, G. (2006). Young children's conception of the biological world. *Current Directions of Psychological Science, 15*, 177-181.

Keil, F. C. (2011). Science starts early. *Science, 331*, 1022-1023.

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

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 461-469.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principles. *Journal of Experimental Psychology: learning, Memory, and Cognition, 34*, 945-949.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*, 80-113.

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

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The contribution of social-perceptual skills and statistical regression to self-enhancement biases. *Journal of Personality and Social Psychology, 82,* 180-188.

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

Kruger, J., & Dunning, D. (2002). Unskilled and unaware—But why? A reply to Krueger and Mueller. *Journal of Personality and Social Psychology*, *82*, 189-192.

Kruglanski, A. (1980). Lay epistemologic process and contents: Another look at attribution theory. Psychological Review, 87, 70-87.

Kruglanski, A. W. (1989). *Lay epistemics and human knowledge: Cognitive and motivational bases*. New York: Plenum.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *Journal of Politics, 62*, 79-816.

Lord, C.G., Lepper, M.R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*, 1231-1243.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 248*, 122-130.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science*, *210*, 1139-1141.

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 771-792.

Mynttinsen, S., Sundstrom, A., Vissers, J., Koivukoski, M., Hakuli, K., & Keskinen, E. (2009). Self-assessed driver competence among novice drivers: A comparison of driving test candidate assessments and examiner assessments in a Dutch and Finnish sample. *Journal of Safety Research, 40*, 301-309.

Nisbett, R., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes." *Psychological Review 84*, 231-259.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misconceptions. *Political Behavior, 32*, 303-330.

Oppenheimer, D., & Edwards, M. (2012). *Democracy despite itself: Why a system that shouldn't work at all works so well*. Cambridge, MA: MIT Press.

Park, Y.-J., & Santos-Pinto, L. (2010). Overconfidence in tournaments: Evidence from the field. *Theory and Decision, 69*, 143-166.

Proffitt, D.R. & Gilden, D.L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 384-393.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). Naïve physics reasoning: A commitment to substance-based conceptions. *Cognition and Instruction, 18*, 1-34.

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

Simons, D. (in press). Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin & Review*.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209-215.

Tetlock, P.E. (2005). *Expert political judgment*: *How good is it? How will we know?* Princeton, NJ: Princeton University Press.

Tucker, L. R. (1964). A suggested alternative formulation in the development of Hursch Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review,* 71, 528-530.

VanLehn, K. (1983). On the representation of procedures in repair theory. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp.201-252). Hillsdale, NJ: Erlbaum.

Wagenaar, W. A., & Timmers, H. (1979). The pond-and-duckweed problem: Three experiments on the misperception of exponential growth. *Acta Psychologica, 43*, 239–251.

Wason, P. C. (1960), On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology,12*, 129-140.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Baltimore: Penguin Press.

Wason, P. C. (1977). Self-contradictions. In P. N. Johnson-Laird & P.C. Wason (eds.) Thinking: Readings in cognitive science. (pp. 114-128). Cambridge: Cambridge University Press.

Zuckerman, M., Koestner, R., & Driver, R. (1981). Beliefs about cues associated with deception. *Journal of Nonverbal Behavior, 6*, 105-114.

Footnotes

[1]We use the term "rational" the way it has been used in the educational psychology literature to indicate rules or procedures that lead to systematic error (Ben Zeev, 1995, 1998; Brown & VanLehn, 1980; VanLehn, 1983). In using the term this way to establish continuity between that past work and this research program, we do not wish to connote that such errors or procedures are rational in a normative or adaptive sense.

Table 1. *Average Self-Evaluation Ratings Versus Actual Performance*.

	Study					
Measure	1	2	3	4	5	6
Self-Evaluation Measure						
Ability Percentile	63.8	64.7	64.5	67.0	68.6	69.0
Test Score Percentile	60.6	63.7	59.7	64.4	68.8	67.6
Raw Score Estimate (%)	65.6	72.3	70.8	67.8	73.7	69.9
Average Item Confidence	--	76.3	73.1	82.1	84.0	82.3
Actual Raw Score (%)	49.1	56.3	42.7	23.5	18.0	23.1

Note: All self-evaluation measure averages exceed their objective metric by at least $t = 5.32$, $ps < .0001$.

Table 2. *Linear and Quadratic Components (β) of the Relationship Between Objective Performance and Perceived Performance*.

	Regression Component	
Study/Measure	Linear	Quadratic
Study 1		
Ability Percentile	.38**	.07
Test Score Percentile	.40**	.26**
Raw Score Estimate (%)	.47**	.21**
Study 2		
Ability Percentile	.33**	.18
Test Score Percentile	.19**	.13
Raw Score Estimate (%)	.29**	.36**
Average Item Confidence	.44**	.23**
Study 3		
Ability Percentile	.24**	.42**
Test Score Percentile	.29**	.33**
Raw Score Estimate (%)	.02	.60**
Average Item Confidence	.14	.33**
Study 4		
Ability Percentile	.17*	.07
Test Score Percentile	.20**	.04
Raw Score Estimate (%)	-.04	.41**
Average Item Confidence	.27**	-.01
Study 5		
Ability Percentile	-.29**	.36**
Test Score Percentile	-.32**	.35**
Raw Score Estimate (%)	-.82**	.91**
Average Item Confidence	-.63**	.70**
Study 6		
Ability Percentile	.13	.23**
Test Score Percentile	.06	.39**
Raw Score Estimate (%)	-.07	.52**
Average Item Confidence	-.04	.40*

*$p < .10$ ** $p < .05$

Table 3. *Relationship between Consistency* (R^2) *and Self-Evaluation before and after controlling for objective performance*.

		Multiple Regression	
Study/Measure	Consistency	Consistency	Objective
Study 1			
Ability Percentile	.30**	.16*	.31**
Test Score Percentile	.46**	.36**	.24**
Raw Score Estimate (%)	.59**	.46**	.27**
Study 2			
Ability Percentile	.20**	.12	.11
Test Score Percentile	.28**	.15	.20
Raw Score Estimate (%)	.43**	.45**	-.02
Average Item Confidence	.48**	.36**	.20*
Study 3			
Ability Percentile	.52**	.45**	.23**
Test Score Percentile	.43**	.34**	.30**
Raw Score Estimate (%)	.68**	.63**	.18**
Average Item Confidence	.49**	.43**	.20**
Study 4			
Ability Percentile	.37**	.33**	.13
Test Score Percentile	.34**	.30**	.15
Raw Score Estimate (%)	.49**	.43**	.17**
Average Item Confidence	.31**	.25**	-.19**
Study 5			
Ability Percentile	.44**	.44**	.11
Test Score Percentile	.43**	.43**	.08
Raw Score Estimate (%)	.66**	.66**	.03
Average Item Confidence	.49**	.49**	.09
Study 6			
Ability Percentile	.39**	.37**	.15
Test Score Percentile	.43**	.41**	.12
Raw Score Estimate (%)	.55**	.48**	.22**
Average Item Confidence	.35**	.29**	.21*

*$p < .10$ ** $p < .05$

Note: Model 1 examines the zero-order relationship between consistency and self-evaluation measures. Model 2 examines simultaneously the relationships of consistency and actual performance to self-evaluation.

Table 4. *Linear and Quadratic Components (β) of the Relationship Between Consistency and Objective Performance*.

		Quadratic Regression	
Study/Measure	Linear Only	Linear Component	Quadratic Component
Study 1			
Test Score Percentile	.46**	.47**	.58**
Raw Score	.51**	.42**	.56**
Study 2			
Test Score Percentile	.64**	.65**	.65**
Raw Score	.61**	.68**	.69**
Study 3			
Test Score Percentile	.17*	.12*	.73**
Raw Score	.29**	-.20**	.86**
Study 4			
Test Score Percentile	.11	-.01	.44**
Raw Score	.31**	-.34**	.77**
Study 5			
Test Score Percentile	-.00	-.73**	.61**
Raw Score	-.05	-1.11**	1.17**
Study 6			
Test Score Percentile	.15	.01	.45**
Raw Score	.33**	-.42**	.87**

**$p < .05$

Table 5. *Relation of Consistency (β) to Self-Evaluations Among Poorest Performers*.

Self-Evaluation Measure	Study					
	1	2	3	4	5	6
Ability Percentile	.20	.05	.44**	.29**	.41**	.49**
Test Score Percentile	.40**	.08	.50**	.28*	.39**	.46**
Raw Score Estimate (%)	.45**	.34**	.65**	.38**	.64**	.43**
Average Item Confidence	--	.40**	.55**	.22	.46**	.26
Composite	.40**	.25	.59**	.35**	.53**	.49**

Note: Composite score represents average of all other self-evaluation measures after standardization.

*$p < .10$ ** $p < .05$

Table 6. *Self-Evaluations of Completely Consistent Participants Getting All Items Right Versus All Wrong*.

	Participant Group		
Study/Measure	All Right	All Wrong	*t*
Study 1			
Ability Percentile	76.0	68.1	1.05
Test Score Percentile	79.2	66.1	1.73*
Raw Score Estimate (%)	89.0	83.0	1.01
Composite	.84	.40	1.47
Study 2			
Ability Percentile	72.6	68.3	0.29
Test Score Percentile	75.2	78.3	-0.25
Raw Score Estimate (%)	86.7	94.5	-0.70
Average Item Confidence	87.6	83.6	0.51
Composite	.56	.55	-0.02
Study 3			
Ability Percentile	83.8	73.0	1.53
Test Score Percentile	76.5	66.8	1.23
Raw Score Estimate (%)	96.3	86.5	2.20**
Average Item Confidence	87.8	81.7	1.30
Composite	.72	.38	1.83*
Study 4			
Ability Percentile	86.0	76.8	1.32
Test Score Percentile	85.0	73.2	1.07
Raw Score Estimate (%)	96.0	84.5	1.54
Average Item Confidence	92.0	83.8	1.22
Composite	.88	.41	1.59
Study 5			
Ability Percentile	88.0	79.1	.83
Test Score Percentile	87.0	81.1	.53
Raw Score Estimate (%)	100.0	92.9	1.05
Average Item Confidence	99.4	90.5	1.30
Composite	.87	.52	1.06
Study 6			
Ability Percentile	87.2	84.4	.44
Test Score Percentile	93.0	82.2	2.13*
Raw Score Estimate (%)	100.0	90.0	2.22*
Average Item Confidence	97.6	85.4	2.37*
Composite	1.01	.56	2.52*

Note: Composite score represents average of all other self-evaluation measures after standardization.
*$p < .10$ ** $p < .05$

Table 7. *Consistency and Self-Evaluations of Participants Who Stated They Did or Did Not Follow an Explicit Rule in Study 4.*

	Followed Explicit Rule		
Measure	Yes	No	*t*
Consistency (R^2)	.65	.42	4.47**
Self-Evaluation Measures			
Ability Percentile	71.1	61.0	2.68**
Test Score Percentile	69.2	57.5	2.77**
Raw Score Estimate (%)	74.4	58.2	3.73**
Average Item Confidence	84.1	79.3	1.55
Composite	.20	-.28	3.15**
Objective Performance			
Percentile	48.4	51.1	-0.50
Test Score	16.0	2.88	-2.23**

Note: Composite score represents average of all other self-evaluation measures after standardization.

** $p < .05$

Table 8. *Variance Explained in Decisions by Explicit Rule Participants Stated They Were Following (Study 4).*

	Explicit Rule Endorsed			
Variance Accounted For By	Correct	Matching	P-Only	*F*(1,61)
Correct Rule	.61	.05	.24	41.56**
Matching Rule	.12	.51	.32	12.96**
P-Only Rule	.22	.27	.66	34.10**
n	26	28	8	

Note: *F*-test is the result of a contrast that pits the variance accounted for by the rule participants said they were following (weighted +2) against the two alternatives (both weighted -1).

** $p < .05$

Table 9. *Consistency, Self-Evaluations, and Objective Performance of Participants Following a Correct, Incorrect, or No Rule (Study 4).*

	Rule Followed			
Measure	Correct	Incorrect	None	$F(1,61)$
Consistency (R^2)	$.75_a$	$.64_a$	$.42_b$	27.78**
Self-Evaluation Measures				
Ability Percentile	77.7_a	69.7_a	61.0_b	11.35**
Test Score Percentile	77.7_a	66.3_b	57.5_b	12.01**
Raw Score Estimate (%)	83.5_a	70.3_b	58.2_c	18.16**
Average Item Confidence	89.3_a	$84.0_{a,b}$	79.3_b	6.15**
Composite	$.53_a$	$.11_b$	$-.29_c$	16.96**
Objective Performance				
Percentile	81.1	33.5	47.8	
Test Score	67.3	6.7	16.0	
n	26	36	50	

Note: *F*-test is the result of a contrast comparing the correct and incorrect rule followed groups (both weighted +1) against the no rule followed group (weighted -2). Means in any row not sharing the same subscript are significantly different from one another according to *t*-test. Composite score represents average of all other self-evaluation measures after standardization.

**$p < .05$

Table 10. *Consistency and Self-Evaluations of Participants in Rule Induction and Control Conditions of Study 5*.

	Rule Induction Condition		
Measure	Yes	No	*t*
Self-Evaluation Measures			
Ability Percentile	73.1	63.7	2.26**
Test Score Percentile	73.6	63.6	2.37**
Raw Score Estimate (%)	79.7	67.2	2.89**
Average Item Confidence	88.8	79.9	2.87**
Composite	.21	-.23	2.93**
Objective Performance			
Percentile	45.6	54.7	-2.12**
Test Score	14.1	22.1	-1.54
n	71	66	

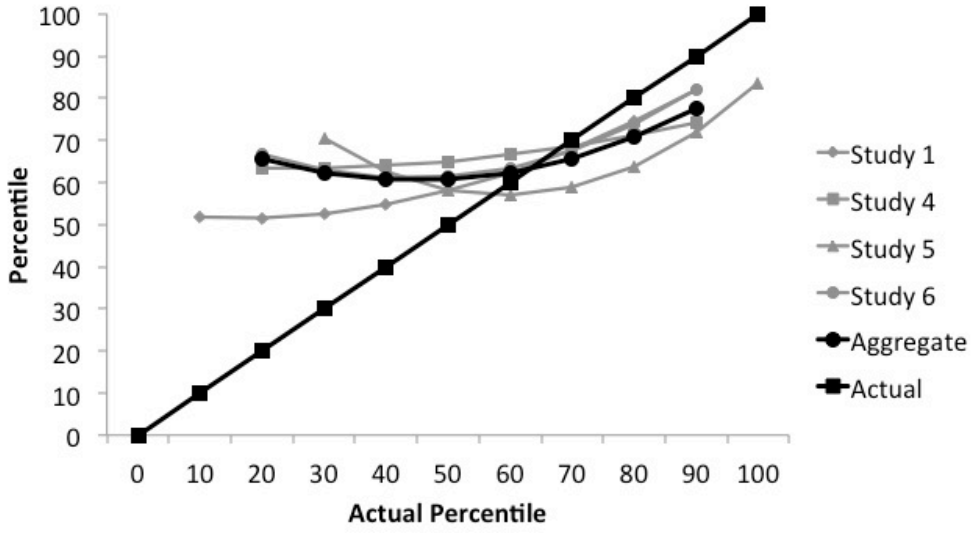
**$p < .05$

Table 11. *Correlations of Consideration of Alternatives (Study 6) with Consistency, Self-Evaluation Measures, and Objective Performance*.

Measure	Consideration of Alternatives
Consistency (R^2)	-.36**
Self-Evaluation Measures	
Ability Percentile	-.43**
Test Score Percentile	-.43**
Raw Score Estimate (%)	-.49**
Average Item Confidence	-.44**
Composite	-.53**
Objective Performance	
Percentile	-.49**
Test Score	-.41**
n	75

**$p < .05$

A. Logical Ability



B. Test Performance

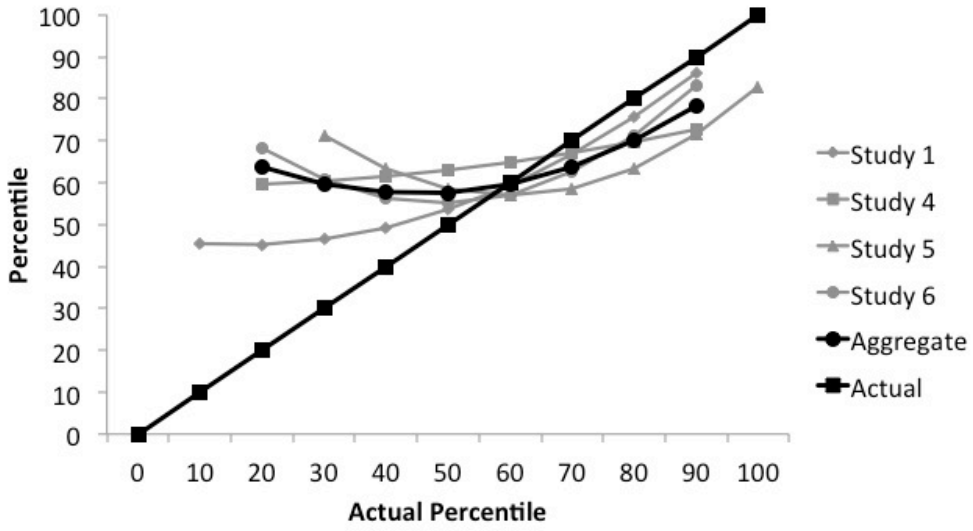
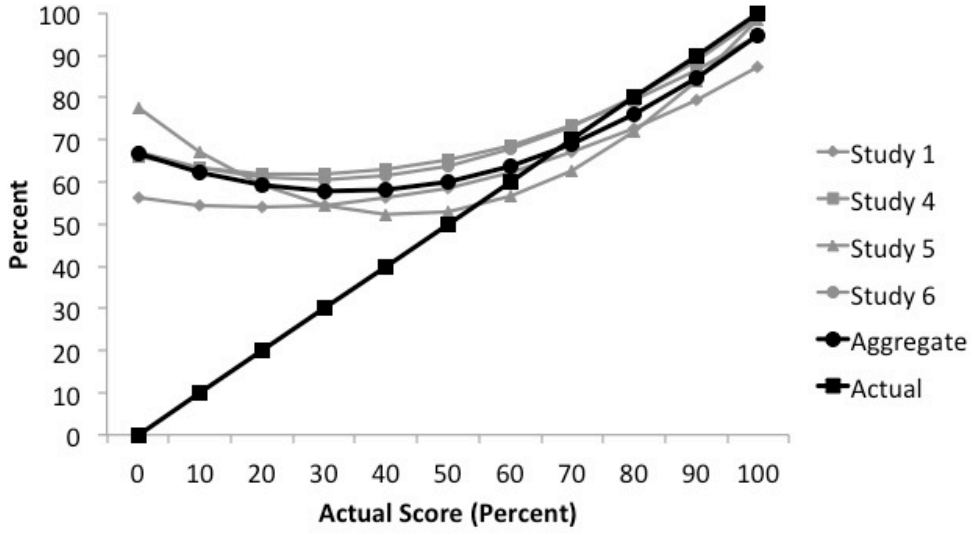


Figure 1. Relationship between actual and perceived performance in percentile terms for the Wason task (Studies 1, 4, 5, and 6). Aggregate refers to an unweighted average of predicted values across the four studies.

A. Raw Score Estimate (Percent)



B. Average Item Confidence

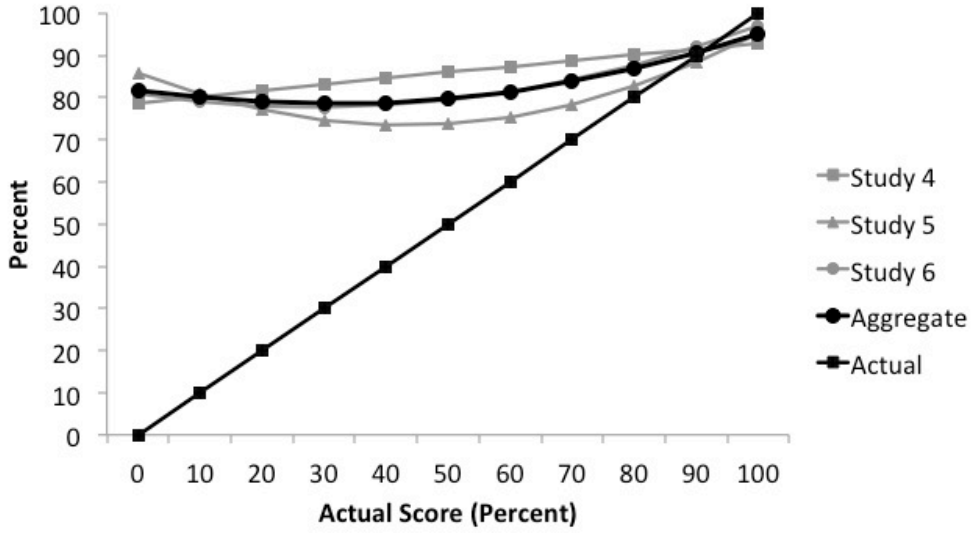
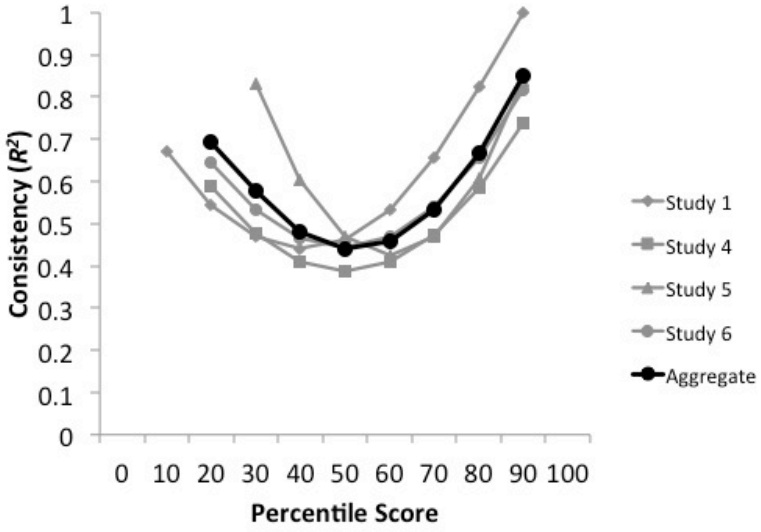


Figure 2. Relationship between actual and perceived performance on absolute measures for the Wason task (Studies 1, 4, 5, and 6). Aggregate refers to an unweighted average of predicted values across the three or four studies.

A. Percentile Measures



B. Raw Score

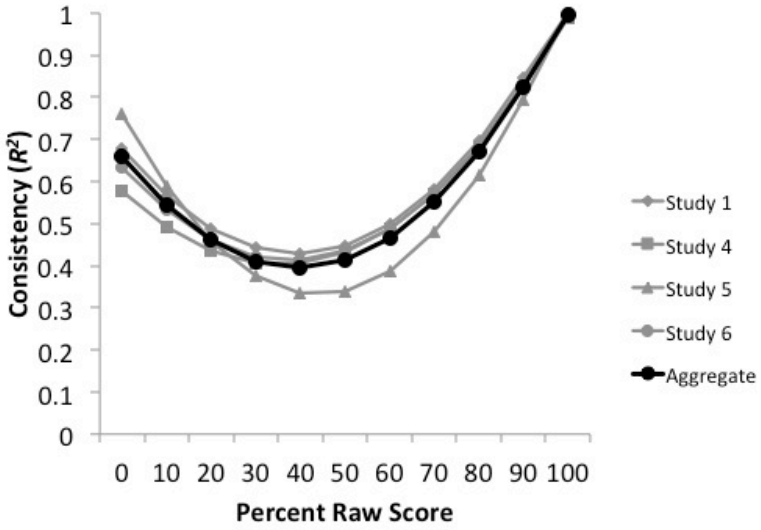


Figure 3. Relationship between actual performance (percentile and raw score) and decision consistency (R^2) for the Wason task (Studies 1, 4, 5, and 6). Aggregate refers to an unweighted average of predicted values across the four studies.

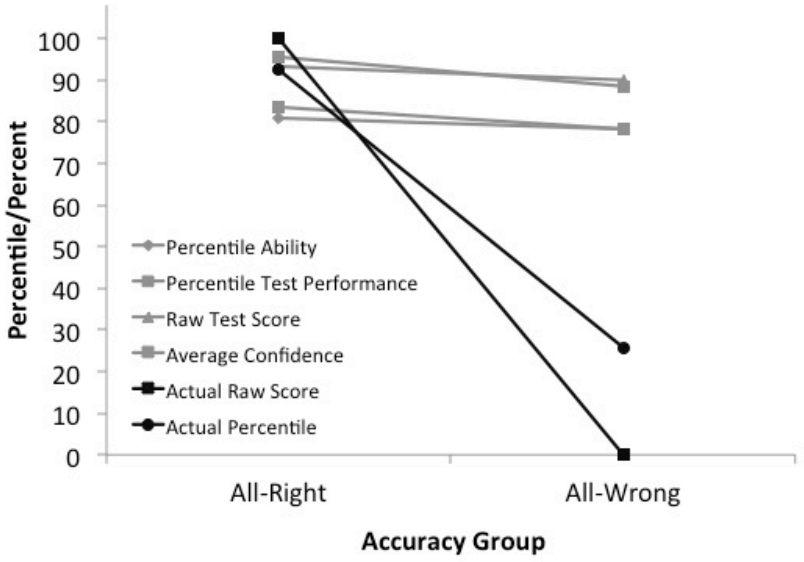
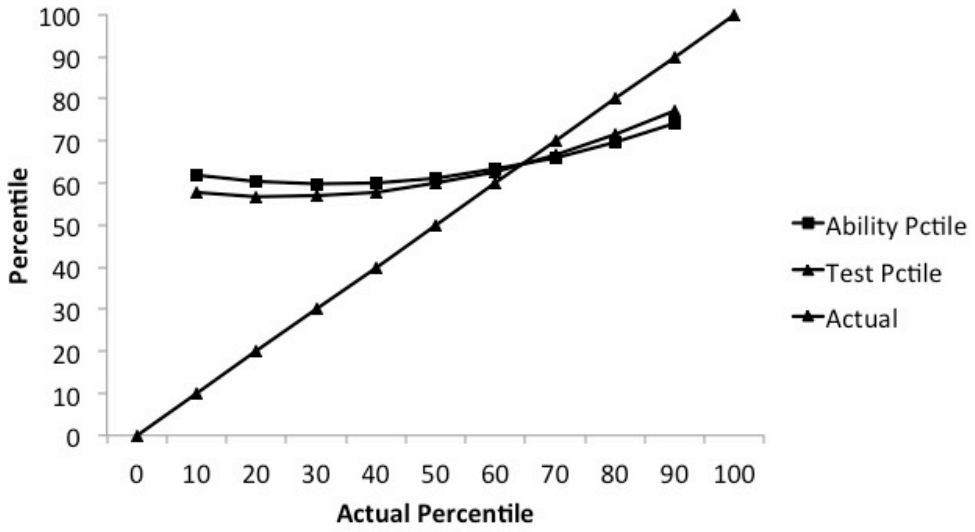


Figure 4. Among completely consistent participants (R^2 = 1.0), differences between perception of performance versus actual differences in performance for those getting all items right versus all wrong.

A. Percentile Measure



B. Raw Score

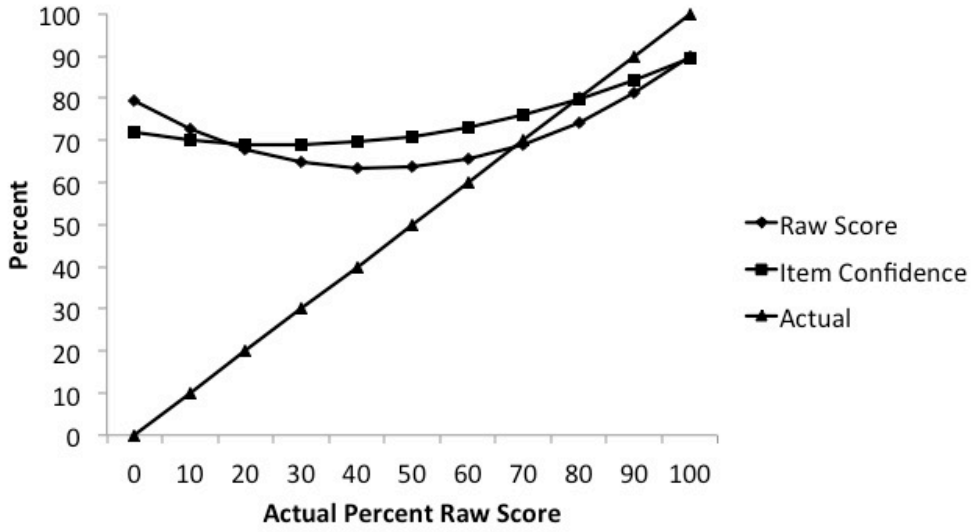
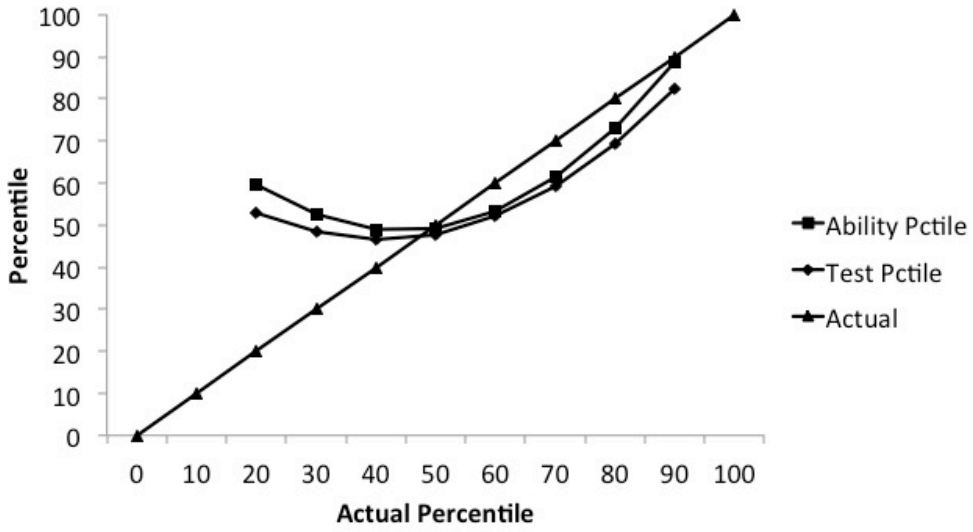


Figure 5. Relationship between perceived and actual performance (percentile and raw score) for the intuitive physics task (Study 2).

A. Percentile Measure



B. Raw Score

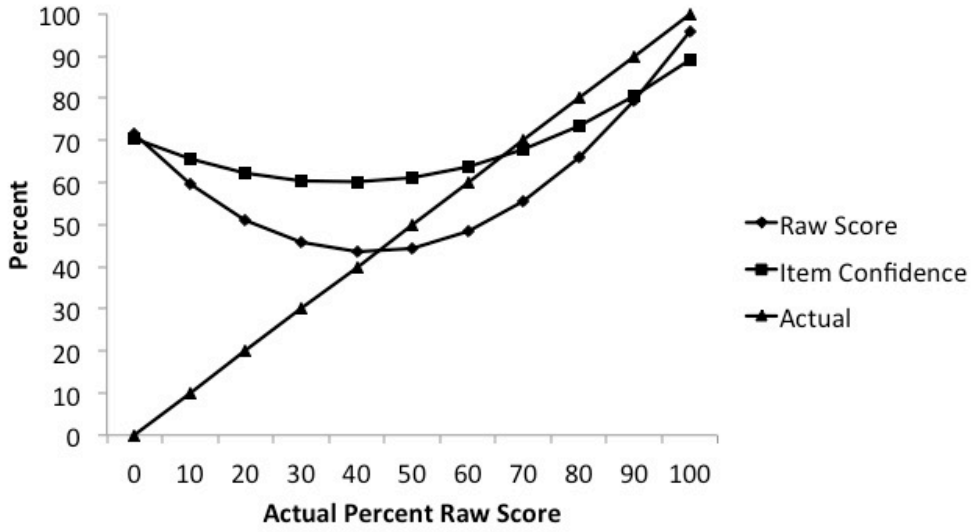
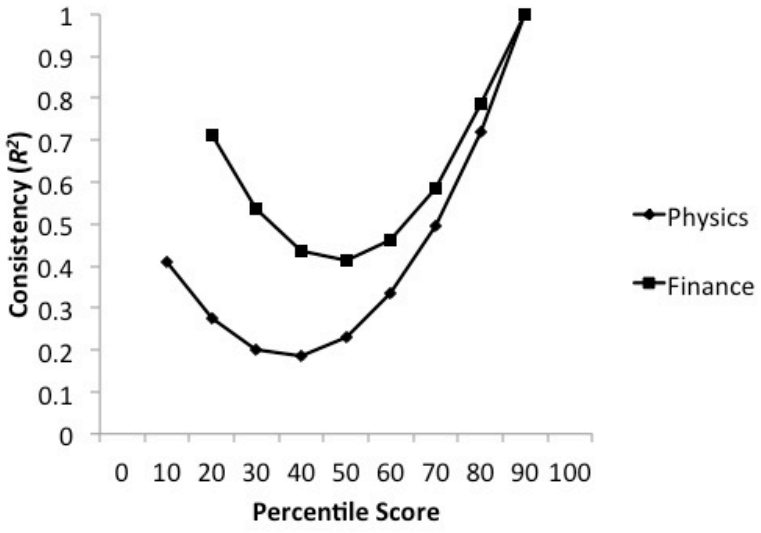


Figure 6. Relationship between perceived and actual performance (percentile and raw score) for the financial investment task (Study 3).

A. Percentile Measures



B. Raw Score

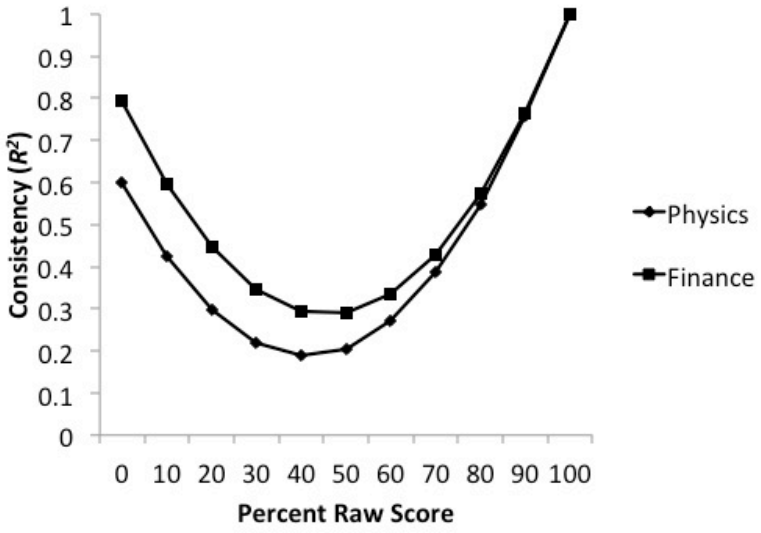


Figure 7. Relationship between actual performance (percentile and raw score) and decision consistency (R^S) for intuitive physics (Study 2) and financial investment (Study 3) tasks.

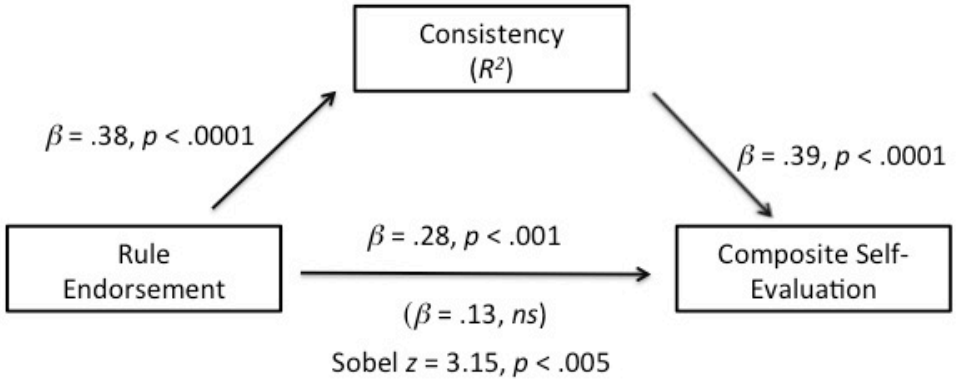


Figure 8. Mediational analysis of the link between explicitly endorsing a rule and favorability of self-evaluation of performance, with actual decision consistency as the mediator (Study 4).

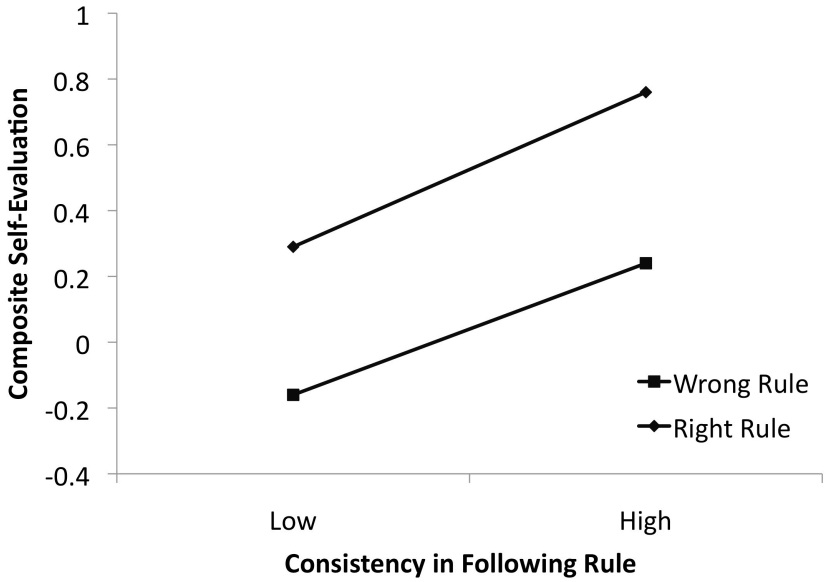


Figure 9. Relationship between decision consistency (R^2) and a composite measure of self-evaluation. Low consistency taken as -1SD and high consistency as +1SD from from overall mean (Study 4).

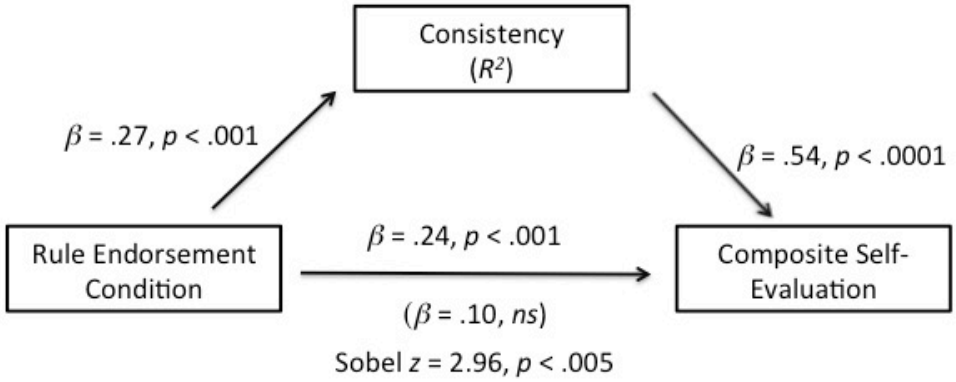
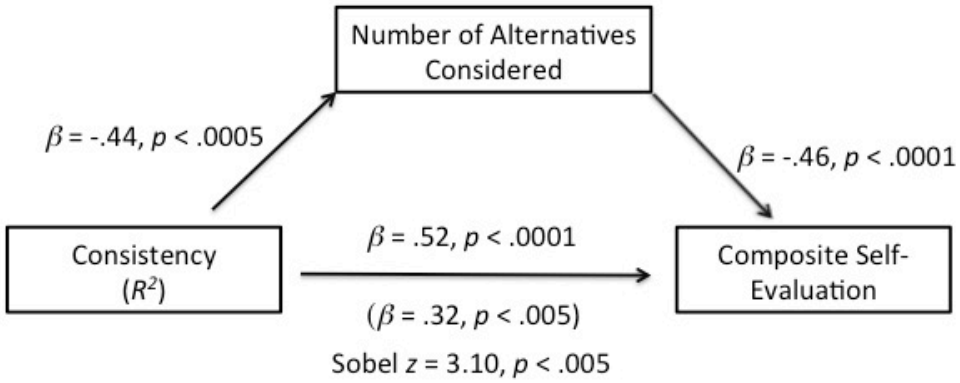


Figure 10. Mediational analysis of the link between rule endorsement condition and favorability of self-evaluation of performance, with actual decision consistency as the mediator (Study 5).

A. R^2 Used as Measure of Consistency



B. Self-Reported Endorsement of Rule Used as Measure of Consistency

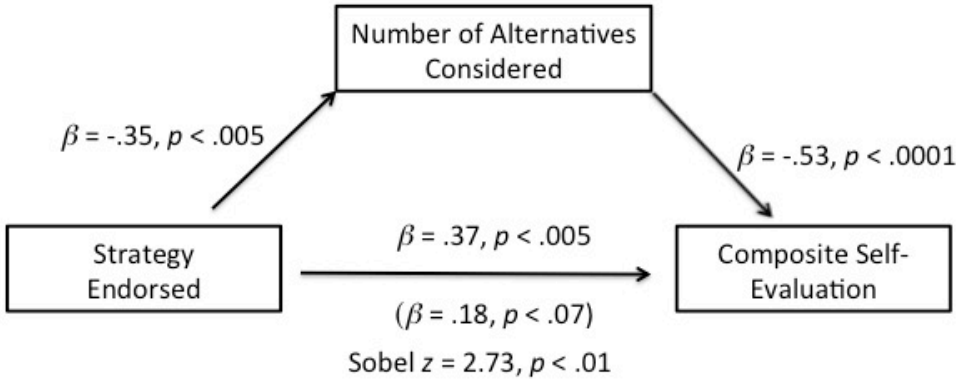


Figure 11. Mediational analysis of the link between decision consistency and favorability of self-evaluation of performance, with neglect of alternatives as the mediator. Panel A defines consistency as actual consistency (R^2), and Panel B as explicit endorsement of rule. (Study 6).