Widely-Used Measures of Overconfidence Are Confounded With Ability

Stephen A. Spiller

UCLA Anderson School of Management

October 2, 2024

Author Note

Stephen A. Spiller https://orcid.org/0000-0001-6951-6046

Data and materials from prior investigations are available at https://osf.io/6tecy/ and https://alpdata.rand.org/. All code is available at https://researchbox.org/1597&PEER_REVIEW_passcode=ORRDVP. I have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Stephen A. Spiller, UCLA Anderson School of Management, 110 Westwood Plaza, Los Angeles, CA, 90095. Email: stephen.spiller@anderson.ucla.edu

Abstract

The overconfidence concept is one of the great success stories of psychological research, influencing research in other disciplines as well as discourse in the popular press, business, and public policy. Relative to underconfidence, overconfidence at various tasks is purportedly associated with greater narcissism, lower anxiety regarding those tasks, higher status, greater savings, more planning, and numerous other differences. Yet much of this evidence may merely reflect that there are associations with ability rather than overconfidence. This results from two underappreciated properties of typical measures of overconfidence. First, performance is an imperfect measure of ability; accounting for performance does not sufficiently account for ability. Second, self-evaluations of performance should reflect ability in addition to performance; because performance is ambiguous, people should use prior beliefs about their ability. I show these uncontroversial principles imply that commonly-used measures of overconfidence are confounded with ability. I support these analytical results by reexamining previously-published results. In the first analysis, I find overconfidence predicts subsequent performance, consistent with overconfidence as a signal of ability but inconsistent with overconfidence as a bias. In the second set of analyses, I find the purported association between overconfidence and other proposed constructs can be adequately explained through ability alone. I close with recommendations on approaches to recognize and reduce the extent of the problem. This model serves as a stark reminder: when researchers propose that differences in overconfidence are associated with other behaviors, beliefs, or evaluations, they must account for the possibility that differences in ability provide a sufficient explanation.

Keywords: overconfidence, ability, knowledge, performance, measurement error

Overconfidence is widely considered to be a ubiquitous bias. It is reliably reproduced in academic research, worthy of chapters in popular business books, and labeled as "the most significant of the cognitive biases" by a founder of the heuristics-and-biases research program (Kahneman, 2011). Overconfidence refers to the state in which one's beliefs regarding one's ability on some dimension exceed one's true ability on that dimension.[1] Correlates of overconfidence on specific tasks have been widely studied. These include narcissism (Ames & Kammrath, 2004; Campbell et al., 2004; John & Robins, 1994), savings (Avdeenko et al., 2019), advice-seeking (Kramer, 2016), financial planning (Anderson et al. 2017; Parker et al., 2012), reduced language anxiety (MacIntyre et al., 1997), social status (Anderson et al., 2012), social class (Belmi et al., 2020), choice of nonlinear incentives (Larkin & Leider, 2012), susceptibility to false news (Lyons et al., 2021), search behavior (Moorman et al., 2004), stock ownership (Ke, 2021), short-term debt (Landier & Thesmar, 2008), and many more.

I prove that under reasonable assumptions, multiple widely-used measures of overconfidence are confounded with ability, and I show that this confound can be consequential. I do not argue that overconfidence, nor differences in overconfidence, nor correlates of overconfidence, do not exist. Instead, I argue that in order to claim that other measures correlate with differences in overconfidence, one must adequately address this confound.

To do this, I extend Moore and Healy's (2008) model of overconfidence to allow different people to approach the same task with different abilities and (possibly imperfect) self-insight into their ability. Moore and Healy present a model in which people's prior beliefs about a task's simplicity, combined with necessary ambiguity in assessing their own performance, lead people to overestimate performance on hard tasks, underestimate performance on easy tasks, overplace performance on easy tasks, and

[1] The definition for manifest overconfidence includes both *overestimation*, self-evaluations that overstate absolute performance, and *overplacement*, self-evaluations that overstate relative performance (Moore & Healy, 2008). The model I develop here does not directly address *overprecision*, or excessive certainty in one's knowledge. The research I focus on here uses a variety of related terms, including overconfidence, biased self-evaluations or self-assessments, unjustified confidence, inappropriate confidence, subjective knowledge when controlling for objective knowledge, and others. There are direct links to the literature on the correlates of positive self-views and self-enhancement (Taylor & Brown, 1988; Colvin et al., 1995). I return to these at the end of the paper.

underplace performance on hard tasks. Whereas their model explains these patterns of estimation and placement across tasks, my extension considers differences across a sample on a given task.

The key issue is that although researchers intend to account for latent ability, they instead account for observed performance. If people are appropriately-sensitive to even imperfectly-calibrated prior beliefs (e.g., incorporate base rates and use Bayesian inferences), their self-evaluations will reflect ability directly, not just through performance. The resulting associations between measures of overconfidence (which account for performance but not ability) and other constructs are therefore systematically biased. Although many reports claim to find evidence that *overconfidence* on a task is associated with various correlates, their evidence may instead indicate that *ability* on a task is associated with those correlates. This confound frequently escapes notice, perhaps because in contrast to other applications of using a noisy covariate, when studying overconfidence, (a) it is not always apparent that performance is merely a measure of ability rather than the focus of inquiry, and (b) it is not always apparent that evaluations ought to regress to one's prior beliefs. This confound can be particularly pernicious because ability is often considered and explicitly ruled out as an alternative explanation of the results based on how the overconfidence measure is constructed.

I begin by describing a typical paradigm used to measure differences in overconfidence, variations on that theme, and why this results in a problem. I next present a mathematical model to formalize and quantify this bias. I then examine whether these theoretical predictions hold in data using previously collected datasets. First, I establish the confound generates correlations where we expect none exist. Using data from Moore and Healy (2008), I find measures of overconfidence predict subsequent performance, consistent with an account in which measures of overconfidence are confounded with ability. Second, I establish the confound could plausibly account for reported findings. Using data from Anderson et al. (2012) and Belmi et al. (2020), I reexamine the relationship between overconfidence and proposed causes and consequences. I find that models in which there is no overconfidence are sufficient to explain the full data pattern. I close with recommendations to recognize and ameliorate the problem, even if eliminating the possibility of the problem in all its forms may be a nearly unattainable target.

Theme and Variations: Measuring Differences in Overconfidence

Research on correlates of overconfidence has used an extensive array of measures. I consider cases of overestimation and overplacement (but not overprecision) in which there is a reality criterion against which to compare. When using a single measure of performance and a single self-evaluation, there are at least 20 different ways that overconfidence may be calculated; adding cases in which self-evaluations reflect future expectations rather than past performance would double this number. Each of these measures of overestimation and overplacement are confounded with ability on the focal task. Such measures vary in terms of whether the measures assess absolute or relative performance, whether self-evaluations assess performance or ability, whether the self-evaluation measure is in the same metric or a different metric as performance, and whether the measures assess overconfidence by including a control variable, calculating a residual, or calculating a difference score.

Base Case

Begin by considering a study designed to assess overconfidence regarding absolute performance using the residual of a self-evaluation measure in the same metric as performance. Participants complete an ability-based task (e.g., a 10-item financial literacy quiz) and then report their self-evaluation of their own performance (e.g., how many of the 10 items do they think that they got correct). The researcher then regresses self-evaluations of performance as the dependent variable on objective performance as the independent variable. The residual of this regression, reflecting how much higher or lower self-evaluations are than is warranted by objective performance, is used as a measure of overconfidence. The researcher then tests whether those residuals are associated with some other measure (e.g., financial planning) in a new analysis.

Residual vs. Control vs. Difference

There are three approaches researchers may use to combine a measure of performance and a self-evaluation to compute overconfidence: they may use residualized self-evaluations, they may control for

performance in a multiple regression analysis, or they may calculate a difference score.[2] The first two are nearly identical. All three are problematic, for related but distinguishable reasons detailed in the model.

Using the *residual* approach, researchers regress self-evaluations on objective performance and keep the residuals as measures of overconfidence. By construction, the residuals have a mean of zero: scores ranging from well-calibrated to overconfident will be indistinguishable from those ranging from underconfident to well-calibrated.[3]

Using the *control* approach, researchers regress the proposed correlate of overconfidence on self-evaluations and control for the performance measure. In this approach, the partial regression coefficient estimate on self-evaluation with performance as a control is precisely the same as the coefficient estimate on the residualized estimate (with or without performance as a control). Controlling for performance rather than (or in addition to) residualizing self-evaluation has the benefit of reducing error variance in the analysis of the outcome measure, thereby providing a more precise estimate.

Using the *difference* score approach, researchers subtract the measure of objective performance from the self-evaluation of performance and keep the difference as a measure of overconfidence. Researchers sometimes use a difference score and also control for objective performance. If they do, the coefficient on the difference score is precisely equivalent to that on self-evaluation when controlling for performance. When researchers control for performance in analysis of some proposed correlate, whether the focal variable is the residual, self-evaluation, or the difference score, the focal coefficient is precisely the same and provides identical inferences.

Table A1 in the supplement characterizes a sample of recent articles in terms of their usage of residuals, difference scores, both, or measures that are equivalent to either (e.g., a measure of confidence without accounting for performance). Of 31 articles, 28 used a difference score or equivalent measure and 21 used a residual score or equivalent measure. This indicates each type of measure is commonly used.

---

[2] Parker and Stone (2014) refer to the residual approach as *unjustified confidence* and the difference score approach as *overconfidence*.

[3] This also precludes the possibility of claims like that from John and Robins's (1994) study of self-enhancement in which they interpret participants with negative residuals as showing a self-diminishment bias (p214).

Self-Evaluate Using the Same vs. Different Metric

The self-evaluation may be assessed in the same metric as performance or in a different metric. Above, both performance (on a 10-item quiz) and self-evaluation (out of 10 items) are in the same metric. Alternatively, researchers may assess self-evaluations in a different metric (e.g., a 1-7 scale). If the self-evaluation is in a different metric, overconfidence may be assessed using the residual or covariate method but should not be assessed using a difference score (even if the variables have been standardized).

Self-Evaluate Performance vs. Ability

Participants may be asked to evaluate their performance or their ability. The cases above represent self-evaluations of task-specific performance. In other cases, the self-evaluation may be an evaluation of ability rather than an evaluation of performance. For example, after completing a 10-item financial literacy quiz, participants may report how well they performed on a 1 to 7 scale (performance), or they may report how knowledgeable they are generally about financial matters on a 1 to 7 scale (ability). Researchers residualize this measure of self-evaluated ability on performance (or control for performance in multiple regression) to consider the role of overconfidence. Although typical examples of self-evaluated ability tend to be in a different metric, it could be assessed in the same metric. For example, in principle researchers could inquire about expected performance on a 10-item test drawn from the same test bank in order to assess ability in the same metric, enabling potential use of difference scores. When participants evaluate expected performance on an upcoming task, the evaluation assesses ability.

Absolute vs. Relative Evaluation

Performance and self-evaluations may be measured in absolute or relative terms. In each case above, the focus is on absolute performance. In Moore and Healy's (2008) parlance, this is overestimation. The same techniques are used when measuring relative performance (i.e., overplacement), such as percentile performance within some specified sample. Self-evaluations of relative standing may be measures of performance on a particular task or measures of evaluations of ability.[4]

[4] In addition to overestimation and overplacement, Moore and Healy (2008) also discuss overprecision: "excessive certainty regarding the accuracy of one's beliefs" (p. 502). The current research does not address overprecision.

Variations on a Theme

These variations may be assembled in any combination as long as it does not involve taking a difference between two measures in different metrics. Evaluations may also be assessed item-by-item to enable assessments of sensitivity or efficiency (e.g., Burson et al., 2006; Fleming & Lau, 2014; Stankov & Crawford 1996). Each of the approaches described above results in a measure of overconfidence that is confounded with ability. As a result, using any of these measures biases measures of the correlation between overconfidence and other measures. The confound and bias is present whether the residual, covariate, or difference approach is taken, for both overestimation and for overplacement, whether self-evaluations are of performance or of latent ability, and whether they use the same or different metrics.

These measures are sometimes, but not always, interpreted as measures of stable, general individual differences in overconfidence. For example, in the finance literature, measures of overconfidence on some unrelated tasks are used as correlates of trading activity or stock ownership (e.g., Grinblatt & Keloharju 2009; Ke 2021); this analysis requires both stability over time and consistency across domains to permit the researchers' preferred interpretation. In other cases, these measures are used as measures of possibly-transient, possibly-domain-specific overconfidence. For example, differences in overconfidence on a particular task with a particular partner are correlated with perceived competence and social status as rated by that partner (Anderson et al. 2012); no assumption of stability or consistency is required. Whether there are stable, general individual differences in overconfidence is a topic of ongoing debate (for arguments and evidence against, see e.g., Li & Moore 2024; Moore & Dev 2017; Moore & Swift 2011; Moore & Schatz 2017; for evidence in favor in favor, see Lawson et al. 2023; Binnedyk & Pennycook 2024). The analysis I present here addresses measured differences in overconfidence, whether they are intended as stable and general or transient and domain-specific.

What's the Problem?

I first give an informal verbal account of the concern, and then provide a formal mathematical proof and simulation results. The problem arises from four properties common to the paradigms described above. First, people typically differ in task-relevant ability (whether that ability is general cognitive

ability, mere test-taking ability, domain-specific knowledge, etc.) Second, people typically have at least partial insight into their ability. Third, performance is typically an imperfect measure of ability: it includes some noise and is unlikely to fully and only assess the construct it is intended to measure. Fourth, performance is typically ambiguous to the participant: people often only receive an imperfect sense of how they did from the task itself prior to evaluating how they did.

Because performance is ambiguous, principles of Bayesian updating require that self-evaluations of performance ought to regress toward people's prior beliefs (Moore & Healy, 2008). This leads to regression toward ability under the weak assumption that people's beliefs about their own ability are correlated with their true ability. This implies that self-evaluations are a (possibly noisy) weighted average of ability and performance. As a result, observing self-evaluations exceed performance signals that ability exceeds performance too. If two quiz-takers who have some insight into their own ability each scored a 70%, and one believes she scored an 80% and the other believes he scored a 60%, there is a sound basis one may use to infer that the first test-taker has greater ability than the second.

Whenever performance is a noisy measure of ability, controlling for differences in performance is not sufficient to control for differences in ability (e.g., Birnbaum and Mellers, 1979; Cohen et al., 2003; Culpepper & Aguinis, 2011; Gillen et al., 2019; Kahneman, 1965; Westfall & Yarkoni, 2016).[5] Because self-evaluations of performance are regressive toward ability, controlling for performance will result in remaining variation in self-evaluation that is attributable to ability. Although the residuals are uncorrelated with performance by construction, they are still correlated with true ability. So measured

[5] The role of error in the use of discrepancy scores and imperfect sampling are repeated themes in the overconfidence literature (e.g., Burson et al., 2006; Erev et al., 1994; Gigerenzer et al., 1991; Juslin, 1993; Klayman et al., 1999). The focus of these critiques has been imperfect calibration and findings regarding aggregate overconfidence rather than the implications for individual-level measures of overconfidence described here. Concerns about inappropriate inferences regarding true scores when relying on measured scores are an old problem in measurement (e.g., Cochran, 1968; Cronbach & Furby, 1970; Lord, 1956, 1958, 1960; McNemar, 1958; Rogosa et al., 1982; Thomson, 1924). This has led to an array of possible approaches to attempt to recover unbiased coefficient estimates (e.g., Cronbach & Furby, 1970; Culpepper & Aguinis, 2011; Kline, 2005; Fuller, 1987). Why has such a critical concern not been central in recent discussions of measures of overconfidence? A key contributing factor may be that both overconfidence measures are uniquely susceptible to an illusion that only performance, not latent ability, matters, because the performance measure is often the target of self-evaluation. But because one's own performance is noisy and ambiguous, requiring people to incorporate their prior beliefs, the conclusion that latent ability does not matter is incorrect.

overconfidence, computed via residuals or by controlling for performance, is confounded with ability. When the self-evaluation is of ability rather than performance (e.g., when assessing expectations), the confound is more severe because the measure directly assesses ability rather than is contaminated by it.

The concern above applies when self-evaluations are residualized or the analysis controls for performance. A related variant applies when difference scores are used. If the measure does not fully and only measure what it is believed to measure, performance will exhibit regression to the mean. People who are very high in ability will perform moderately highly, and people who are very low in ability will perform moderately poorly. As a result, the difference measure will also be confounded with ability. Consider again a 10-item quiz designed to measure financial literacy. But, unbeknownst to researchers or participants, four of the items inadvertently assess trust in institutions instead. A financially-literate but average-trusting participant expected to get 9 answers correct, actually got 7 correct (5 of the 6 financial literacy questions and 2 of the 4 trust questions), and, due to the inherent ambiguity, reported that they got 8 correct. A less-literate but more-trusting participant expected to get 5 answers correct, actually got 7 correct (3 of the 6 financial literacy questions and all 4 of the trust questions), and due to the inherent ambiguity, reported that they got 6 correct. In this example, the apparent overconfidence of the first participant and underconfidence of the second participant reflect true differences in financial literacy, not a surplus nor deficit of confidence.[6]

Because people's private information regarding their own ability affects their evaluations of their own performance but not their evaluations of others' performance, this same logic leads to the same confound for overplacement.

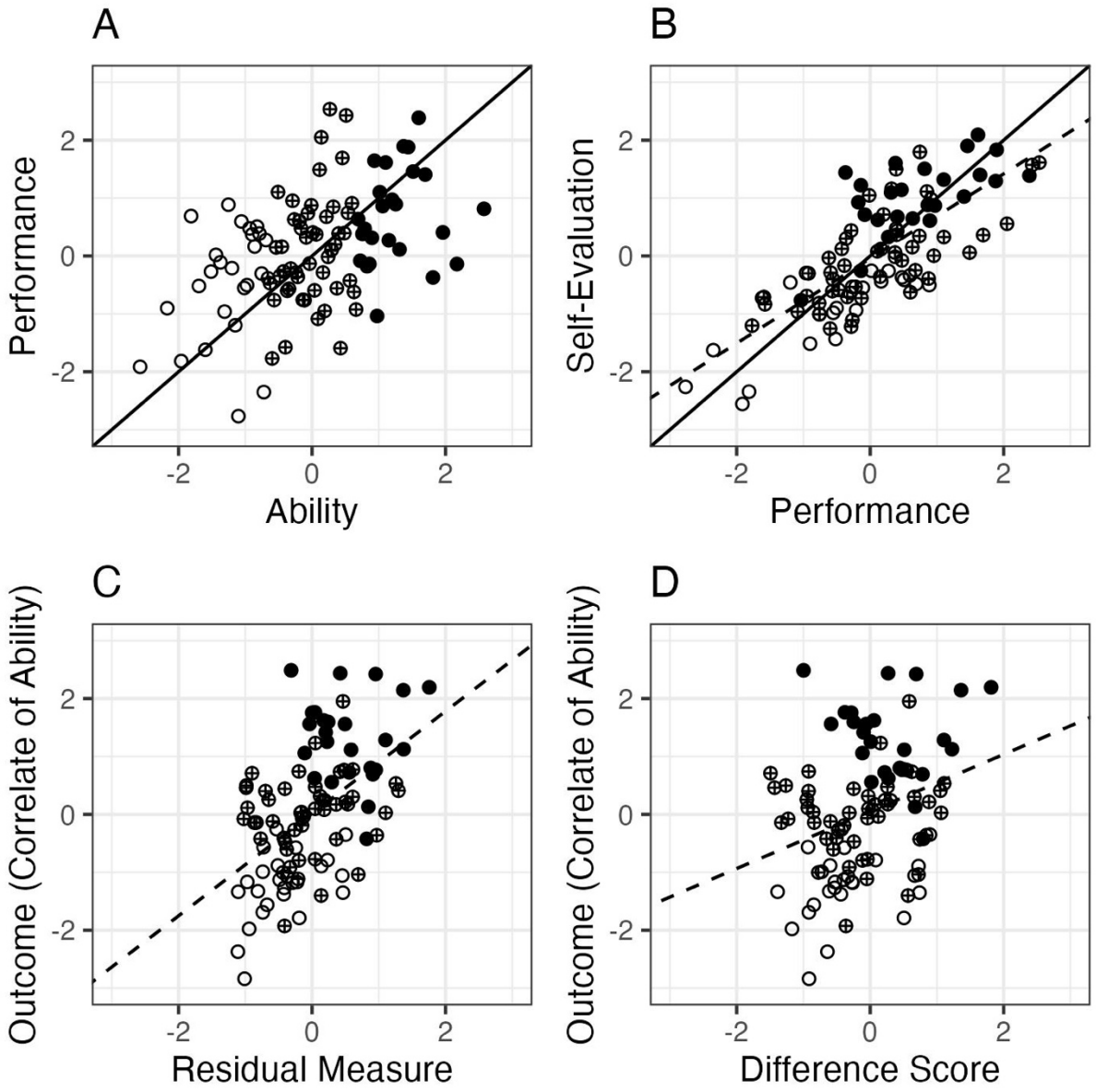
A Simulated Example of the Problem

A visual depiction of the problem is given in Figure 1. In this example, participants vary in ability, have perfect self-insight into their own ability, and have ambiguity about their own performance.

[6] It would be inappropriate to attribute the error to the participant: the fact that the participant uses prior beliefs about financial literacy rather than a linear combination of financial literacy and trust in institutions should not be interpreted as overconfidence if they rely on the very construct the researchers themselves believe they are measuring. I return to this point at the end of paper.

Figure 1

Visual Depiction of How Measures of Overconfidence Are Confounded With Ability



Note. The 25 highest-ability individuals are depicted as filled circles. The 25 lowest-ability individuals are depicted as open circles. The 50 middle-ability individuals are depicted as crossed circles. Solid lines depict 45-degree lines; dashed lines depict best-fit regression lines. The parameters used for this example from the model described later are $\rho = 1, \lambda = .5, \sigma_v^2 = 1, \alpha = .5, \sigma_v^2 = .25, \beta = 1, \sigma_\epsilon^2 = .25$.

In evaluating their performance, they rely on a noisy signal of their own performance and beliefs about their own ability. The outcome is correlated with ability, and only correlated with beliefs because their

beliefs are accurate. There is no overconfidence in this example. Details regarding the model parameters used to simulate these data are given in the figure caption; the model is described in the next section.

Panel A shows the relationship between ability and performance for 100 people. In this example, the performance measure is both noisy and regressive; the 45-degree line is given by the solid line. The 25 highest-ability people are depicted as filled circles and the 25 lowest-ability people are depicted as open circles; the 50 middle-ability people are depicted as crossed circles.

Panel B depicts self-evaluations as a function of performance when self-evaluations are regressive toward ability. The 50 individuals classified as overconfident by the residual score (i.e., the vertical difference between each point and the dashed best-fit line) include 22 of the 25 highest-ability individuals and only 5 of the 25 lowest-ability individuals, and the 50 individuals classified as underconfident by the residual score include 20 of the 25 lowest-ability individuals and only 3 of 25 highest-ability individuals. Similarly, the 44 individuals classified as overconfident by the difference score (i.e., the vertical difference between each point and the solid 45-degree line) included 17 of the 25 highest-ability individuals and only 6 of the 25 lowest-ability individuals, whereas the 56 individuals classified as underconfident by the difference score included 19 of the 25 lowest-ability individuals and only 8 of the 25 highest-ability individuals. In this example, classifying individuals by overconfidence effectively, albeit imperfectly, classifies them by ability ($r_{ability,residual} = 0.60$, $r_{ability,difference} = 0.37$).

Panels C and D plot the correspondence between the residual measure (C) and difference score (D) with an arbitrary correlate of ability. As is evident in this example, these correlates of ability are positively correlated with both measures of overconfidence, despite the fact that both measures of overconfidence account for performance and in this example there is no true overconfidence. As is derived below, the problem in (C) arises from the simulated measurement error in performance whereas the problem in (D) arises from the simulated regression to the mean in performance.

Modeling the Bias in Measures of Overconfidence

I next formalize the model described qualitatively above. A straightforward extension of Moore and Healy's (2008) model of overconfidence permits a focus on differences, so I adapt their notation.[7]

Model Setup

Latent Ability and Beliefs

People, denoted by *i*, differ in their task-specific ability or skill:

$$S_i \sim D(0, 1) \quad (1)$$

where D(0, 1) represents any distribution that has been standardized to have a mean of 0 and variance of 1. Depending on the task, ability may represent general cognitive ability, mere test-taking ability, domain-specific knowledge, current caffeination status, etc.

People's beliefs, $\tilde{S}_i$, are a function of their own ability:

$$\tilde{S}_i = \rho S_i + \zeta_i \quad (2)$$

where $0 \leq \rho \leq 1$ and ζ is independently drawn from any distribution with mean 0 and variance $1 - \rho^2$, such that $\tilde{S}$ has unit variance and the correlation between S_i and $\tilde{S}_i$ is given by ρ.[8] Latent overconfidence is then given by $\tilde{S}_i - S_i$.

People often can and do have meaningful insight into their own ability, suggesting we ought to expect $\rho > 0$. The Subjective Numeracy Scale (Fagerlin et al., 2007) was developed to find a way for people to self-report their own numeracy using a less-burdensome task than a math test. Objective financial literacy shows correspondence with subjective financial literacy (Lusardi & Mitchell 2017).

---

[7] This relates to a discussion in Healy and Moore (2007) and footnote 2 in Moore and Healy's (2008) in which luck is separated from expectations of ability. The implication for bias in the measure of overconfidence is not addressed.

[8] This sets the mean level of overconfidence to 0 by construction, but this assumption is merely for convenience. As the current discussion only addresses relatively more or less overconfident individuals, one can arbitrarily shift beliefs by changing the mean of ζ without any substantive impact on the argument developed here regarding the confound with ability. This also equates the variance of beliefs to the variance of ability. Relaxing this constraint and freeing σ_ζ^2 so that it may take on other values would lead ρ in equations 6 and 8 to be replaced by $\rho\sqrt{\rho^2 + \sigma_\zeta^2}$. If the variance of beliefs exceeds that of ability, the residual's confound is larger and the difference score's confound is more-positive or less-negative. If the variance of beliefs is less than that of ability, the residual's confound is smaller and the difference score's confound is less-positive or more-negative.

Objective knowledge and subjective knowledge are correlated across a range of domains (Alba & Hutchinson, 2000; Carlson et al., 2009). Across multiple domains, there is good reason to expect people have at least partial insight into their own abilities. Partial but incomplete insight into their own skill can be modeled as $0 < \rho < 1$.

If $\rho = 1$, then $\tilde{S}_i = S_i$ and there is no latent overconfidence. This perfect self-insight condition is of interest not because it is likely to be accurate, but rather because it presents an important null model to address: Is there evidence of a correlation with measures of overconfidence even when latent overconfidence does not exist? An answer of "yes" would be troubling. The assumption of perfect self-insight is assuredly wrong in many—if not all—cases. But if the methods we use to assess overconfidence and its correspondence with other constructs find evidence of a correlation in the absence of overconfidence, we must rethink those findings.

Observable Performance and Self-Evaluations

Although ability varies across people, it is not directly observable. Instead, people's performance, P_i, is assessed on a particular task. Performance reflects a combination of ability and luck:

$$P_i = \lambda S_i + \nu_i \quad (3)$$

where $0 \leq \lambda \leq 1$ and luck, ν_i, is independently drawn from any distribution with mean 0 and variance σ_ν^2. λ represents performance's loading on ability. A perfect measure that fully and only captures the focal ability (possibly with classical measurement error) has $\lambda = 1$; an invalid measure (e.g., pure noise or a measure of an unrelated construct) has $\lambda = 0$. Consider a researcher measuring individual differences in intelligence using either (a) a test consisting of three of Raven's progressive matrices, or (b) a phrenologist's head measurements. Both measures contain noise, but for Raven's matrices we expect $\lambda > 0$ (whether or not $\lambda = 1$) whereas for the phrenologist's head measurements we expect $\lambda = 0$.

People's self-evaluations, $\widetilde{P}_i$, are their noisy attempts to evaluate their own performance, P_i. After complete feedback, performance may be unambiguous. But prior to such feedback, people generally have some ambiguity regarding how they performed. As Moore and Healy (2008) persuasively argue, the

presence of such uncertainty should lead to self-evaluations that incorporate prior beliefs through Bayesian-like reasoning (whether or not people are proper Bayesian updaters). For Moore and Healy, these prior beliefs represented beliefs about the simplicity of the task. In the current model, these prior beliefs represent beliefs $\tilde{S}_i$ about one's own ability S_i. The key extension is thus the relevant and systematic variability in those prior beliefs. The result of this Bayesian-like reasoning is that people ought to evaluate their own performance as lying somewhere between their prior beliefs and their true performance, plus noise, where the weight on prior beliefs increases with ambiguity:[9]

$$\tilde{P}_i = \alpha\widetilde{S_i} + (1-\alpha)P_i + \upsilon_i \quad (4)$$

where $0 \leq \alpha \leq 1$ and υ_i is independently drawn from any distribution with mean 0 and variance σ_υ^2. α represents the ambiguity of someone assessing their own performance. As ambiguity increases, α approaches 1, and self-evaluations of performance reflect beliefs about ability to a greater extent. When self-evaluations are measures of ability rather than measures of performance, $\alpha = 1$, as the measure is only a measure of ability and is not designed to assess performance at all.

The measurement model is depicted in Figure 2. Panel A depicts the general version, in which beliefs are imperfectly correlated with ability. Panel B represents a simplified version of the model implied by constraining $\rho = 1$.

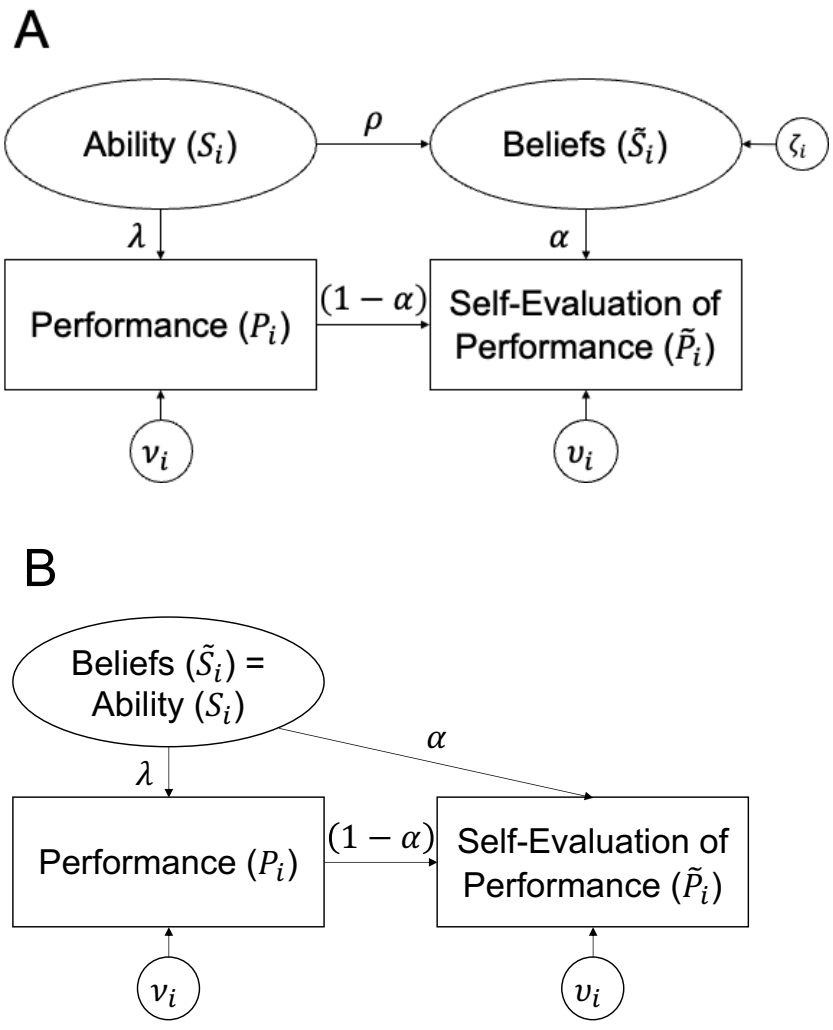
There are two different reasons why P_i and $\tilde{P}_i$ may be uncorrelated. First, they will be uncorrelated if $\alpha = 1$ and $\rho = 0$. This leads to the typical interpretation: performance (may) provide a noisy signal of ability, and people can assess neither how they performed nor their true ability. Second, they will also be uncorrelated if $\alpha = 1$ and $\lambda = 0$. In this case, they are uncorrelated because performance does not provide a signal of ability, even though people have insight into their true ability. The difference in these two cases implies observing no correlation between P_i and $\tilde{P}_i$ is not sufficient to claim that $\rho = 0$.

[9] If people knew their true performance, P_i, they could simply report it directly. The fact that P_i enters their beliefs but is not used directly reflects the fact that participants receive a noisy signal of their performance. That noise is then folded into υ_i, leaving the signal to enter the equation directly. See Moore and Healy (2008).

Also note that if $\rho = 1$, then $\tilde{S}_i = S_i$ because $E[\zeta] = 0$ and $\sigma_\zeta^2 = 0$. In this case, there are no differences in overconfidence. If $\rho < 1$, there may be differences in overconfidence, but that does not imply they are *relevant* differences in overconfidence. When researchers introduce a third variable and argue for its relation to overconfidence, they typically imply a causal arrow to or from $\tilde{S}_i$. For $\rho > 0$, beliefs may be correlated with the third variable entirely due to their mutual correlations with ability.

Figure 2

Measurement Model of Relationships Among Ability, Beliefs, Performance, and Self-Evaluations



Note. Panel A enables imperfect but correlated beliefs. Panel B depicts the equivalent model when constraining $\rho = 1$. In Panel B, $S_i = \tilde{S}_i$ because $\rho = 1$ so $E[\zeta] = 0$ and $\sigma_\zeta^2 = 0$.

Computing Measures of Overconfidence

The researcher aims to isolate the role of overestimation by: (a) regressing $\tilde{P}_i$ on P_i and keeping the residual, (b) including both self-evaluation $\tilde{P}_i$ and performance P_i in a single multiple regression model, or (c) taking the difference between $\tilde{P}_i$ and P_i. I first address the residual and regression approaches together, as they result in equivalent coefficients, and then the difference score.[10] Each of these approaches lead to confounded measures under the conditions specified below.

Residual and Multiple Regression Approaches

To calculate overconfidence via residuals, researchers regress self-evaluations on performance:

$$\tilde{P}_i = \gamma P_i + \epsilon_i \quad (5)$$

The residuals, $e_i = \hat{\epsilon}_i$, are kept as measures of overconfidence.[11] Because (a) performance is noisy, and (b) self-evaluations incorporate priors on self-evaluations of ability, the expected errors (and thus the residuals) vary with ability:

$$E[\epsilon|S] = \rho\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_v^2}\right)\alpha S \quad (6)$$

The derivation in the Supplement.[12] For reasonable sample sizes such that $e_i \cong \epsilon_i$, the residual from regressing self-evaluation on performance is positively confounded with ability if $\rho\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_{v_i}^2}\right)\alpha > 0$. In other words, there is a confound if three conditions hold. First, true ability and beliefs about one own ability must be positively correlated ($\rho > 0$); the confound is maximized if people have perfect self-insight and there are no differences in overconfidence ($\rho = 1$). Second, there must be error in the performance measure that is not attributable to ability ($\sigma_v^2 > 0$, making $\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_{v_i}^2}\right) > 0$). The absence of measurement error is the exception, not the rule, so this condition is likely to be met.[13]

[10] Using difference scores controlling for performance results in the same coefficient as if one used residuals.
[11] Throughout, I exclude intercepts for simplicity. Because my focus is on individual differences in overconfidence rather than mean levels of overconfidence, intercepts can be accounted for through recentering.
[12] If we relax the constraint that $\sigma_\zeta^2 = 1 - \rho^2$, equation 6 would be: $E[\epsilon|S] = \rho\sqrt{\rho^2 + \sigma_\zeta^2}\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_v^2}\right)\alpha S$.
[13] Note that under this interpretation, we are assuming all variance in performance is attributable to ability or else to measurement error. In the discussion I consider when σ_v^2 might include systematic error (e.g., other constructs).

Third, self-evaluations must be related to beliefs conditional on performance ($\alpha > 0$), not just through performance. Any application of basic Bayesian logic in the presence of uncertainty will lead to a direct effect of beliefs on self-evaluations, as will commonly-used measures that include self-evaluations of ability rather than just self-evaluations of performance, so this condition is likely to be met as well.

Multiple regression can be rewritten as a regression of residuals on residuals, so when predicting a candidate correlate, the regression coefficient on evaluations controlling for performance will be precisely the same as the regression coefficient on residualized evaluations. The multiple regression estimate will typically be more-precise.

Given the broader literature on measurement error in predictors (e.g., Birnbaum & Mellers, 1979; Cohen et al., 2003; Culpepper & Aguinis 2011; Gillen et al., 2019; Kahneman, 1965; Westfall & Yarkoni, 2016), why does the current paradigm deserve special consideration? First, unlike subjective responses to 7-point scales or preferences as measured by intertemporal choice or risk preference tasks, performance measures contain a veneer of precision and objectivity that may wrongly evoke less concern regarding its status as a noisy measure of ability. Some researchers may interpret it as a mere observable rather than a measure of a latent construct. Second, without the extension of Moore and Healy's (2008) model, it may not be transparent to all researchers that self-evaluations are directly affected by beliefs about ability. This grants a false sense of security regarding the impact of any measurement error in performance.

Difference Score Approach

To calculate overconfidence using a difference score, one simply subtracts performance from self-evaluation:

$$\Delta_i = \tilde{P}_i - P_i \tag{7}$$

In expectation, this difference score is also a function of ability:

$$E[\Delta|S] = (\rho - \lambda)\alpha S \tag{8}$$

The derivation is in the Supplement.[14] The difference is positively confounded with ability if $(\rho - \lambda)\alpha > 0$. Once again, it is positively confounded if a specifiable set of conditions hold. First, beliefs must be positively correlated with ability ($\rho > 0$). Second, performance must load sufficiently poorly on ability ($\lambda < \rho$). Third, self-evaluations must be related to beliefs conditional on performance ($\alpha > 0$), not just via performance.[15] If people hold accurate beliefs, then in the idealized case in which the measure of performance fully and only measures the construct that researchers and participants think it measures, $\lambda = 1$ and there is no association between the difference score and ability. (There will also be no association if beliefs are as imperfectly related to ability as performance is, i.e., if $\rho = \lambda$.) Similarly, as in the residual case, if self-evaluation only depends on performance and not beliefs, then $\alpha = 0$ and there is no relationship between the difference and ability.

Unlike the residual measure, and aligning with typical critiques of difference scores, the difference score measure can be negatively confounded with ability if $\rho < \lambda$.[16] This implies that for certain parameter configurations, there can be cases in which the residual measure of overconfidence is positively correlated with some outcome, the difference score measure is negatively correlated with that same outcome, and each correlation is entirely attributable to the measure's confound with ability.

[14] If we relax the constraint that $\sigma_\zeta^2 = 1 - \rho^2$, equation 8 would be: $E[\Delta|S] = \left(\rho\sqrt{\rho^2 + \sigma_\zeta^2} - \lambda\right)\alpha S$.

[15] In this idealized case, measurement error (σ_v^2) is inconsequential. In practice however, P_i is often bounded such that $\sigma_v^2 > 0$ would likely drive λ down.

[16] Prior critiques have addressed the fact that difference scores are confounded with their component measures: what appears to be a property of the difference may instead reflect a property of one of the components (e.g., Cronbach & Furby 1970; Cohen et al., 2003; Edwards & Parry 1993; Griffin et al., 1999; Johns 1981; Wall & Payne 1973; Zuckerman & Knee 1996). Response Surface Analysis via polynomial regression (e.g., Edwards 1994; Barranti et al., 2017; Humberg, Dufner, et al. 2019; Humberg, Nestler, & Back 2019) and Condition-based Regression Analysis (e.g., Humberg et al., 2018a, 2019) seek to establish alternative conditions to establish whether a discrepancy vs. a positive self-evaluation is the active ingredient. The concern I raise regards a confound of self-evaluations with ability, and so is relevant whether one is interested in the discrepancy score or positive self-evaluation. In addition to other concerns regarding condition-based regression analysis in their standard form (Krueger et al., 2017; Fiedler, 2021; cf. Humberg et al., 2018b, 2022), these regression-based approaches do not distinguish between performance and ability, and so are equally susceptible to the concerns I raise here.

Comparing the Two Approaches

Comparing equations 6 and 8 reveals the relative magnitude of the bias for residual and difference scores. For $\lambda = 0$, both equations simplify to the same expectation, $\rho\alpha S$. This is the relevant case if one only uses confidence and ignores performance, as arises in several documented cases in Table A1. For $\lambda > 0$, the bias from the residual is greater (in signed magnitude) than the bias from the difference score if either (a) ability's effect on performance is larger than its effect on beliefs ($\lambda > \rho$), or (b) error in performance is large enough ($\sigma_v^2 > \lambda(\rho - \lambda)$). If one were to reasonably constrain variance of performance to be no less than variance of ability ($\sigma_v^2 \geq 1 - \lambda^2$), this condition always holds except for the edge case of $\rho = \lambda = 1$.

If researchers use a difference score but control for performance, the covariate-adjusted regression coefficient and statistical test of the difference score is precisely equivalent to that using the multiple regression approach. The intuition is that the coefficient on the difference score is interpreted as "all else constant," and when "all else" includes performance, the only way the difference (evaluation-performance) changes holding performance constant is by the evaluation changing. Thus, because the multiple regression approach leads to the same bias as the residual score approach, using difference scores while controlling for performance behaves like the residual score calculation.

For both residual and difference measures, the same confound holds for both overestimation of absolute performance and overplacement of relative performance. Evaluations of one's own performance are a function of one's own idiosyncratic ability and idiosyncratic beliefs; evaluations of others' are not. As a result, the conditional expectation of evaluations of others' performances are not a function of one's own ability. The additional terms in overplacement drop out, leaving a bias in one's relative performance (overplacement) which matches that in one's absolute performance (overestimation).[17]

[17] This may appear to be at odds with Moore and Healy (2008) who show systematic reversals between overestimation and overplacement. Because S in their model represents a quiz's simplicity, it takes a common value for every individual such that one's prior beliefs are relevant for both oneself and for others. In my model, S_i represents an individual's ability, so it takes an independent value for every individual such that one's prior beliefs are relevant for oneself but not for others.

Simulation Results

Simulations show that these asymptotic results hold for reasonable sample sizes. For each of 7,776 combinations of parameter values (i.e., all factorial combinations of each of ρ, λ, α, σ_v^2, and σ_v^2 taking a value in [0, 0.2, 0.4, 0.6, 0.8, 1.0]), I simulated 1,000 samples of 100 observations each. In each sample, skill was drawn from a standard normal distribution, and error terms were drawn from standard normal distributions scaled by the corresponding variance. Performance and self-evaluations followed the model. Code will be posted: https://researchbox.org/1597&PEER_REVIEW_passcode=ORRDVP.

Simulation results are depicted in Figure 3. Repeating the simulations using log-normal distributions recovered the same results, reflecting that the variables need not be normally distributed.

Two Empirical Applications

Common measures of overestimation and overplacement are confounded with ability in reasonable null models in which overconfidence does not exist or does not matter (i.e., it only relates to candidate constructs through ability). The magnitude of the confound depends on the parameter values and the model is undoubtedly a simplification. Does the proposed confound actually matter in practice? Two empirical applications indicate a resounding yes.

In the first application, I examine a case in which an outcome measure (performance on a related task) should be related to ability and should not be related to overconfidence. If we observe that the outcome is related to measures of overconfidence, this can be explained parsimoniously through the confound with ability. Given strong prior beliefs about true relationships among ability, overconfidence, and the outcome measure, this demonstrates that the confound can hold in practice.

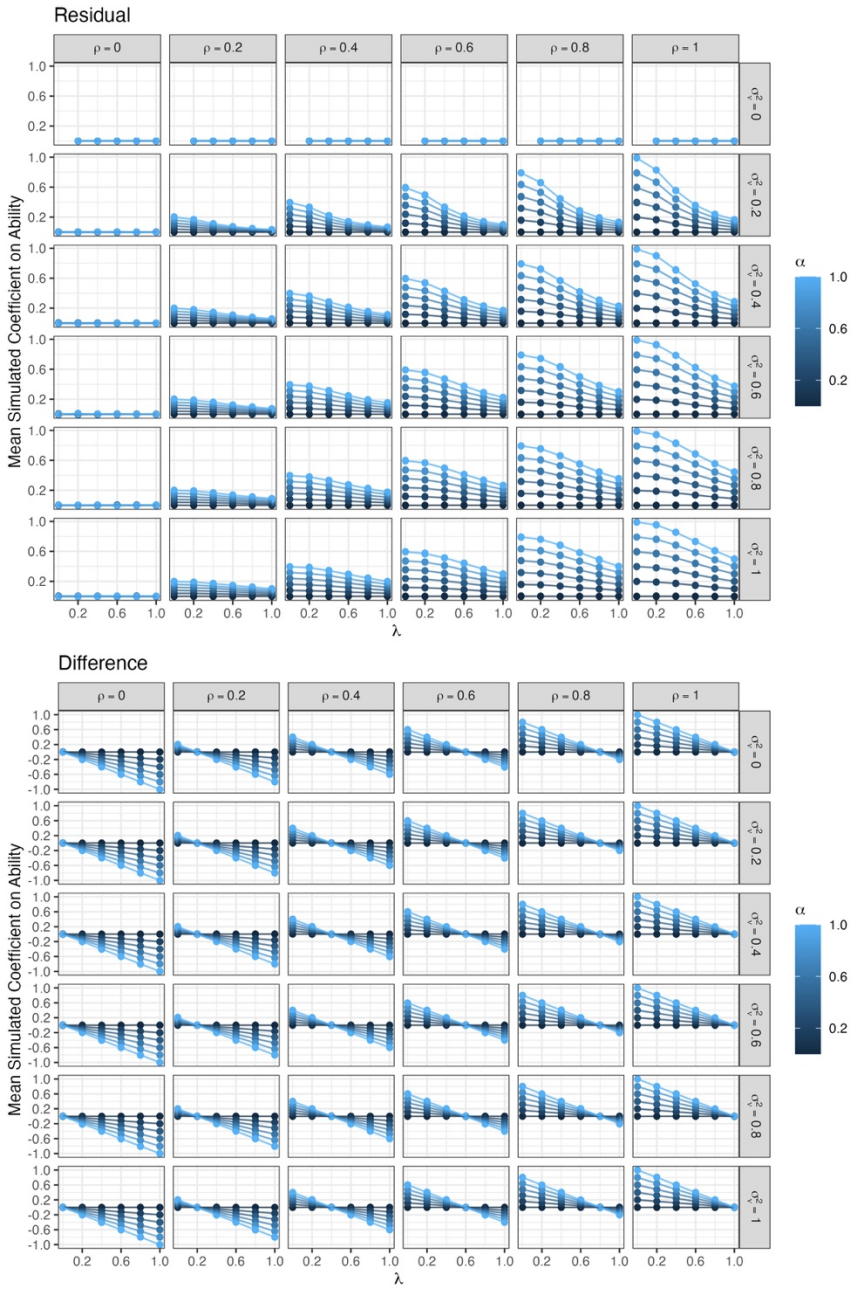
In the second application, I re-examine a related pair of published findings in which results are interpreted in terms of overconfidence. Using the current model, I find that they are also fully compatible with plausible parameter values in which there is no consequential overconfidence (or no overconfidence at all). This indicates the confound provides a plausible alternative interpretation of published results.

The first application indicates measures of overconfidence can lead to problematic conclusions. The second application indicates substantive claims from the literature which have been interpreted in

terms of overconfidence may not be sufficient to imply a causal or consequent role of overconfidence, and are even consistent with perfectly-calibrated beliefs (i.e., the absence of overconfidence).

Figure 3

Simulation Results of Bias in Residual and Difference Scores as a Function of ρ, λ, α, and σ_v^2



Note. σ_v^2 is plotted but not apparent as it does not affect the bias. For the residual panel, when $\sigma_v^2 = \lambda = 0$, the estimate is missing: performance does not vary and so cannot be used as a covariate.

Empirical Application I: Overconfidence Predicts Performance

The analysis above indicates that overconfidence measures are biased under plausible conditions. Is this bias likely to distort inferences? The answer depends on typical parameters: how strongly performance loads on ability, how much measurement error is in performance, whether people exhibit (even imperfect) Bayesian updating, and the degree to which people have self-insight into their own ability. To test this, I first examine a case where: (a) there is a measure of performance, (b) there is a self-evaluation of that performance, and (c) there is an outcome measure which is a priori likely related to ability and unrelated to overconfidence.

Using data from Moore and Healy (2008), I consider a case where the relevant outcome is a future measure of performance.[18] Future performance cannot cause past overconfidence, and it is unlikely that past overconfidence causes future performance in a way that is not attenuated as the number of intervening tasks increases. As a result, finding that past residuals or difference scores predict future performance suggests that past residuals or difference scores are confounded with skill.

Overconfidence Paradigm

Moore and Healy (2008) collected data from 82 college undergraduates on many measures. I describe the relevant components here and refer the reader to Moore and Healy for complete details. Participants completed 18 10-item trivia quizzes: an easy, medium, and hard quiz on each of six topics. The quizzes were presented sequentially in six blocks. Each block contained an easy, a medium, and a hard quiz on different topics, in randomized order. In addition to other measures, for each quiz, participants: (a) provided a pre-quiz measure of expected performance, (b) took the quiz, and (c) provided a post-quiz measure of estimated performance.

Analysis of these data requires addressing two key complications. First, Moore and Healy's (2008) quizzes systematically differed in difficulty. Performance on other quizzes provides a proxy for ability, such that if trivia quiz ability exists, performing well on one quiz should predict performing well

[18] Moore and Healy do not make the inferential error reported here. Rather, the availability and richness of their data (https://osf.io/6tecy/) present an opportunity to examine whether the error can affect real inferences.

on another, all else equal. But performing well on one quiz is a signal not only that ability may be high, but also that difficulty may be low. Indeed, the block-randomized design leads to negative autocorrelation in difficulty between successive quizzes: in expectation in the first 5 blocks, there is a 44% chance that a hard quiz is followed by an easy quiz but only an 11% chance that a hard quiz is followed by another hard quiz. This mechanically generates a negative correlation between performance on one quiz and performance on the subsequent quiz. To address this design characteristic, I consider expectations, estimates, and performance for *blocks* (where each block is a triplet of quizzes) rather than for *quizzes*. A block always consists of one easy, one medium, and one hard quiz, reducing the extent to which performance on one block is negatively correlated with performance on other blocks.

Second, the data provide rich within-subject data but with a modest sample size for between-subject analyses by current standards (82 participants). To exploit the within-subject data, I consider performance on sets of sequential quizzes, clustering errors by subject. For example, to examine skill, I regress block performance on prior block performance; each participant contributes five observations: block 2 performance as a function of block 1 performance, block 3 performance as a function of block 2 performance, etc. The analysis accounts for non-independence through clustered errors using the *lm_robust()* function from the *estimatr* package (Blair et al., 2022).

A Puzzle: Overconfidence Predicts Subsequent Performance

Using the first five quiz blocks to provide measures of performance and self-evaluations, I follow the approaches from the literature to construct three measures of overestimation: residualized self-evaluations, self-evaluation controlling for performance, and difference scores. I similarly construct the three corresponding measures of overplacement using Moore and Healy's (2008) overplacement measure.

Overestimation as assessed via residualized self-evaluations predicted performance in the next block. The partial coefficient on self-evaluations controlling for performance also predicted performance in the next block. This coefficient is necessarily equal to that on the residual, but estimated more precisely because performance accounts for additional variance in the dependent variable in the multiple regression analysis but not the residual analysis. Overestimation as assessed via the difference did not significantly

predict performance in the next block. Similarly, overplacement as assessed via residualized relative self-evaluations predicted relative performance in the next block, as did overplacement as assessed via the partial coefficient on relative self-evaluations controlling for relative performance. Overplacement as assessed via the difference score also predicted performance in the next block, but the coefficients were significantly negative. Each regression coefficient, test, and 95% confidence interval is given in Table 1.

Table 1

Regression Coefficients Predicting Subsequent Performance

		Overestimation		
Measure Type	Predictor Lag	Coef (SE)	Test statistic	95% CI
Residual Score	1 (Primary)	0.298 (0.141)	t(38) = 2.11, p = .042	[0.012, 0.584]
	2	0.381 (0.164)	t(33) = 2.33, p = .026	[0.048, 0.715]
	3	0.323 (0.208)	t(28) = 1.56, p = .130	[-0.102, 0.748]
	4	0.434 (0.233)	t(20) = 1.86, p = .077	[-0.051, 0.918]
	5	0.295 (0.403)	t(15) = 0.73, p = .476	[-0.562, 1.151]
Multiple Regression	1 (Primary)	0.298 (0.088)	t(38) = 3.40, p = .002	[0.121, 0.476]
	2	0.381 (0.092)	t(33) = 4.14, p < .001	[0.194, 0.569]
	3	0.323 (0.164)	t(28) = 1.97, p = .058	[-0.012, 0.659]
	4	0.434 (0.155)	t(20) = 2.81, p = .011	[0.111, 0.756]
	5	0.295 (0.365)	t(15) = 0.81, p = .432	[-0.482, 1.071]
Difference Score	1 (Primary)	0.085 (0.121)	t(37) = 0.71, p = .485	[-0.160, 0.330]
	2	0.137 (0.131)	t(31) = 1.05, p = .302	[-0.129, 0.404]
	3	0.077 (0.177)	t(25) = 0.43, p = .668	[-0.287, 0.440]
	4	0.052 (0.174)	t(17) = 0.30, p = .771	[-0.316, 0.419]
	5	-0.001 (0.357)	t(14) = -0.00, p = .997	[-0.765, 0.763]

		Overplacement		
Measure Type	Predictor Lag	Coef (SE)	Test statistic	95% CI
Residual Score	1 (Primary)	0.401 (0.103)	t(42) = 3.89, p < .001	[0.193, 0.609]
	2	0.419 (0.106)	t(42) = 3.95, p < .001	[0.205, 0.632]
	3	0.417 (0.120)	t(41) = 3.48, p = .001	[0.175, 0.658]
	4	0.437 (0.138)	t(33) = 3.16, p = .003	[0.156, 0.717]
	5	0.400 (0.249)	t(22) = 1.61, p = .122	[-0.116, 0.915]
Multiple Regression	1 (Primary)	0.401 (0.056)	t(42) = 7.13, p < .001	[0.287, 0.514]
	2	0.419 (0.062)	t(42) = 6.73, p < .001	[0.293, 0.544]
	3	0.417 (0.085)	t(41) = 4.91, p < .001	[0.246, 0.588]
	4	0.437 (0.093)	t(33) = 4.68, p < .001	[0.247, 0.626]
	5	0.400 (0.176)	t(22) = 2.28, p = .033	[0.036, 0.764]
Difference Score	1 (Primary)	-0.188 (0.062)	t(52) = -3.05, p = .004	[-0.311, 0.064]
	2	-0.152 (0.062)	t(49) = -2.45, p = .018	[-0.277, -0.027]
	3	-0.191 (0.077)	t(48) = -2.49, p = .016	[-0.346, -0.037]

4	-0.206 (0.073)	t(36) = -2.83, p = .007	[-0.354, -0.059]
5	-0.202 (0.123)	t(24) = -1.64, p = .114	[-0.456, 0.052]

Note. All degrees of freedom estimated given cluster-robust standard errors. In all analyses, residuals were calculated using only observations included in the relevant analysis.

One might argue that overconfidence improves subsequent trivia quiz ability (e.g., via self-efficacy). Most such explanations would predict the correlation is stronger for adjacent blocks. Yet there is no evidence that coefficients from the residual or covariate analyses diminish with lags. See Table 1.

Instead, the conceptual model proposed above provides a parsimonious explanation: performance in the current and future blocks are both driven by ability, and the measure of overconfidence is confounded with ability. The fact that difference scores did not predict future performance (for overestimation) or negatively predicted future performance (for overplacement) may be attributable to imperfect self-insight (i.e., $\rho = \lambda$ for overestimation or $\rho < \lambda$ for overplacement).[19]

To examine whether the conceptual model is consistent with the residual and multiple regression analyses, I examine whether four necessary components are in place: (a) Are there differences in ability? (b) Do participants have insight into their own ability? (c) Does trivia quiz performance contain error as a measure of trivia quiz ability? and (d) Does a proxy for ability predict self-evaluations beyond performance? I focus on the overestimation results.

Are There Differences in Ability? Yes

If performance is correlated across blocks, there is evidence of systematic differences in trivia quiz ability.[20] I regress performance in block *t* on prior performance in block *t*-1, clustering errors by subject. The coefficient on lagged performance was 0.754 (SE = 0.045, $t(31) = 16.59$, $p < .001$; 95% CI: [0.662, 0.847]), indicating high performance on one block is strongly associated with high performance on the next block. When an analogous approach was used with block *t*-2, *t*-3, etc., there was no evidence of a relationship that decays with lag (lag 2: b = 0.783, SE = 0.057; lag 3: b = 0.788, SE = 0.076; lag 4: b

[19] The current model makes no attempt to relate beliefs about others on a given quiz to their own beliefs or to their own ability and does not provide an immediate explanation to reconcile the overestimation difference score and overplacement difference score results. One could add a parameter, but it would not address the current confound.
[20] Ability includes skill, knowledge, and other necessary inputs that remain stable during the course of the study.

= 0.796, *SE* = 0.088; lag 5: *b* = 0.756, *SE* = 0.094). These results are consistent with the presence of individual differences in ability at trivia quizzes ($\sigma_S^2 > 0$), which are noisily measured by each quiz.

Do Participants Have Insight Into Their Own Ability? Yes

If participants can predict how they will perform on a quiz without knowing the specific content of that quiz, it suggests they have some insight into their own trivia quiz ability. I regress pre-quiz expectations on subsequent performance, clustering errors by subject. (At the time of the pre-quiz expectation, subjects had little information on which to base their predictions, as neither the quiz difficulty nor the quiz topic was known yet.) The coefficient on performance was 0.467 (*SE* = 0.068, *t*(31) = 6.87, $p < .001$, 95% CI: [0.329, 0.606]). This suggests participants have partial insight into how they will perform.[21] Given the limited information available, this is most readily attributable to self-insight into their own ability ($\rho > 0$).

Does Trivia Quiz Performance Contain Error as a Measure of Trivia Quiz Ability? Yes

It is difficult to imagine that three 10-item quizzes could constitute an errorless measure of trivia quiz skill. So in regressing performance on lagged performance, it comes as no surprise that indeed, $R^2 = 0.538 < 1$, indicating that it is not the case that both measures are errorless indicators of the same construct. Performance as a measure of ability contains error ($\sigma_v^2 > 0$).

Does a Proxy for Ability Predict Self-Evaluations Beyond Performance? Yes

The last required component is that participants' self-evaluations of performance are regressive toward their own ability. Ability is not directly observable, but subsequent performance provides a noisy proxy. For this particular purpose the only requirement is that it contains sufficient signal, not that it excludes sufficient noise. I regress self-evaluations on current performance and subsequent performance, where subsequent performance serves as a noisy proxy for skill. The coefficient on current performance

[21] One may be concerned that participants are aware of the difficulty of the third quiz in each block, thereby inflating this relationship. If the first quiz was hard and the second was medium, participants could determine that the third would be easy. The main result also holds using only the first quiz from each block (adjusted for difficulty), which was completely randomized (*b* = 0.233, *SE* = 0.048, *t*(31) = 4.80, $p < .001$, 95% CI: [0.134, 0.332]). The coefficient was no stronger when considering the third quiz (*b* = 0.177, *SE* = 0.041, *t*(45) = 4.33, $p < .001$).

was 0.880 (*SE* = 0.035, *t*(49) = 24.94, *p* < .001; 95% CI: [0.809, 0.951]), indicating that participants indeed have some idea of how well they did on each block, though this coefficient also partially captures the role of ability given the presence of noise in both measures of performance. Critically, the coefficient on subsequent performance was 0.099 (*SE* = 0.029, *t*(49) = 3.43, *p* = .001, 95% CI: [0.041, 0.156]): controlling for current performance, future performance is predictive of current self-evaluations. The magnitude of this coefficient did not attenuate with more intervening blocks (1 block intervening: *b* = 0.119, *SE* = 0.034; 2 blocks intervening: *b* = 0.101, *SE* = 0.052; 3 blocks intervening: *b* = 0.141, *SE* = 0.044; 4 blocks intervening: *b* = 0.099, *SE* = 0.107). This suggests that post-quiz estimates are regressive toward idiosyncratic ability ($\alpha > 0$) in addition to assessing performance as intended.

Subsequent Performance Illustrates the Problem Regarding Other Correlates of Overconfidence

Although the puzzle suggests that overconfidence predicts future performance, a more parsimonious (and perhaps more probable) explanation is that there are differences in ability, people have self-insight, performance is a noisy measure of ability, and self-evaluations pick up ability in addition to performance. Therefore the measure of overconfidence is confounded with ability and ability is what predicts future performance. A key problem is that many findings in the literature of an association between overconfidence and other correlates use an approach equivalent to that in the puzzle above, but do not sufficiently consider the relevant alternative explanation after incompletely accounting for ability.

Empirical Application II: Relating Overplacement to Status and Social Class

The first application finds that measures of overconfidence predict subsequent performance. This can be readily explained by the confound between measures of overconfidence and ability. While this indicates a problematic conclusion one might draw from the data, it alone does not require us to reinterpret prior findings. Perhaps this whole endeavor is a statistical curiosity with little connection to substantive claims. Using another pair of findings, I examine how the model provides a potential alternative explanation for how differences in overconfidence correlate with other important constructs. The goal is to consider whether this model could plausibly account for the results, not whether it does account for the results, nor whether it rules out the original interpretations.

In both cases, I report a set of parameters compatible with the reported correlations. Two important caveats are in order. First, given the degrees of freedom, the parameter values I report are a subset of those that fit the data, not the only ones that fit the data. Second, and more importantly, one should not interpret the specific values as precise point estimates. This model, like all models, is wrong. Even if the links were correct, it is unlikely they perfectly capture the correct functional form. Instead, it informs qualitative conclusions about the relevant components: good vs. poor self-insight, high vs. low construct validity, high vs. low ambiguity, strong vs. weak relationship. Though I report numeric estimates, the qualitative pattern is what matters.

IIA: Overplacement, Perceived Competence, and Status

In six studies, Anderson et al. (2012) propose that holding ability constant, greater confidence generates higher status in the eyes of others. Study 1 was correlational and well-suited to examine using the current framework.[22] Could ability alone be responsible for the correlation with status in Study 1?

In Study 1, participants took a geography quiz and reported (a) a self-evaluation of their quiz performance percentile, and (b) a self-evaluation of their general geography knowledge percentile. They then repeated the task with a partner. Their partner then rated their perceived competence using a similar percentile measure regarding geography knowledge and status. Self-evaluation was computed as the average of the task-specific and general-knowledge self-evaluation percentiles. Overplacement was measured as the residual of self-evaluation percentile regressed on actual percentile. Overplacement as assessed via the residual was correlated with both partner-rated perceived competence and partner-rated status; these correlations were nearly as strong as the correlations of actual performance with perceived competence and status.

The paper refers to overconfidence as “a genuine yet flawed perception of one’s own abilities” (p719) and notes that the measure of overconfidence “reflects bias in self-perceptions” (p721). Might the results instead be explainable through links between ability and perceived competence and status? I

[22] The model does not attempt to account for the results of the remaining studies, and therefore should not be interpreted as providing a counter-explanation for the paper’s complete set of results.

considered sets of $\rho, \lambda, \alpha, \beta_{PC}$ (representing a causal impact of ability on perceived competence) and β_S (representing a causal impact of ability on rated status). For simplicity, I constrained error variances such that each variable had unit variance.

First consider the case of perfect self-insight, such that $\rho = 1$ and $\tilde{S}_i = S_i$. It is possible to recover nearly-identical correlations to those observed in the data with $\rho = 1$, $\lambda = .65$, $\alpha = .6$, $\beta_{PC} = .6$ and $\beta_S = .5$. Given the number of parameters, multiple configurations fit similarly well (e.g., $\rho = .75, \lambda = .55, \alpha = .75,\ \beta_{PC} = .75, \beta_S = .6$).[23] The empirical correlations and the correlations implied by these two sets of parameters are given in Table 2.

Table 2

Empirical Application IIA: Observed and Modeled Correlations

Measure	Perceived Competence	Status
Actual Performance	.39	.33
Residual	.36	.26
$\rho = 1, \lambda = .65, \alpha = .6$	$\beta_{PC} = .6$	$\beta_S = .5$
Actual Performance	.39	.32
Residual	.34	.28
$\rho = .75, \lambda = .55, \alpha = .75$	$\beta_{PC} = .75$	$\beta_S = .6$
Actual Performance	.41	.33
Residual	.35	.28

This paper is sometimes referenced for its methodology as a recent paper that uses residual scores to assess overconfidence (e.g., Belmi et al. 2020; Cheng et al. 2021; Lyons et al. 2021; Murphy et al. 2015). As a result, in addition to its substantive finding, its methodological approach is uniquely influential. Although the present model has sufficient free parameters to readily fit the data, it still has fewer free parameters than implied by the analysis in the original paper which would necessarily include additional links between beliefs and both perceived competence and status. These are fixed to 0 in the current exploration.

[23] The paper's text also reports the correlation between actual performance and self-evaluation was .56. The correlation implied by this latter set of parameters is identical at .56 ($\lambda\rho\alpha + (1 - \alpha) = .55 * .75 * .75 + .25$).

IIB: Overplacement and Social Class

The second finding builds on the first. Whereas Anderson et al. (2012) argued overconfidence causes status, Belmi et al. (2020) argue social class causes overconfidence due to the pursuit of status. (The model predicts a pattern of correlations, no matter the direction of the causal arrow, and in both cases the target studies assessed correlations.) This example provides a more-stringent test of what the model is able to explain given a larger set of correlations to be fit, including both residual score measures and difference score measures of overconfidence. Could social class cause test-taking ability instead? (Note that ability may merely refer to ability at taking tests, which may be biased by social factors.)

Across four studies including three different tests, the paper uses four measures of social class (self-report, income, education, and parental education), and three ways of measuring overplacement on each test. The prioritized measure is the residual score measure: self-evaluated percentile is regressed on actual performance percentile, and the residual is used as a measure of overplacement. The second measure is the difference score measure: the difference between self-evaluated and actual percentiles. Finally, the paper uses Edwards' (1995) recommended approach of calculating Wilks' lambda from two regressions, one predicting self-evaluated percentile from social class and the other predicting actual percentile from social class. Although I do not detail that approach above, if the model properly captures the correlations with perceived and actual scores in addition to residual and differences, it will necessarily account for the results of this last test too. This suggests this last test would not fully resolve the problem.

Across the four studies, three different instruments are used to assess overplacement. In Study 1, the test was a flashcard game in which participants assess whether two successive images match. In Study 2, the test was a test of general cognitive ability (the Wonderlic Personnel Test). In Studies 3 and 4, the tests were trivia quizzes. Each relies on a combination of skills, including test-taking ability which may be biased by social factors.

The paper reports meta-analytic estimates in Tables 14 and 15. Across studies, the authors find that each measure of social class is more-correlated with perceived performance than actual performance, that each measure of social class correlates with residualized self-evaluations, and that self-reported social

class (though not income, education, nor parental education) correlates with the difference between perceived and actual percentiles. I examine whether plausible parameter values are consistent with these estimates. For convenience, I rely on the paper's estimates in the absence of covariates. These correlations, along with their 95% confidence intervals, are reproduced in the top panel of Table 3.[24]

Table 3

Empirical Application IIB: Observed and Modeled Correlations Between Social Class and Measures of Perceived and Actual Performance and Overconfidence

Measure	Self-Report	Income	Education	Parental
Perceived	.24 [.20, .28]	.11 [.11, .12]	.10 [.01, .18]	.14 [.08, .19]
Actual	.00 [-.11, .12]	.02 [-.12, .21]	.07 [-.03, .16]	.00 [-.12, .12]
Residual	.23 [.20, .26]	.11 [.04, .18]	.08 [-.01, .16]	.13 [.06, .19]
Difference	.13 [.05, .22]	.06 [-.09, .20]	-.01 [-.09, .08]	.08 [-.03, .20]

$\rho = 1, \lambda = .2, \alpha = .7$	$\beta = .3$	$\beta = .15$	$\beta = .1$	$\beta = .15$
Perceived	.23	.11	.08	.11
Actual	.06	.03	.02	.03
Residual	.22	.11	.07	.11
Difference	.16	.08	.05	.08

$\rho = .6, \lambda = .05, \alpha = .95$	$\beta = .4$	$\beta = .2$	$\beta = .15$	$\beta = .2$
Perceived	.23	.11	.09	.11
Actual	.02	.01	.01	.01
Residual	.23	.11	.09	.11
Difference	.15	.08	.06	.08

As with the previous analysis, I examine whether any parameter configurations could generate these results. I again constrain error variances such that each variable had unit variance. The correlations resulting from two such sets of parameters are reported in the lower two panels of Table 3. The middle panel represents a case in which people hold accurate beliefs ($\rho = 1$ and $\tilde{S}_i = S_i$) and there is no latent overplacement. The bottom panel represents a case in which people hold correlated beliefs ($0 < \rho < 1$), so while some people may exhibit overplacement and others underplacement, social class only relates to

[24] Despite working with the meta-analytic estimates, there are substantive differences across studies, and a single set of parameters may not be able to account for every study's results individually. This is not unique to this model, as this applies to any attempt to fit the same parameters across studies. The present model's constraints are simply more explicit. In addition, we need not expect λ, α, or even ρ to be consistent across studies.

beliefs through their respective correlations with test-taking ability. Across the four measures of social class, the parameters relating test-taking ability, beliefs, performance, and self-evaluations (ρ, λ, α) are held constant, but the parameter relating social class to test-taking ability (β) is allowed to vary given the different measures of class.

Note several aspects of this analysis. First and most relevantly, the fitted parameters can adequately account for the entire pattern of meta-analytic results without any role for latent overplacement (middle and bottom panels) and even with perfectly-calibrated beliefs (middle panel, in which $\rho = 1$). The model uses 6 (middle panel) or 7 (bottom panel) free parameters for 16 correlations, though it captures the full pattern at least as well as the proposed theory, which would add 1 to 4 more parameters (i.e., to relate latent overconfidence to each measure of social class). Every correlation is well-within the 95% confidence interval, and more than two-thirds are within the implied 67% confidence interval (i.e., half the 95% confidence interval). This is despite the fact that both the residual score (for every measure of social class) and the difference score (for the self-report measure of social class) implicated a relationship between overplacement and social class.

Second, this account neither rules out the authors' hypothesis nor indicate an impact of social class on any form of ability. It merely presents an alternative interpretation of the data in which social class may not be related to overplacement. It is important to reinforce once again that under this interpretation, ability could refer to mere test-taking ability, not general cognitive ability.

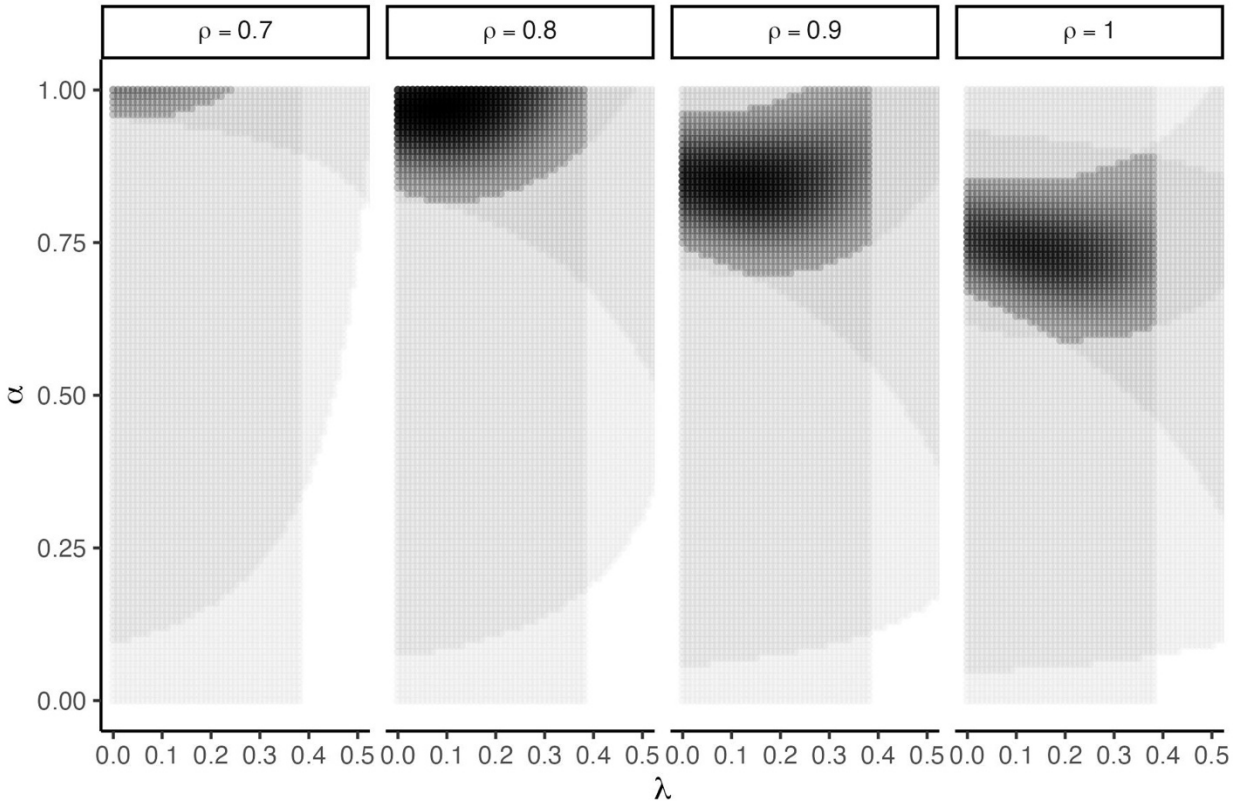
Third, multiple parameter configurations are each nearly-equally compatible with data. The two in the table are not the only two possibilities. Generally speaking, compatible parameters have low λ, high α, and a tradeoff between ρ and β. Figure 4 presents a set of compatible values of $\rho, \lambda,$ and α for self-reported social class given $\beta = 0.3$. Similar values are compatible for income, education, and parental education by using β = 0.15, 0.10, and 0.15, respectively.

Finally, one should probe whether these parameter values are plausible. For low values of β, fitting these data requires high ρ, low λ, and high α. Whether these are themselves plausible depends on

outside knowledge and cannot be resolved using these data alone. (For example, finding little correlation between self-evaluations and actual performance is not sufficient to infer that ρ is low, as it may be compatible with low λ instead.) Alternative approaches are required to rule out the model's null. Table A2 in the supplement provides a possible parameter configuration consistent with findings for 16 articles. Many of the parameter configurations are plausible, but some are not.

Figure 4

Empirical Application IIB: Compatible Parameter Values for Self-Reported Social Class Given $\beta = .3$



Note. Dark shaded regions indicate regions where all of the four target correlations implied by the set of parameters falls within the 95% confidence interval of the corresponding correlation for self-reported social class. Cases where each value is closer to the center of the confidence interval are shaded darkest. Lightly-shaded regions indicate regions where only one, two, or three of the four target correlations fall within the 95% confidence interval.

General Discussion

Both in theory and in practice, widely-used measures of overconfidence are confounded with ability. This applies to both overestimation and overplacement. While the residual and difference score

approaches are driven by different theoretical properties (the first by error in measurement, the second by imperfect construct validity), both are attributable to the fallacy of equating performance with ability. Below I discuss some nuances of the model that are elided over above and end with recommendations on how one might begin to address these concerns.

Model Nuances

Model Specifications

First, the measure of performance is modeled as attributable to ability (S_i, weighted by λ) and luck (v_i). I have largely treated luck as though it is random measurement error. But it is possible to decompose luck into task-specific stable luck (e.g., someone taking a financial literacy quiz may be lucky that in part it assesses institutional trust rather than just financial literacy) as well as temporally inconsistent random luck (e.g., someone was distracted in the moment and misread an answer). Once luck is thus decomposed, it becomes apparent that addressing measurement error alone is unlikely to completely address the problem for the residual; given this representation, reducing measurement error to 0 will not push σ_v^2 to 0.

Second, if there were only measurement error, lack of reliability in the performance measure could lead to lack of validity in residual measure. If that were the case, then increasing reliability of the performance measure could increase validity of the residual measure. I address this in greater detail in the second recommendation.

Third, the variance on beliefs is assumed to be equal to that of ability by scaling the error term relative to ρ. This was to meet a balance of parsimony to avoid proliferation of parameters with realism in considering the implications even when people do not have perfect self-insight (when $\rho \neq 1$). As noted in the footnotes, relaxing this constraint would affect the magnitude of each confound by replacing ρ with $\rho\sqrt{\rho^2 + \sigma_\zeta^2}$. If beliefs tend to vary more than ability, this would mean the confounds reported here tend to be underestimates of the extent of the problem.

Finally, this model represents but one functional form. It is surely not perfectly specified. As one example, it posits a simple linear relationship between ability and beliefs. Yet people who are less-skilled may be less-well-calibrated regarding their own skill (Kruger & Dunning 1999); despite statistical critiques (Krueger & Mueller 2002), there is indeed evidence for the unskilled-and-unaware effect that addresses those critiques (Feld et al. 2017; Jansen et al. 2021). Such a relationship will change the specific quantification of the confound, though it will not alleviate the problem. This is important to note, because it reinforces that the parameter values in the exercises above represent *a* possible null to be ruled out, not *the only* possible null to be ruled out.

Residuals vs. Difference Scores

As noted throughout, different researchers have relied on residuals, or difference scores, or difference scores that in practice act like residuals, and there remains considerable debate about the relative status of the two (see e.g., Belmi et al. 2020; Lyons et al. 2021; Parker & Stone 2014 for recent discussions of the two approaches by researchers interested in the substantive questions). The current model reinforces that neither one can be considered right and the other wrong; their relative biases depend on the relative contributions of ability and luck to the measure of performance.

Other proposals, particularly in the domain of positive self-views and self-enhancement, have proposed the use of polynomial regression, response surface analysis, or condition-based regression analysis (e.g., Edwards 1994; Humberg et al. 2018, 2019) to address the concerns regarding difference scores and residuals. But by themselves, these are not sufficient to account for the concern. This is because in their base form they do not account for measurement error or construct mismatch. Only in conjunction with a strategy to address measurement error will they address reliability (relevant for the residual score measure), and even then construct validity remains a concern (for both the residual score and difference score measures).

Is $\lambda < 1$ Just a Form of Overconfidence?

Throughout, I have returned to the notion that the measure of performance must fully and only measure the target construct to make use of the difference score measure. This matters because the prior

to which people are regressing ($\tilde{S}$) must align with the construct the performance measure assesses (S). A mismatch, as in the case of a financial literacy scale with items that measure trust instead, is equivalent to construct invalidity, or $\lambda < 1$, which leads to the focal problem for difference scores. (Recall that the measure may be highly reliable even with low validity.) A potential critique is that this simply represents a different form of improper confidence: people confidently rely upon a prior that should not apply and so regress to the wrong belief as a result. I argue we cannot be so quick to attribute such a problem to the participant's updating strategy (and therefore a variant of overconfidence) rather than the researcher's inferential strategy.

Consider again the phrenologist introduced earlier. Both the phrenologist and the participant may earnestly believe that the phrenologist is generating a diagnostic measure of intelligence. If the participant is asked how they perform on this measure of intelligence, but they have substantial ambiguity about their own head measurements ($\alpha = 1$), they will report their beliefs about their own intelligence. Of course, their score on the phrenology examination will be unrelated to their intelligence ($\lambda = 0$). If people have partial self-insight ($\rho > 0$), then on average, people who think they received a higher score from the phrenologist than they truly did will be more intelligent. A skeptic may argue: "That is overconfidence! The participant is regressing their self-evaluation of head size to their beliefs about their own intelligence when they should be regressing to their own beliefs about the shape of their head." But in such a case, it would be inappropriate to fault the participant for regressing to the very construct the researcher claims to be measuring with a worthless instrument. Thankfully, most researchers are not phrenologists and are using instruments with greater validity. But greater validity than phrenology is a low bar.

This raises a thorny question regarding whether the effects of using misleading labels for a performance task should be considered overconfidence. Overconfidence should not depend on what the researcher believes a task measures. If we do not accept the overconfidence label in the case described above (in which the participant earnestly believes the measure is measuring the same construct the researcher earnestly believes it measures), we perhaps ought to be cautious in accepting an

overconfidence label in the presence of a mislabeled instrument (when the participant earnestly believes the measure is measuring the construct the researcher tells them it measures; Ehrlinger & Dunning, 2003).

Beyond Overconfidence

This paper focused on overconfidence. But these concerns regarding the use of residual and difference scores represent a broader concern. They apply anytime there are two noisy measures of correlated constructs (or possibly the same construct, as when beliefs are equal to ability) where one is of interest and the other is to be ruled out. This is directly relevant in the literature on self-enhancement (Taylor & Brown, 1988; Colvin et al., 1995) and correlates of narcissism (see Grijalva & Zhang 2016 for a meta-analysis considering residual scores vs. difference scores in the study of correlates of narcissism). But it may also speak to other disparate applications with structural similarities (e.g., as when a researcher has two measures of product quality and is interested in the role of one, net of the other). While this manuscript addresses overconfidence, the problem is a more-general one.

Recommendations

What solutions are available? No easy ones. But the absence of an easy solution does not provide cover to carry on as though there is no problem. I suggest four recommendations. Used by themselves or in concert, they have the potential to reduce the extent of problematic inferences.

1. Use Reliable, Valid Measures

Most importantly, this serves as a(nother) call to ensure the use of reliable and valid measures. This recommendation ought to go without saying: no one touts the use of an unreliable or invalid measure. But given the strong theoretical reasons to believe there ought to be a confound without such measures, this research accentuates the importance of using them. This is particularly important given that performance measures are often moderately reliable at best. While reliabilities are not reported in Belmi et al., Anderson et al. report scale reliability of 0.66 for a narrowly-scoped geography quiz. For typical trivia quizzes, Krueger and Mueller (2002) report split-half correlations ranging from 0.17 to 0.56 and Burson et al. (2006) report split-half correlations ranging from -0.24 to 0.52. I note these examples because the data were readily reported, not because they were uniquely low.

2. Account for Measurement Error

To provide an unbiased test, it is useful to recall the conditions under which there is no bias for the role of self-evaluation when controlling for performance (i.e., when using the residual measure or multiple regression measure). The bias is eliminated if any of three conditions hold: (a) $\alpha = 0$, meaning there is no ambiguity and participants have no reason to regress their self-evaluations towards their prior beliefs, (b) $\rho = 0$, meaning latent beliefs are unrelated to latent ability, or (c) when $\frac{\lambda^2}{\lambda^2+\sigma_v^2} = 1$, meaning there is no error and performance is at least partially related to ability.

This latter concern is affected by a classic problem in which measurement error in one independent variable (performance) biases both its own coefficient and the coefficients of measures of correlated latent variables. Possible solutions to address this include structural equation models (e.g., Kline 2005) and errors-in-variables (Culpepper and Aguinis 2011). These approaches only help to the extent that σ_v^2 only represents measurement error and not other constructs (as with stable luck described above). The supplement presents additional details regarding both approaches. Using a third empirical application (Parker et al's 2012 study of the relationship between inappropriate confidence and financial planning), I reanalyze the data, first using a structural equation model, and separately using an adjustment for errors-in-variables. In this application, the standard analysis finds an apparent correlation between financial planning and overconfidence; accounting for measurement error largely eliminates it.

3. Bound the Parameter Space

Rather than attempting to rule out this alternative explanation, researchers may instead relax the strength of their claims by acknowledging the conditions under which it may hold. Given the ability to characterize the magnitude of the bias, one can plausibly specify parameter configurations that could and could not account for the observed results under the null model. In some cases, no parameter values may be able to account for the full set of correlations without a role for overconfidence. In other cases, there may be parameter configurations which could account for the observed results but are implausible: while they are mathematically plausible, they may be ruled out based on theory.

Consider Figure 4. Figure 4 depicts the parameter values, given $\beta = 0.3$, that would reconcile the results regarding self-reported social class with the model presented here. As can be seen by the crosscutting lightly shaded areas, while many parameters may be compatible with any one correlation, a much smaller set is compatible with the full set of correlations. By constraining the model to have the same parameters across each measure of social class, except for β which varies with the measure, the plausible shaded region narrows further. In this case, a region of parameter values remains. In some cases, even when compatible values remain, some may be implausible.

In other cases, one can rule out the alternative explanation altogether. First, if there truly is no relationship between ability and the candidate correlate, then although the measure is confounded, the confound has no bite to it. Note that no correlation between *performance* and the outcome measure of interest is not sufficient: such a lack of correspondence could merely indicate that performance is a poor measure of ability even if it is a reliable measure of something else. This would again lead to a biased estimate of the effect of overconfidence.

Second, if the relationship between ability and the outcome measure of interest and the relationship between residualized overconfidence and the outcome measure of interest have opposite signs, the core bias described here could not account for such a pattern of results. For example, as depicted in Table A2 in the supplement, while the results of Lyons et al. (2021) regarding false news could be accounted for with the reported correlations, it would require an unlikely (and possibly implausible) sign on β, suggesting it is unlikely to account for the results. This does not mean that the bias is inconsequential: indeed, it may suggest that the magnitude of the relationship between overconfidence and the outcome measure of interest is underestimated. Note this is not guaranteed to hold for the difference score if beliefs are imperfectly correlated with ability due to the patterns displayed in Figure 3.

To implement this, one might consider whether overconfidence on one set of items predicts performance on a separate set of items, as in the first empirical application. If it does, this would suggest either that: (a) the instrument has the problems described here, or (b) that overconfidence varies with

ability (an inverted Dunning-Kruger effect; Burson et al. 2006). But several notes are critical here. First, the items used to calculate performance must be distinct from those used to calculate overconfidence. Second, this analysis can provide evidence of a problem, but by itself it cannot provide evidence of the lack of a problem (though it may help to bound parameters). Third, one must not fall into a trap of accepting the null hypothesis, particularly when considering uncertainty around the estimates. In establishing such bounds, it is important to consider the uncertainty regarding one's estimate, not merely the point estimate itself. Further, and as described above, these bounds are with respect to this rather strictly-specified null model. Other null models (e.g., one in which an unskilled-and-unaware effect holds but ability is the only correlate of behavior) may not be so readily ruled out.

4. Use Alternative Measurement Approaches

Finally, one may opt to use a different measurement approach altogether. A variety of measures have been introduced which may be less susceptible to the problems described above. Direct measures of overclaiming (Paulhus et al. 2003), e.g., indicating one recognizes people, objects, or events that do not exist, present an intriguing possibility for reducing the problems described here. One interpretation of such measures in terms of the current model is that ability is known to be constant and minimal (i.e., no one has the requisite knowledge to recognize things that do not exist.) Yet concerns remain regarding the role of inferences in the face of ambiguity. As a result of their ambiguity regarding individual items, people likely rely on their priors, which may again lead high-ability people to be more likely to overclaim than low-ability people in certain circumstances.

Similarly, Binnendyk and Pennycook (2023) present a measure of individual differences in general overconfidence. These are based on estimated performance on a task for which performance is at or near chance and difficult to ascertain. One interpretation of this measure in terms of the current model is that ability at this task ought to be unrelated to other correlates of interest. As with the Paulhus et al. (2003) measure, there is reason to be more-optimistic regarding this task, and to potentially prefer it over the other methods described here, but it may not completely address the problems laid out here. To the extent that some people accurately believe themselves to be generally more successful than others at a

variety of tasks, the same problem could persist. In such a case, the task may operate as a poor measure of the underlying construct that participants rely upon (see "Is $\lambda < 1$ Just a Form of Overconfidence?"). The plausible range of parameters may lead to smaller biases in such a case and so be of negligible concern.

Summary and Conclusion

Research suggests that overconfidence on particular tasks varies across people. Yet widely-used measures of overconfidence that are used to study its correlates are confounded with the very thing they aim to rule out: ability. This is because measures of performance are imperfect, so accounting for performance is insufficient to account for ability. Given ambiguity regarding performance, measures of confidence ought to regress towards prior beliefs about ability even when they are intended to be self-evaluations of task performance. Because performance itself is an imperfect measure, the variance of self-evaluation attributable to ability is not fully partialed out. The result is that both residual and difference overconfidence measures are confounded with ability. In an idealized model, this bias can be quantified.

These confounds imply that it is possible to observe surprising results. In the first application, I find overconfidence predicts later performance even after several intervening tasks. They also provide an alternative interpretation of findings from the literature: one need not posit a role for (or even the presence of) differences in overconfidence when considering relations with social status and social class. Instead, the entire pattern of results could be driven test-taking ability alone. If researchers are willing to make strong assumptions regarding construct validity and estimate or assume reliability of each measure, it is possible to address these concerns through structural equation modeling or error-in-variables adjustments. However, these partial solutions are not an automatic panacea, as a number of complications may arise regarding construct validity and unstable estimates. Instead, design-based solutions (e.g., experimental manipulations or using other measurement approaches) or accepting alternative interpretations of the results (i.e., plausible parameter configurations) may ultimately prove necessary. This work may serve as an stark reminder to further improve our collective attempts to measure differences in overconfidence (whether stable or transient) and their true associations with traits, decisions, and behaviors.

References

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research, 27*(2), 123-156.

Anderson, A., Baker, F., & Robinson, D. T. (2017). Precautionary savings, retirement planning and misperceptions of financial literacy. *Journal of Financial Economics*, *126*(2), 383-398.

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718-735.

Ames, D. R., & Kammrath, L. K. (2004). Mind-reading and metacognition: Narcissism, not actual competence, predicts self-estimated ability. *Journal of Nonverbal Behavior*, *28*(3), 187-209.

Avdeenko, A., Bohne, A., & Frölich, M. (2019). Linking savings behavior, confidence and individual feedback: A field experiment in Ethiopia. *Journal of Economic Behavior & Organization*, *167*, 122-151.

Barranti, M., Carlson, E. N., & Côté, S. (2017). How to test questions about similarity in personality and social psychology research: Description and empirical demonstration of response surface analysis. *Social Psychological and Personality Science*, *8*(4), 465-475.

Belmi, P., Neale, M. A., Reiff, D., & Ulfe, R. (2020). The social advantage of miscalibrated individuals: The relationship between social class and overconfidence and its implications for class-based inequality. *Journal of Personality and Social Psychology*, *118*(2), 254-284.

Benoît, J. P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, *79*(5), 1591-1625.

Benoît, J. P., Dubra, J., & Moore, D. A. (2015). Does the better-than-average effect show that people are overconfident?: Two experiments. *Journal of the European Economic Association*, *13*(2), 293-329.

Binnendyk, J., & Pennycook, G. (2024). Individual differences in overconfidence: A new measurement approach. https://doi.org/10.31234/osf.io/ugb3s

Birnbaum, M. H., & Mellers, B. A. (1979). Stimulus recognition may mediate exposure effects. *Journal of Personality and Social Psychology*, *37*(3), 391-394.

Blair G, Cooper J, Coppock A, Humphreys M, Sonnet L (2022). *estimatr: Fast Estimators for Design-Based Inference*. https://declaredesign.org/r/estimatr/, https://github.com/DeclareDesign/estimatr.

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*(1), 60-77.

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making*, *17*(4), 297-311.

Carlson, J. P., Vincent, L. H., Hardesty, D. M., & Bearden, W. O. (2009). Objective and subjective knowledge relationships: A quantitative analysis of consumer research findings. *Journal of Consumer Research*, *35*(5), 864-876.

Cheng, J. T., Anderson, C., Tenney, E. R., Brion, S., Moore, D. A., & Logg, J. M. (2021). The social transmission of overconfidence. *Journal of Experimental Psychology: General*, *150*(1), 157.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, *10*(4), 637-666.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: negative implications for mental health. *Journal of Personality and Social Psychology*, *68*(6), 1152.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we?. *Psychological Bulletin*, *74*(1), 68.

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, *16*(2), 166-178.

Edwards, J. R. (1994). Regression analysis as an alternative to difference scores. *Journal of Management*, *20*(3), 683-689.

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management journal*, *36*(6), 1577-1613.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*(1), 5-17.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672-680.

Feld, J., Sauermann, J., & De Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of Behavioral and Experimental Economics*, *68*, 18-24.

Fiedler, K. (2021). Suppressor effects in self-enhancement research: A critical comment on condition-based regression analysis. *Journal of Personality and Social Psychology*, *121*(4), 792-795.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(443), 1-9.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506.

Gillen, B., Snowberg, E., & Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *Journal of Political Economy*, *127*(4), 1826-1863.

Griffin, D., Murray, S., & Gonzalez, R. (1999). Difference score correlations in relationship research: A conceptual primer. *Personal Relationships*, *6*(4), 505-518.

Grinblatt, M., & Keloharju, M. (2009). Sensation seeking, overconfidence, and trading activity. *The Journal of Finance*, *64*(2), 549-578.

Healy, P. J., & Moore, D. A. (2007). Bayesian overconfidence. *Available at SSRN 1001820*.

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., Küfner, A. C., van Zalk, M. H., Denissen, J. J. A., Nestler, S. & Back, M. D. (2019). Is accurate, positive, or inflated self-perception most advantageous for psychological adjustment? A competitive test of key

hypotheses. *Journal of Personality and Social Psychology*, *116*(5), 835-859.

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., van Zalk, M. H., Denissen, J. J. A., Nestler, S., & Back, M. D. (2018a). Enhanced versus simply positive: A new condition-based regression analysis to disentangle effects of self-enhancement from effects of positivity of self-view. *Journal of Personality and Social Psychology*, *114*(2), 303-322.

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., van Zalk, M. H., Denissen, J. J. A., Nestler, S., & Back, M. D. (2018b). Why Condition-Based Regression Analysis (CRA) is Indeed a Valid Test of Self-Enhancement Effects: A Response to Krueger et al. *Collabra: Psychology*, *4*(1), 26.

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., van Zalk, M. H., ... & Back, M. D. (2022). The true role that suppressor effects play in condition-based regression analysis: None. A reply to Fiedler (2021). *Journal of Personality and Social Psychology*, *123*(4), 884-888.

Humberg, S., Nestler, S., & Back, M. D. (2019). Response surface analysis in personality and social psychology: Checklist and clarifications for the case of congruence hypotheses. *Social Psychological and Personality Science*, *10*(3), 409-419.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756-763.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*(1), 206.

Johns, G. (1981). Difference score measures of organizational behavior variables: A critique. *Organizational Behavior and Human Performance*, *27*(3), 443-463.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*(2), 226-246.

Kahneman, D. (1965). Control of spurious association and the reliability of the controlled

variable. *Psychological Bulletin*, *64*(5), 326-329.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Ke, D. (2021). Who wears the pants? Gender identity norms and intrahousehold financial decision-making. *The Journal of Finance*, *76*(3), 1389-1425.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216-247.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*. 2nd ed. New York: Guilford.

Kramer, M. M. (2016). Financial literacy, confidence and financial advice seeking. *Journal of Economic Behavior & Organization*, *131*, 198-217.

Krueger, J. I., Heck, P. R., & Asendorpf, J. B. (2017). Self-enhancement: Conceptualization and assessment. *Collabra: Psychology*, *3*(1), 28.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*(2), 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*(6), 1121-1134.

Landier, A., & Thesmar, D. (2008). Financial contracting with optimistic entrepreneurs. *The Review of Financial Studies*, *22*(1), 117-150.

Larkin, I., & Leider, S. (2012). Incentive schemes, sorting, and behavioral biases of employees: Experimental evidence. *American Economic Journal: Microeconomics*, *4*(2), 184-214.

Lawson, M. A., Larrick, R. P., & Soll, J. B. (2023). Individual differences in overconfidence and their psychological bases. http://dx.doi.org/10.2139/ssrn.4558486

Li, S. & Moore, D. (2024). Is overconfidence an individual difference? *Behavioral Decision Research in*

Management.

Lord, F. M. (1956). The measurement of growth. *ETS Research Bulletin Series*, *1956*(1), i-22.

Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement*, *18*(3), 437-451.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, *55*(290), 307-321.

Lusardi, A., & Mitchell, O. S. (2017). How ordinary consumers make complex economic decisions: Financial literacy and retirement readiness. *Quarterly Journal of Finance, 7*(3), 1750008.

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, *118*(23), e2019527118.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*(2), 265-287.

McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement*, *18*(1), 47-55.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502-517.

Moore, D. A., & Dev, A. S. (2017). Individual differences in overconfidence. *Encyclopedia of Personality and Individual Differences. Springer. Retrieved from http://osf. io/hzk6q*.

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*(8), e12331.

Moore, D., & Swift, S. A. (2011). The three faces of overconfidence in organizations. In *Social Psychology and Organizations*, 179-216.

Moorman, C., Diehl, K., Brinberg, D., & Kidwell, B. (2004). Subjective knowledge, search locations, and consumer choice. *Journal of Consumer Research*, *31*(3), 673-680.

Murphy, S. C., von Hippel, W., Dubbs, S. L., Angilletta Jr, M. J., Wilson, R. S., Trivers, R., & Barlow, F.

K. (2015). The role of overconfidence in romantic desirability and competition. *Personality and Social Psychology Bulletin*, *41*(8), 1036-1052.

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., & Wirth, K. (2016). Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy*, *9*(1), 4.

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: numeracy underlies better alternatives. *Numeracy*, *10*(1), 4.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric Theory New York. *NY: McGraw-Hill*.

Parker, A. M., Bruin De Bruin, W., Yoong, J., & Willis, R. (2012). Inappropriate confidence and retirement planning: Four studies with a national sample. *Journal of Behavioral Decision Making*, *25*(4), 382-389.

Parker, A. M., & Stone, E. R. (2014). Identifying the effects of unjustified confidence versus overconfidence: Lessons learned from two analytic methods. *Journal of Behavioral Decision Making*, *27*(2), 134-145.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological bulletin*, *92*(3), 726.

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*(6), 971-986.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*(2), 143-148.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193-210.

Thomson, G. H. (1924). A formula to correct for the effect of errors of measurement on the correlation of initial values with gains. *Journal of Experimental Psychology*, *7*(4), 321.

Wall, T. D., & Payne, R. (1973). Are deficiency scores deficient?. *Journal of Applied Psychology*, *58*(3), 322.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, *11*(3), e0152719.

Zuckerman, M., & Knee, C. R. (1996). The relation between overly positive self-evaluation and adjustment: a comment on Colvin, Block, and Funder (1995). *Journal of Personality and Social Psychology*, *70*(6), 1250-1.