Widely-Used Measures of Overconfidence Are Confounded With Ability

Stephen A. Spiller

UCLA Anderson School of Management

July 11, 2026

Author Note

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

All simulation and analysis code is available at https://researchbox.org/1597. Data and materials from prior investigations are available via the cited articles. I have no conflicts of interest to disclose.

The author thanks Jack Soll, the AE, three anonymous reviewers, John Lynch, Don Moore, and Oleg Urminsky for valuable feedback on this research, as well as the UCLA BDM Lab, seminar participants at Chicago, Virginia, UCR, Monash, Florida, BU, Northwestern, Cornell, OSU, MIT, SMU, Yale, Texas, Tulane, Stanford, and INSEAD, and conference participants at the JDM Winter Symposium, LAX Workshop, BDRM, and the ACR, SCP, and SJDM annual meetings.

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

Measures of overconfidence are used both to assess biases in people's beliefs regarding their own abilities and to identify correlates of those biases. Proposed correlates of overconfidence include narcissism, social status, (low) task anxiety, savings, planning, and many more. Research on these correlates typically explicitly or implicitly claims the associations with overconfidence are not attributable to ability (i.e., latent inputs enabling successful task performance, including inputs that vary across people, states, or contexts). I argue that several typical approaches used to measure overconfidence are not sufficient to support such claims. This argument follows from two underappreciated properties of these 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 incorporate prior beliefs about their ability. These properties imply multiple commonly-used measures of overconfidence are confounded with ability. I support these analytical results by reexamining previously-published findings. In the first analysis, I find overconfidence predicts subsequent performance, consistent with measures of overconfidence serving as a signal of ability. In the second set of analyses, I find some purported associations between overconfidence and other constructs could be attributed to ability instead. I close with recommendations to recognize and reduce the potential for mistaken inferences. When research indicates that differences in overconfidence are associated with other behaviors, beliefs, or evaluations, researchers should account for the possibility that differences in ability could provide a sufficient explanation.

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

Overconfidence refers to the state of holding positively-biased beliefs about one's own ability on one or more dimensions. It is reliably reproduced in academic research, written about 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). As a result, research in many disciplines has documented correlates of overconfidence.[1] These include narcissism (Ames and Kammrath 2004; Campbell et al. 2004; John and 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 and Leider 2012), susceptibility to false news (Lyons et al. 2021), search behavior (Moorman et al. 2004), stock ownership (Ke, 2021), short-term debt (Landier and Thesmar 2008), and many more.

Do associations with measures of overconfidence necessarily imply associations with true latent overconfidence (i.e., beliefs that exceed ability)? No. In this research, I build on a key insight from Moore and Healy's (2008) model of overconfidence: self-evaluations of performance are sensitive to one's own prior beliefs.[2] By modeling the consequences of differences in ability (i.e., latent inputs enabling successful task performance, including inputs that vary across people, states, or contexts), I show that under reasonable assumptions, widely-used measures of overconfidence are confounded with ability. Reanalysis of published findings indicate this confound could be consequential. I do not argue that people are unbiased, nor that everyone is equally overconfident, nor that overconfidence is unrelated to important outcomes. Rather, I propose a plausible null model that implies that causes and consequences of ability would correlate with measures of overconfidence even if there were no latent overconfidence.

The key issue is that although researchers intend to account for latent ability, they instead account for observed performance. If people incorporate even partially-calibrated prior beliefs in their self-

[1] Researchers use a variety of labels to refer to the concepts assessed using the measures discussed here. These include overconfidence, biased self-evaluations, biased self-assessments, unjustified confidence, inappropriate confidence, subjective knowledge when controlling for objective knowledge, and self-enhancement.

[2] In Moore and Healy's (2008) model, when performance is ambiguous, prior beliefs about a task's simplicity leads to overestimation on hard tasks, underestimation on easy tasks, overplacement on easy tasks, and underplacement on hard tasks. Their model explains variance across *tasks*. My model considers variance across *people*.

evaluations (e.g., they use even imperfect Bayesian-like reasoning), their self-evaluations will directly assess ability. Here, ability refers to a broad set of latent inputs that enable successful task performance, including those that may be relatively stable over time (e.g., knowledge acquired through education) as well as those that may be sufficiently stable during the task, evaluation, and outcome, but not on a longer time scale (e.g., temporary lack of concentration due to a cold). Because self-evaluations reflect ability and accounting for performance is insufficient to account for ability, the resulting associations between measures of overconfidence and other constructs can be systematically biased by ability. The biases depend on how overconfidence is measured and how strongly the other constructs correlate with ability. Although many reports claim that *overconfidence* on a task is associated with various correlates, that evidence could instead indicate that *ability* for the task is associated with those correlates. This alternative interpretation frequently escapes notice. In contrast to other uses of noisy covariates, when studying overconfidence: (a) it is not always apparent that performance is a noisy measure of ability rather than the focus of inquiry, and (b) it is not always apparent that self-evaluations ought to regress to one's prior beliefs. This confound can be particularly pernicious because the role of ability is often considered and explicitly ruled out as an alternative explanation due to how overconfidence is measured.

I begin by detailing a typical paradigm used to measure differences in overconfidence, variations on that theme, and why each one results in a bias. I next present a mathematical model to formalize and quantify this bias. I then examine whether these theoretical concerns can impact inferences from real 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 that measures of overconfidence predict later 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 one could recreate the reported correlations through a confound with ability even if there were no overconfidence. I close with recommendations to recognize and ameliorate the problem, even if eliminating the problem may be a nearly unattainable target.

Theme and Variations: How Differences in Overconfidence Are Measured

Research on overconfidence's correlates has used a wide array of different measures. I consider cases of *overestimation* of absolute performance and *overplacement* of relative performance for which there is a reality criterion to compare against (Moore and Healy 2008; I do not address *overprecision*, or excessive certainty in one's knowledge). When using a single performance measure and a single self-evaluation, possible combinations of the factors below result in more than 20 ways that associations with overconfidence may be assessed. Adding cases in which self-evaluations reflect future expectations rather than past performance or including item-by-item assessments of confidence would increase it further. Each of these approaches results in a potential confound with ability for the focal task. The 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. In each case, I focus on variation in overconfidence across people on a given task. Such variation may coexist with mean-level overconfidence, underconfidence, or calibrated confidence, as correlations with outside constructs are unaffected by shifts in the mean.

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 answered correctly). The researcher then regresses self-evaluations of performance as the dependent variable on objective performance as the independent variable. The residuals from this regression, reflecting how much higher or lower self-evaluations are than is expected in the sample given objective performance, are used as measures 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.[3] 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.[4]

Using the *control* approach, researchers regress the proposed correlate of overconfidence on self-evaluations and control for performance. 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 resulting in smaller standard errors.

Using the *difference* approach, researchers subtract the measure of objective performance from the self-evaluation of performance and keep the difference as a measure of overconfidence. Researchers will sometimes use a difference score and 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, so the estimate is equivalent to that from the residual approach. When researchers control for performance in analysis of a proposed correlate, whether the focal variable is the residual, self-evaluation, or difference, the focal coefficient is precisely the same.

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.

[3] Parker and Stone (2014) distinguish *unjustified confidence* (residual) and *overconfidence* (difference score).
[4] As a result, participants with negative residuals cannot be said to exhibit a self-diminishment bias: they may instead merely exhibit less of a self-enhancement bias or no bias at all (cf. John and Robins 1994, p214).

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 of performance in a different metric (e.g., a subjective 1-7 scale). If the self-evaluation is in a different metric, overconfidence may be assessed using the residual or covariate method, but not difference scores (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-evaluations may instead be evaluations of ability. 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 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 in principle be assessed in the same metric. For example, 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.

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. Moreover, self-evaluations might be assessed on an item-by-item basis (thereby enabling measures of not just *bias* but also *sensitivity*, or the ability to discern

correct answers from incorrect answers; e.g., Burson et al. 2006; Fleming and Lau 2014; Stankov and Crawford 1996; Rahnev 2025; Van Marcke et al. 2024).[5] Each of the measurement approaches described above could result in a measure of overconfidence that is confounded with ability. As a result, using any of these measures has the potential to bias estimates of the correlation between overconfidence and other variables. The confound and potential for bias are 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.

Such measures of overconfidence are sometimes interpreted as measures of stable, general individual differences in overconfidence. For example, in finance, measures of overconfidence on unrelated tasks are used as correlates of trading activity or stock ownership (e.g., Grinblatt and 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, measures of overconfidence are correlated with perceived competence and social status as rated by randomly-paired partners (Anderson et al. 2012). Here, no assumption of stability or consistency is required. Whether there are stable general individual differences in overconfidence is a topic of ongoing debate (Binnendyk et al. 2025; for arguments and evidence against, see e.g., Li et al. 2025; Moore and Dev 2017; Moore and Swift 2011; Moore and Schatz 2017; for evidence in favor, see e.g., Lawson et al. 2024, 2025; Binnendyk and Pennycook 2024). The analysis I present here addresses measured differences in overconfidence, whether they are stable or transient, general or domain-specific.

Is the use of either residuals or difference scores as measures of overconfidence unusual? A systematic (though narrowly scoped and non-representative) search of articles published since 2000 in a

[5] Research on metacognition uses item-level evaluations to construct measures of *bias* and *sensitivity* (e.g., Fleming and Lau 2014, Rahnev 2025). Overconfidence corresponds to bias. If sensitivity were assessed using tasks on which there is *intra*personal variability in ability, an analogous concern may apply. But for tasks that draw on the same ability, the argument proposed here does not directly speak to such measures of sensitivity (see e.g., Gigerenzer et al. 1991, Juslin 1994, Klayman et al. 1999 for cue-based models of overconfidence addressing sensitivity).

set of influential journals revealed that 31 of 60 identified articles used at least one such measure; see Supplement for details. The data in these articles drew from laboratory samples, large surveys, and administrative data. Many of the other 29 articles used options-based measures of CEO overconfidence (which I return to in the General Discussion), assessed overprecision, or developed an analytical model without data. Table A1 in the Supplement indicates whether each of the relevant 31 articles used residuals, difference scores (some of which controlled for performance, making their usage equivalent to residuals), both residuals and difference scores, or measures that are equivalent to either (e.g., a measure of confidence without accounting for performance). 28 used difference scores or an equivalent measure and 20 used residuals or an equivalent measure. This suggests each measure is commonly used. In principle, the present concern applies to the interpretation of each measure, as the correlations between overconfidence and other measures could be due to ability (or other latent input for task performance) rather than latent overconfidence. I revisit this possibility following the empirical applications. Together, this search and classification suggests the present concern could apply widely.

The Potential Confound

Before presenting a formal proof of the problem and simulation results, I provide an informal description. The problem arises from four properties common to the paradigms described above. First, people typically differ in relevant ability. Here, ability refers to any latent task-relevant input that remains sufficiently stable throughout the task and may include mere test-taking ability, general cognitive ability, domain-specific knowledge, capacity for concentration, etc. Second, people typically have at least partial insight into their ability. Third, performance is typically an imperfect measure of ability: it includes 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 have an imperfect sense of how they performed.

Because performance is ambiguous, principles of Bayesian reasoning require that self-evaluations of performance regress toward people's prior beliefs (Moore and Healy 2008). This leads to regression toward ability on average under the weak assumption that people's beliefs about their own ability are correlated with their true ability. Self-evaluations are therefore a noisy weighted average of ability and

performance. If a researcher observes that self-evaluations exceed performance, that signals that ability likely 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%, an observer may reasonably infer the first test-taker has higher ability than the second.

When 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). Because self-evaluations of performance regress toward ability, controlling for performance will result in residual variation in self-evaluation that is attributable to ability. Although the residuals must be uncorrelated with performance, they are still correlated with true ability. Measured overconfidence, computed via residuals or by controlling for performance, is therefore confounded with ability. When people self-evaluate ability rather than performance (e.g., when reporting expectations about future performance), the confound is more severe because the measure directly assesses ability rather than merely being contaminated by it.

The concern above applies when self-evaluations are residualized or the analysis controls for performance. A variant applies when researchers use difference scores instead. 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 well, and people who are very low in ability will perform moderately poorly. As a result, the difference score measure will also be confounded with ability. Consider again a 10-item quiz designed to measure financial literacy. 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, got 7 correct (5 of the 6 financial literacy questions and 2 of the 4 trust questions), and due to ambiguity reported that they got 8 correct. A less-literate but more-trusting participant expected to get 5 answers correct, got 7 correct (3 of the 6 financial literacy questions and all 4 of the trust questions), and due to the 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.[6,7] Private information about one's own ability affects evaluations of one's own performance but not evaluations of others' performance, so the same confound holds for overplacement.

This confound implies that measures of overconfidence act as imperfect measures of ability. If latent ability affects some outcome and everyone maintains perfectly calibrated beliefs about their own ability (i.e., no one exhibits latent under- or overconfidence), the overconfidence measure may appear to be a useful predictor beyond performance. This is not because latent overconfidence plays a role (or even exists), but rather because measured overconfidence provides a second noisy measure of ability.

A Simulated Example of the Problem

Figure 1 depicts the problem. In this example, participants vary in ability, have perfect self-insight into their own ability, and have ambiguity about their performance. In evaluating their performance, they rely on a noisy signal of their 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 latent overconfidence in this example. 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 a noisy and regressive relationship between ability and performance for 100 people. The 25 highest-ability people are depicted as filled circles (with one exception shown as a red triangle to enable tracking across panels) and the 25 lowest-ability people are depicted as open circles; the 50 middle-ability people are depicted as crossed circles. The solid line depicts the 45-degree line.

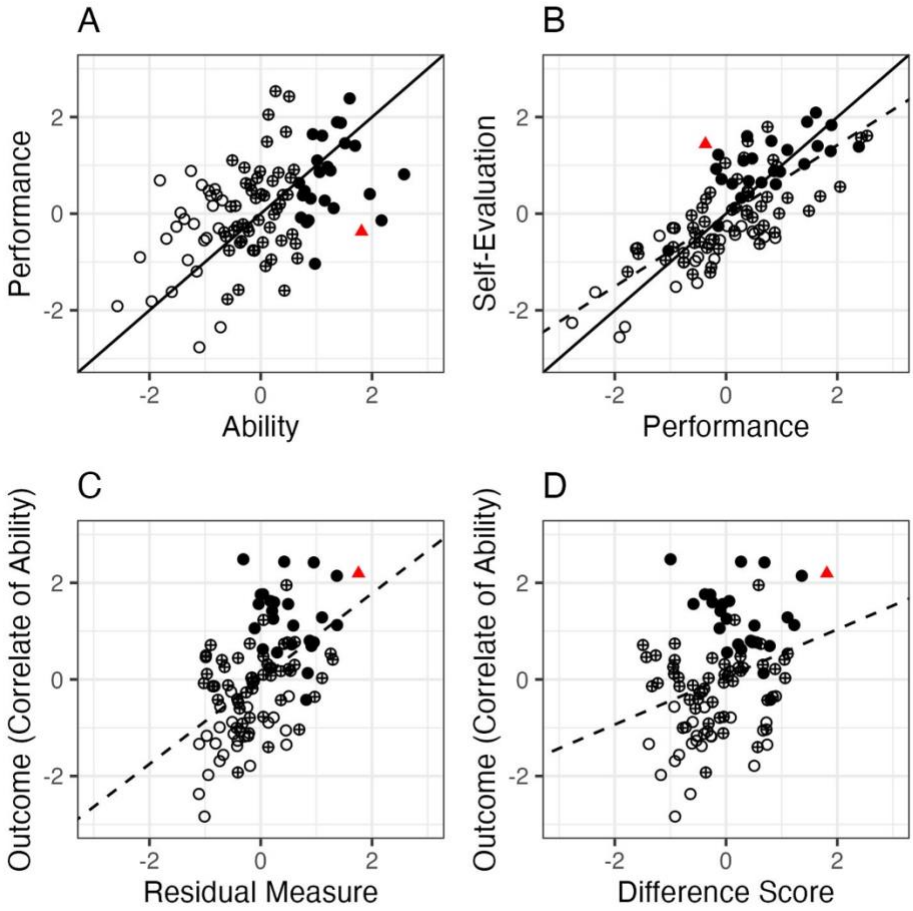
---

[6] Imperfect sampling and the role of error in the use of discrepancy scores are repeated themes in overconfidence (e.g., Burson et al. 2006; Erev et al. 1994; Gigerenzer et al. 1991; Juslin 1994; Klayman et al. 1999). The focus of these critiques has been imperfect calibration and findings regarding aggregate overconfidence rather than implications for individual-level measures of overconfidence. Concerns about inappropriate inferences regarding true scores when relying on measured scores are an old problem in measurement (e.g., Cochran 1968; Cronbach and Furby 1970; Lord 1956, 1958, 1960; McNemar 1958; Rogosa et al. 1982; Thomson 1924), leading to an array of approaches to attempt to recover unbiased coefficient estimates (e.g., Cronbach and Furby 1970; Culpepper and Aguinis 2011; Kline 2005; Fuller 1987). Why has this concern not been central in discussions of measures of overconfidence? I suspect overconfidence measures are susceptible to an illusion that because performance is the target of self-evaluation, latent ability does not matter. But because performance is noisy and ambiguous, requiring people to use their prior beliefs, the conclusion that latent ability does not matter is incorrect.

[7] The fact that the participants use beliefs about financial literacy alone rather than also incorporating trust should not be interpreted as overconfidence if they rely on the same construct the researchers believe they are measuring. I return to this consideration toward the end of the paper.

Figure 1

Depiction of How Measures of Overconfidence Are Confounded With Ability



Note. One of the highest-ability individuals is depicted as a red triangle to enable matching across panels. The other 24 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 detailed later are $\rho = 1, \lambda = .5, \sigma_v^2 = 1, \alpha = .5, \sigma_v^2 = .25, \beta = 1, \sigma_\epsilon^2 = .25$. β represents the coefficient on ability predicting outcome, and σ_ϵ^2 represents the error variance of the outcome.

Panel B depicts self-evaluations as a function of performance when self-evaluations are noisy and regressive toward ability. The 50 individuals classified as overconfident by the residual score (i.e., those above the dashed best-fit line) include 22 of the 25 highest-ability individuals and only 5 of the 25 lowest-ability individuals. The 50 individuals classified as underconfident by the residual score (i.e., those below the dashed best-fit line) include 20 of the 25 lowest-ability individuals and only 3 of the 25 highest-ability individuals. Similarly, the 44 individuals classified as overconfident by the difference score (i.e., those

above the solid 45-degree line) include 17 of the 25 highest-ability individuals and only 6 of the 25 lowest-ability individuals. The 56 individuals classified as underconfident by the difference score (i.e., those below the solid 45-degree line) include 19 of the 25 lowest-ability individuals and only 8 of the 25 highest-ability individuals. In this example, classifying individuals by overconfidence effectively (but imperfectly) classifies them by ability ($r_{ability,residual} = 0.60$, $r_{ability,difference} = 0.37$).

Panels C and D plot the correspondence between an arbitrary correlate of ability and the residual measure (C) and difference score (D), respectively. The correlate of ability is positively correlated with each measure of overconfidence, even though in this example there is no true latent overconfidence and both measures of overconfidence account for performance.

Modeling the Bias in Measures of Overconfidence

This section formalizes the model. A straightforward extension of Moore and Healy's (2008) model of overconfidence permits a focus on differences between people, so I adapt their notation.[8]

Model Setup

Latent Ability and Beliefs

People, denoted by *i*, vary in some target latent ability or skill, S. For each person *i*, this ability S_i is a random variable with an unknown distribution with mean 0 and variance $\sigma_S^2 = 1$. Depending on the application, ability may represent mere test-taking ability, general cognitive ability, domain-specific knowledge, current caffeination status, etc.

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

$$\tilde{S}_i = \rho S_i + \zeta_i \quad (1)$$

where $0 \leq \rho \leq 1$ and ζ is independently drawn from any distribution with mean 0 and variance $\sigma_\zeta^2 = 1 - \rho^2$, such that $\tilde{S}$ has a variance $\sigma_{\tilde{S}}^2$ of 1 and the correlation between S and $\tilde{S}$ is given by ρ.[9] Latent

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

[9] This sets mean overconfidence to 0 for convenience. One can arbitrarily shift beliefs by changing the mean of ζ without any substantive impact on the main argument. This also equates the variance of beliefs to the variance of ability. Relaxing this constraint and freeing the constraint on σ_ζ^2 would make the correlation between S and $\tilde{S}$ equal

overconfidence is then given by $\tilde{S}_i - S_i$.

People often can and do have insight into their own ability, suggesting we ought to expect $\rho > 0$. Consider a few examples. The Subjective Numeracy Scale (Fagerlin et al., 2007) was developed to let people self-report their own numeracy using a less-burdensome task than a math test; it correlates with numeracy. Objective financial literacy shows correspondence with subjective financial literacy (Lusardi and Mitchell 2017). Objective knowledge and subjective knowledge are correlated in many domains (Alba and Hutchinson 2000; Carlson et al. 2009). People often have at least partial insight into their own abilities. Partial but incomplete insight into one's own ability can be modeled as $0 < \rho < 1$.

If $\rho < 1$, there are differences in latent overconfidence. If $\rho = 1$, then $\tilde{S}_i = S_i$ and there is no latent overconfidence. This latter 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 other constructs correlate 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 cases. But if our approach to assessing overconfidence and its correspondence with other constructs finds evidence in its absence, we must rethink that approach.

Observable Performance and Self-Evaluations

Although ability, S_i, 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:

$$P_i = \lambda S_i + \nu_i \quad (2)$$

where $0 \leq \lambda \leq 1$ and luck, ν_i, is independently drawn from any distribution with mean 0 and variance σ_ν^2[10]. λ represents performance's loading on ability. A perfect measure that fully and only captures the

to $\frac{\rho}{\sqrt{\rho^2+\sigma_\zeta^2}}$, but has no consequential implications for the model. In that case, for $\rho = 1$, ζ_i would represent latent overconfidence. Allowing ζ_i to correlate with S_i would enable latent overconfidence to be associated with ability.

[10] Setting $E[\nu] = 0$ is for convenience and does not otherwise impact the key findings of the model. If there are two versions of the task that each measure the same ability, one of which is hard and one of each is easy, one may take $E[\nu]$ of the hard task to be lower than $E[\nu]$ of the easy task.

focal ability (possibly with 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 we expect $\lambda > 0$ for Raven's matrices (whether or not $\lambda = 1$), whereas we expect $\lambda = 0$ for the phrenologist's head measurements. Alternatively, consider the example introduced above of a 10-item financial literacy quiz containing a subset of items that inadvertently assessed trust in institutions. Participants' scores on the quiz would not only be noisy, they would also be regressive toward participants' trust in institutions, implying $\lambda < 1$.

People's self-evaluations, $\widetilde{P_i}$, are their noisy attempts to evaluate their own performance, P_i. After feedback, performance may be unambiguous. But prior to feedback, people generally have ambiguity regarding how they performed. Following Moore and Healy (2008), such uncertainty should lead to self-evaluations that incorporate prior beliefs through Bayesian-like reasoning (even if people are not proper Bayesian updaters). Indeed, experimental manipulations of prior beliefs affect confidence (Van Marcke et al. 2024). For Moore and Healy, these prior beliefs represented beliefs about the simplicity of the task. In the current model, these prior beliefs represent beliefs about one's own ability. The key extension is thus the variability in prior beliefs across people. The result of this Bayesian-like reasoning is that people ought to evaluate their own performance as a weighted average of their prior beliefs and their performance, plus noise, where the weight on prior beliefs increases with ambiguity:

$$\tilde{P}_i = \alpha\widetilde{S_i} + (1 - \alpha)P_i + \upsilon_i \quad (3)$$

where $0 \leq \alpha \leq 1$ and υ_i is independently drawn from any distribution with mean 0 and variance σ_υ^2.[11] 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 as they are for expectations rather than measures of performance,

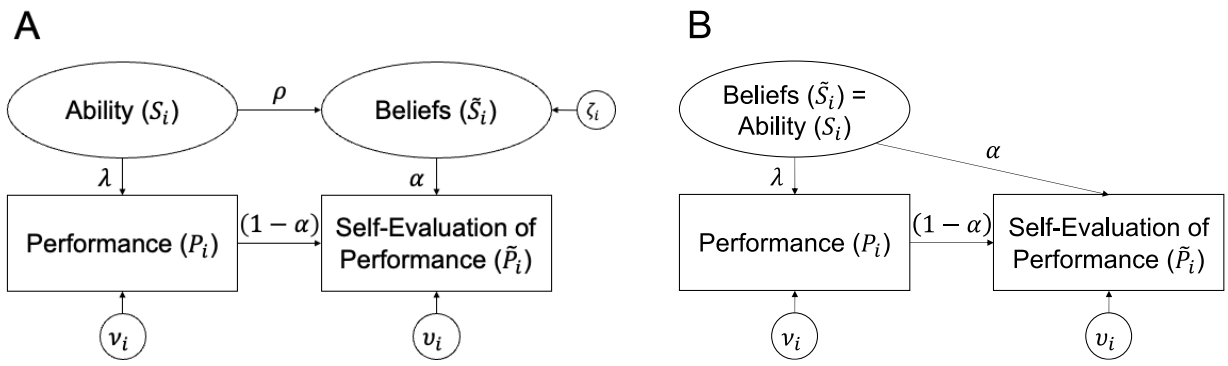
[11] If people knew their true performance, P_i, they could simply report it directly. P_i enters their beliefs but is not used directly because participants receive a noisy signal of their performance. That noise is then folded into υ_i, leaving the signal to enter the equation directly. See Moore and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

$\alpha = 1$, because 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 accurate beliefs by constraining $\rho = 1$. The representation in Panel A distinguishes two reasons why P_i and $\tilde{P}_i$ may not correlate. First, they will not correlate if $\alpha = 1$ and $\rho = 0$. This leads to the typical interpretation: performance (may) provide a noisy signal of ability, but people can assess neither their performance nor their ability. Second, they will also not correlate if $\alpha = 1$ and $\lambda = 0$. Here, they do not correlate because performance does not provide a signal of ability, even though people have insight into their ability. The existence of this second case means that observing that P_i and $\tilde{P}_i$ are uncorrelated is not sufficient to infer that $\rho = 0$.

Figure 2

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



Note. Panel A depicts imperfect but correlated beliefs. Panel B depicts the equivalent model when constraining $\rho = 1$. This implies $S_i = \tilde{S}_i$ given that $E[\zeta] = 0$ and $\sigma_\zeta^2 = 1 - \rho^2 = 0$.

Computing Measures of Overconfidence

In an attempt to isolate the role of overestimation from that of ability, a researcher will take one of three general approaches: (a) regress $\tilde{P}_i$ on P_i and keep the residual, (b) include both self-evaluation $\tilde{P}_i$ and performance P_i in a single multiple regression model predicting the candidate correlate, or (c) take the difference between $\tilde{P}_i$ and P_i. I first address the residual and multiple regression approaches together, as they result in equivalent coefficients, and then address the difference score.

Residual and Multiple Regression Approaches

To calculate overconfidence via residuals, researchers regress self-evaluations on performance:[12]

$$\tilde{P}_i = \gamma P_i + \epsilon_i \quad (4)$$

The residuals, $e_i = \hat{\epsilon}_i$, are kept as measures of overconfidence. Because performance is noisy and self-evaluations use prior beliefs about ability, the expected errors (and thus the residuals) vary with ability:

$$E[\epsilon|S] = \rho\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_v^2}\right)\alpha S \quad (5)$$

The derivation is in the Supplement. For reasonably large sample sizes such that $e_i \cong \epsilon_i$, the residual from regressing self-evaluation on performance is therefore positively confounded with ability if $\rho\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_v^2}\right)\alpha > 0$. That is, there is a confound if three conditions hold. First, true ability and beliefs about one own ability must be positively correlated ($\rho > 0$). Ironically, the confound is maximized if people have perfect self-insight such that there is no latent overconfidence ($\rho = 1$). Second, there must be error in the performance measure that is not attributable to ability ($\sigma_v^2 > 0$, making $\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_v^2}\right) > 0$). The absence of measurement error is the exception, not the rule, so this condition is likely to be met.[13] Third, self-evaluations must be related to beliefs conditional on performance ($\alpha > 0$). Any use of Bayesian-like logic given uncertainty about performance will lead to a direct effect of beliefs on self-evaluations (as will measuring self-evaluations of ability rather than performance), so this condition is likely to be met too.

Multiple regression can be reexpressed as a regression of residuals on residuals. When predicting a candidate correlate, the coefficient on evaluations controlling for performance is the same as the coefficient on residualized evaluations. The multiple regression estimate will typically be more precise.

These measures are intended to assess overconfidence conditional on ability. But instead, the measures are confounded with ability. As a result, each overconfidence measure will correlate with variables that correlate with ability even if those variables are uncorrelated with beliefs conditional on

[12] Throughout, I exclude intercepts. Because my focus is on differences in overconfidence rather than mean levels, intercepts can be accounted for through recentering.

[13] Under this interpretation, all variance in performance is attributable to either ability or measurement error. In the general discussion I consider when σ_v^2 includes systematic error.

ability. (For variables that do correlate with beliefs conditional on ability, estimates will be biased.)

Given the literature on measurement error in predictors (e.g., Birnbaum and Mellers 1979; Cohen et al. 2003; Culpepper and Aguinis 2011; Gillen et al. 2019; Kahneman 1965; Westfall and Yarkoni 2016), why does the current paradigm deserve special consideration? First, because performance is the target of self-evaluations, researchers may mistakenly believe that performance's role as a measure of ability is immaterial. Second, without the extension of Moore and Healy's (2008) model, it may not be transparent that self-evaluations are directly affected by beliefs. Together, these neglected properties grant a false sense of security regarding the impact of measurement error in performance.

Difference Score Approach

To assess overconfidence via a difference score, one subtracts performance from self-evaluation:

$$\Delta_i = \tilde{P}_i - P_i \quad (6)$$

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

$$E[\Delta|S] = (\rho - \lambda)\alpha S \quad (7)$$

The derivation is in the Supplement. 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 ($\alpha > 0$).[14] 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 $\rho = \lambda$.) If self-evaluation only depends on performance and not beliefs, then $\alpha = 0$ and there is no relationship between the difference and ability.

Like the residual measure, the difference score can correlate with other variables that correlate with ability even if those other variables are uncorrelated with beliefs conditional on ability (and for

[14] In this case, measurement error σ_v^2 does not matter. In practice, P_i is often bounded so $\sigma_v^2 > 0$ likely reduces λ.

variables that do correlate with beliefs conditional on ability, estimated correlations will be biased.) Unlike the residual measure, and aligning with typical critiques of difference scores, the difference score measure can also be negatively confounded with ability if $\rho < \lambda$.[15] As a result, for certain parameter configurations, the residual measure of overconfidence will be positively correlated with some outcome, the difference score measure will be negatively correlated with that outcome, and both correlations will be entirely attributable to the measures' confounds with ability.

Comparing the Biases and Additional Variations

Figure 2 Panel B depicts the null model in which participants have perfectly-calibrated beliefs ($\rho = 1$). In this case, the residual bias is given by $\left(1 - \frac{\lambda^2}{\lambda^2 + \sigma_v^2}\right)\alpha S$ and the difference score bias is given by $(1 - \lambda)\alpha S$. For the conditions described above, these act as measures of ability. Under that model, there is no overconfidence. These measures are therefore systematic measures of a construct unrelated to the construct they are intended to measure, as the construct they are intended to measure does not exist.

If there is ambiguity ($\alpha > 0$) and performance loads on ability ($\lambda > 0$), the extent to which the residual score varies with ability will be more-positive than the extent to which the difference score varies with ability if $\sigma_v^2 > \lambda(\rho - \lambda)$.[16] This will be the case if any of three conditions are met: (a) there is enough error in performance; (b) self-insight is low enough; or (c) the loading of performance on ability is not too close to $\frac{\rho}{2}$. On the one hand, this may suggest that researchers are better off using difference scores, given they will generally exhibit less of a positive association with ability. Yet problems with difference scores are described elsewhere (see footnote 15). This particular problem may be smaller for

[15] Prior critiques have addressed difference scores' confound 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 and Furby 1970; Cohen et al. 2003; Edwards and Parry 1993; Griffin et al. 1999; Johns 1981; Wall and Payne 1973; Zuckerman and Knee 1996). Response Surface Analysis via polynomial regression (e.g., Edwards 1994; Barranti et al. 2017; Humberg, Dufner, et al. 2019; Humberg, Nestler, and Back 2019) and Condition-based Regression Analysis (e.g., Humberg et al. 2018a, 2019) aim to establish alternative conditions to assess whether the active ingredient is a discrepancy or positive self-evaluation. The present concern is a confound of self-evaluations with ability that matters whether one is interested in the discrepancy or positive self-evaluation. In addition to other concerns regarding these regression-based approaches (Krueger et al. 2017; Fiedler 2021; cf. Humberg et al. 2018b, 2022), they do not distinguish between performance and ability and so are equally susceptible to the concerns I raise here.

[16] If $\alpha = 0$, neither coefficient is biased; if $\lambda = 0$, they each exhibit the same bias.

difference scores, but others (like the negative bias explored elsewhere) may be larger.

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 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 has the same bias as 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, but evaluations of others' performances 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) that matches that in absolute performance (overestimation).[17]

Simulation Results

Simulations show that these asymptotic results hold in reasonable sample sizes. For all factorial combinations 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, ability was drawn from a standard normal distribution, and error terms were drawn from standard normal distributions scaled by the corresponding variance. Residuals and difference scores were predicted by ability as derived in the model. Average coefficients are depicted in Figure 3.[18] Simulations using log-normal distributions recovered the same averages.

[17] Moore and Healy (2008) show reversals between overestimation and overplacement. S in their model represents a quiz's simplicity, so it takes a common value such that prior beliefs are relevant for both oneself and others. In my model, S_i represents an individual's ability, so prior beliefs are not relevant for others' performance.

[18] The model and simulations assume heterogeneous ability and beliefs but homogeneous parameters. Modeling heterogeneity in parameters is outside the scope of the current investigation. Yet work on metacognitive accuracy suggests there is such heterogeneity (e.g., Grabman & Dodson 2024). Do the current results persist with heterogeneity? Limited investigation by simulating individual-level parameters to be 0 or 1, each with probability 0.5, revealed two key insights. First, if parameter values are uncorrelated, patterns are similar to those for values of

Two Empirical Applications

Common measures of overestimation and overplacement are confounded with ability in reasonable null models in which overconfidence either does not exist or does not matter (i.e., its association with candidate correlates is only 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 yes. Analysis and simulation code is available via ResearchBox: https://researchbox.org/1597.

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 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 the results are interpreted in terms of overconfidence. Using the current model, I find that the results are also compatible with plausible parameter values in which there is no consequential overconfidence (or no overconfidence at all). These analyses indicate the confound can provide a plausible alternative interpretation of published results. Whereas the first application indicates measures of overconfidence can lead to problematic conclusions, the second application indicates evidence from the literature which has been interpreted in terms of overconfidence may not be sufficient to imply a causal or consequent role of overconfidence. Instead, it is consistent with perfectly-calibrated beliefs (i.e., no overconfidence).

Empirical Application I: Overconfidence Predicts Performance

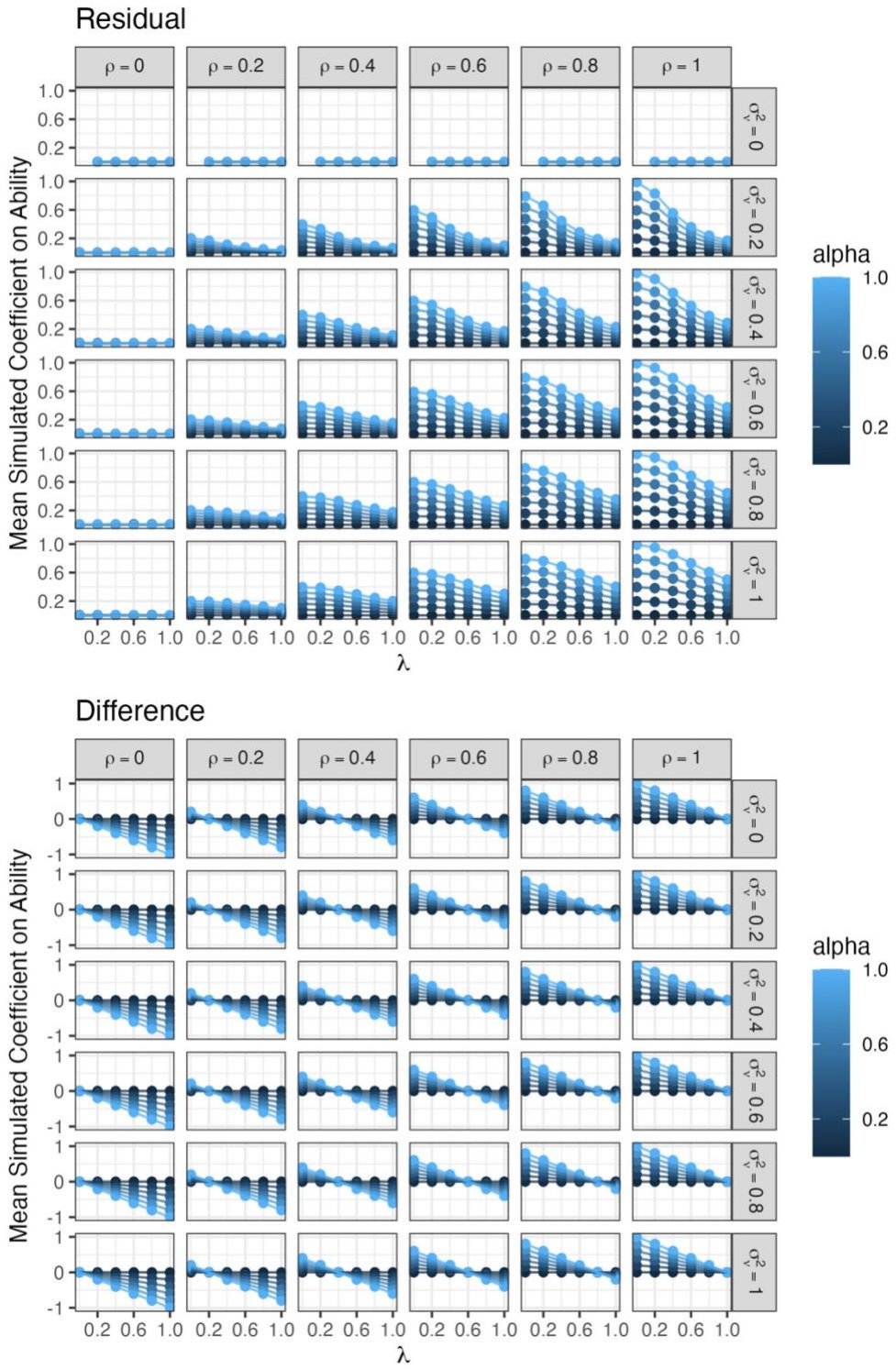
The analysis above indicates overconfidence measures are biased under plausible conditions. Can

---

.4 or .6 (i.e., the expected value). Second, when parameter values are correlated, biases may decrease or become negative under any of three conditions: ambiguity and loading are positively correlated ($\sigma_{\alpha,\lambda} > 0$); ambiguity and self-insight are negatively correlated ($\sigma_{\alpha,\rho} < 0$); or ambiguity and measurement error are negatively correlated ($\sigma_{\alpha,\sigma_v^2} < 0$; only for residual and multiple regression measures). For less-extreme heterogeneity (e.g., uniform rather than bimodal distributions), the change in bias is correspondingly less extreme even in the correlated case.

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 $\sigma_v^2 = \lambda = 0$, estimates are missing. This is because performance does not vary and so cannot be used as a covariate.

this bias affect 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 reasoning, and the degree to which people have self-insight into their own ability. To test this, I first examine a case where there is a measure of performance, people self-evaluate their performance, and 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. Future performance cannot cause past overconfidence, and it is arguably 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 would suggest that past residuals or difference scores are confounded with ability. Note that Moore and Healy do not make the inferential error reported here, as they do not claim to present evidence in favor of correlates of latent overconfidence. Rather, the availability (posted at https://osf.io/6tecy/) and richness of their data present an opportunity to examine whether the error can affect real inferences.

Overconfidence Paradigm

Moore and Healy (2008) collected data from 82 undergraduates. I describe the core components and refer readers to Moore and Healy for more details. Participants completed 18 10-item trivia quizzes: an easy, medium, and hard quiz on each of six topics (science, movies, history, sports, geography, and music). The quizzes were structured 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, participants reported pre-quiz performance expectations immediately preceding each quiz and post-quiz self-evaluations of estimated performance immediately following each quiz. For each of these measures, participants reported full probability distributions of the probability of getting 0 items correct, 1 item correct, etc. Following Moore and Healy, I use the means of the distributions as the self-evaluation measures. Participants also reported analogous assessments of beliefs about another randomly selected participants' performance. The gap between self-evaluations and other-evaluations provided a measure of estimated placement. The gap between actual performance and others' actual performance provided a measure of actual placement. The

estimated and actual measures of performance can be used to construct residual, multiple regression, and difference score measures of overestimation, and the estimated and actual measures of placement can be used to construct corresponding measures of overplacement.

Analysis of these data requires addressing two key complications. First, the quizzes differed in difficulty. If trivia quiz ability exists, performing well on one quiz should predict performing well 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*. That is, I sum across the quiz-level measures within-block to create block-level measures of each metric. A block always contains one easy, one medium, and one hard quiz, substantially reducing the extent to which performance on one block is mechanically 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 ($N = 82$). To exploit the within-subject data, I consider performance on sets of sequential quizzes, clustering errors by subject. For example, to examine ability, I regress block performance on prior block performance such that 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) in *R* (R Core Team 2025).

Key Finding: Overconfidence Predicts Subsequent Performance

Using the first five quiz blocks to provide measures of performance and post-quiz 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) estimated and actual placement measures.

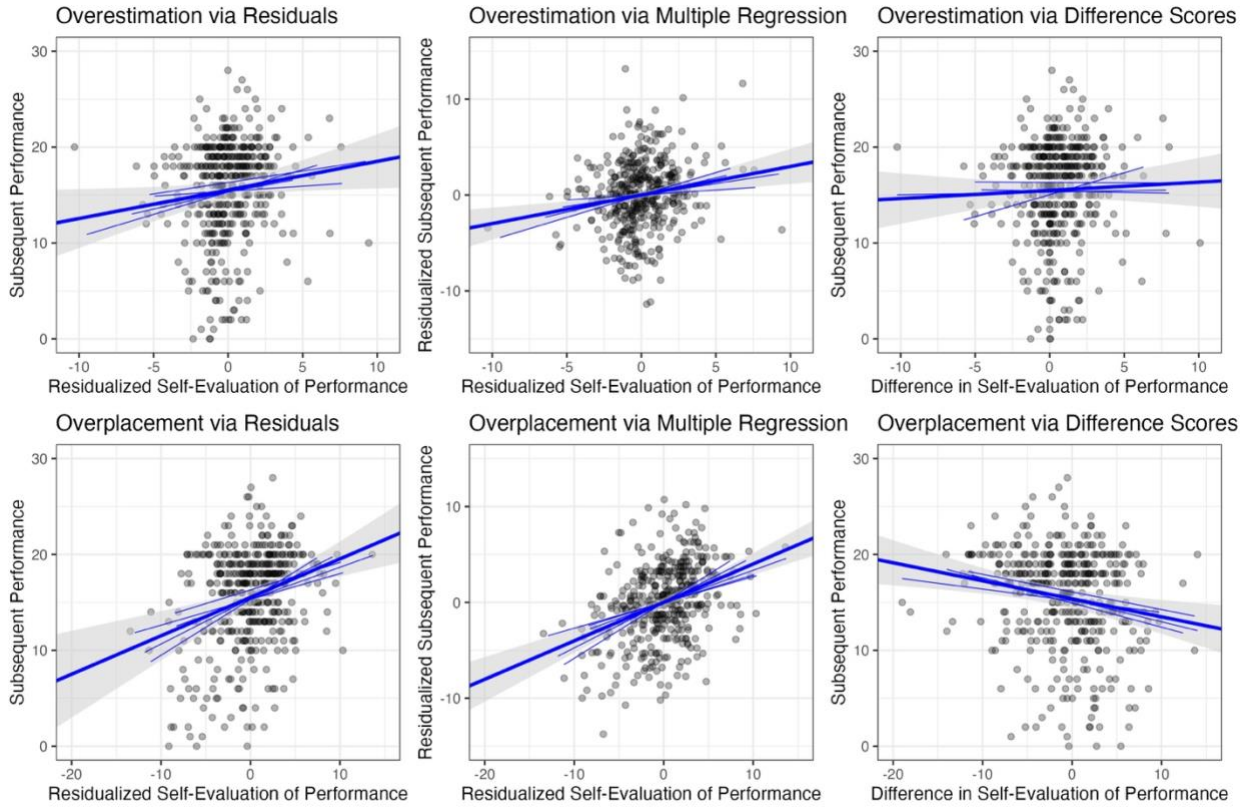
To examine whether overconfidence predicted performance, I conducted six regressions. In each, I regressed subsequent performance on the key predictor(s) with errors clustered by subject. To assess the residualized method, I regressed performance on the previous block's residualized evaluations. To assess the covariate method, I regressed performance on the previous block's evaluations and performance. Note this must give the same regression coefficient on overestimation, potentially with greater precision. To assess the difference score method, I regressed performance on the previous block's difference between evaluations and performance. Overestimation as assessed via residualized self-evaluations predicted performance in the next block ($b = 0.298$, $SE = 0.141$, $t(38) = 2.11$, $p = .042$), as did overestimation as assessed via self-evaluations controlling for performance ($b = 0.298$, $SE = 0.088$, $t(38) = 3.40$, $p = .002$). [19] This latter coefficient is necessarily equal to that on the residual. It is estimated more precisely because performance accounts for substantial variance in the dependent variable in the multiple regression analysis. Overestimation as assessed via the difference score did not significantly predict subsequent performance ($b = 0.085$, $SE = 0.121$, $t(37) = 0.71$, $p = .485$). Overplacement as assessed via residualized relative self-evaluations predicted relative performance in the next block ($b = 0.401$, $SE = 0.103$, $t(42) = 3.89$, $p < .001$), as did overplacement as assessed via relative self-evaluations controlling for relative performance ($b = 0.401$, $SE = 0.056$, $t(42) = 7.13$, $p < .001$). Overplacement as assessed via the difference score also predicted subsequent performance, but the coefficient was significantly negative ($b = -0.188$, $SE = 0.062$, $t(52) = -3.05$, $p = .004$). Results are depicted in Figure 4.

One might argue overconfidence improves subsequent trivia quiz ability (e.g., via self-efficacy). If there were such effects and they were transient, the coefficient would be larger for adjacent blocks. There is no evidence that coefficients from the residual or covariate analyses diminish as there are more intervening blocks. See Table A4 in the Supplement.

[19] All degrees of freedom estimated given cluster-robust standard errors.

Figure 4

Measures of Overconfidence Predict Subsequent Performance in Moore & Healy (2008)



Note. Thick line and confidence band represent pooled model with cluster-robust errors. Thinner lines represent analyses of individual blocks.

This analysis is unable to conclusively rule out an effect of overconfidence on later trivia quiz ability, but such an effect is not necessary to explain the data. Instead, the present model 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. This is consistent with reasonable parameter values for which the positive bias will be greater for the residual measure than the difference score measure.

The results of the residual and multiple regression analyses can be understood by considering four key questions: (1) Are there differences in ability? (2) Do participants have insight into their own ability?

(3) Does trivia quiz performance contain error as a measure of trivia quiz ability? and (4) Does a proxy for ability predict self-evaluations beyond performance? I focus on the overestimation results.

1. 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, $t(31) = 16.59$, $p < .001$), 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 = 0.796, SE = 0.088; lag 5: b = 0.756, SE = 0.094). These results are consistent with the presence of differences in trivia quiz ability ($\sigma_S^2 > 0$), which are noisily measured by each quiz.

2. Do Participants Have Insight Into Their Own Ability? Yes

If participants can predict how they will perform on a quiz, it suggests they have insight into their own trivia quiz ability. I regress expectations before a quiz on subsequent performance, clustering errors by subject. When expectations were measured, neither the difficulty nor the topic of the quiz was known. The coefficient on performance was 0.467 (SE = 0.068, $t(31) = 6.87$, $p < .001$) indicating participants have partial insight into how they will perform.[21] Given the limited information available to participants, this is most readily attributable to self-insight into their own ability ($\rho > 0$).

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

In regressing performance on lagged performance, it comes as no surprise that $R^2 = 0.538 < 1$,

[20] Trivia quiz ability includes skill, knowledge, and other necessary inputs that remain stable during the study.

[21] Participants could have anticipated the difficulty of the third quiz in each block, which could inflate this relationship. If the first quiz was hard and the second was medium, participants could anticipate the third would be easy. The main result holds using only the completely-randomized first quiz from each block, adjusted for difficulty (b = 0.233, SE = 0.048, $t(31) = 4.80$, $p < .001$, 95% CI: [0.134, 0.332]). The coefficient was no stronger when using only the third quiz (b = 0.177, SE = 0.041, $t(45) = 4.33$, $p < .001$). Note there are more between-quiz shocks to actual performance (e.g., quiz topic) that may average out within a block than there are between-quiz shocks to expected performance that may average out within a block (e.g., quiz topic is not known when generating an expectation). This results in greater relative variance in performance vs. expectations at the quiz-level than the block level, meaning these coefficient magnitudes (0.233, 0.177) are not directly comparable to that in the main text (0.467).

indicating that the measures are not errorless indicators of the same construct. Performance as a measure of ability contains error ($\sigma_v^2 > 0$).

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

The final key question is whether self-evaluations of performance regress toward ability. Ability is not directly observable, but subsequent performance provides a noisy proxy. I therefore regress retrospective self-evaluations on subsequent performance as a proxy for ability, controlling for concurrent performance. The coefficient on concurrent performance was 0.880 (SE = 0.035, $t(49) = 24.94$, $p < .001$). This suggests participants likely have some insight into how well they did on each block (although this coefficient also partially captures the role of ability). The partial coefficient on the proxy for ability, subsequent performance, was 0.099 (SE = 0.029, $t(49) = 3.43$, $p = .001$).[22] The size of this coefficient did not attenuate with more intervening blocks (1 block: $b = 0.119$, $SE = 0.034$; 2 blocks: $b = 0.101$, $SE = 0.052$; 3 blocks: $b = 0.141$, $SE = 0.044$; 4 blocks: $b = 0.099$, $SE = 0.107$). This suggests post-quiz self-evaluations assess not only performance, but also idiosyncratic ability directly ($\alpha > 0$).

Subsequent Performance Illustrates the Problem Regarding Other Correlates of Overconfidence

Overconfidence, as measured via residuals or multiple regression, predicts future performance. The proposed confound is a sufficient explanation if four conditions hold: (1) people differ in ability, (2) people have some self-insight, (3) performance measures ability with error, and (4) self-evaluations regress toward beliefs about ability. These conditions are reasonable: it would be surprising if ability did not differ, if people had no self-insight, if performance were a perfect measure of ability, or if people avoided using beliefs about their own ability. Many reported associations between overconfidence and other variables use approaches equivalent to those above, but do not sufficiently rule out ability even while claiming to do so. (Given that difference scores were not positively related to future performance, the prospect of a positive confound for difference scores may be less concerning in this case.)

[22] This and the Key Finding represent the same shared variance between block *t+1* performance and block *t* self-evaluations after partialling out block *t* performance. To ensure non-redundant tests, one could simply use block *t+2* performance instead. But substantively, this reflects the core insight that overconfidence predicts later performance precisely because later performance is a proxy for ability and overconfidence is confounded with ability.

Empirical Application II: Correlating Overplacement With Status and Social Class

The first application found that measures of overconfidence predict subsequent performance. This can be readily explained by the confound between measures of overconfidence and ability. This indicates a problematic conclusion one might draw from the data, but it does not require us to reinterpret prior findings. Perhaps the 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 constructs. The goal is to consider whether this model could account for the results, not 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 are correct, it is unlikely they 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. The qualitative pattern is what matters.

IIA: Overplacement, Perceived Competence, and Status

In six studies, Anderson et al. (2012) propose that holding ability constant (as operationalized via performance), greater confidence generates higher status in the eyes of others. Study 1 was correlational and well-suited to analysis using the current framework.[23] Could ability alone be responsible for the correlation with status in Study 1?

Participants took a geography quiz by placing 15 U.S. cities on a blank map of North America. Accuracy was defined by placing a city within 150 miles of its true location. Participants reported (a) a self-evaluation of their quiz performance percentile and (b) a self-evaluation of their general geography

[23] I do not attempt to account for the results of the remaining studies, and so this should not be interpreted as a counter-explanation for the paper's complete set of results. The paper is sometimes cited for its use of residual scores to assess overconfidence (cf., Belmi et al. 2020; Cheng et al. 2021; Lyons et al. 2021; Murphy et al. 2015).

knowledge percentile. Next, they repeated the quiz with a partner. The partner rated the participant's perceived competence using two analogous percentile measures as well as the participant's status using a 4-item Likert scale. 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.

Could the results be explainable through links between *ability* and perceived competence and status instead? I consider sets of $\rho, \lambda, \alpha, \beta_{PC}$ (i.e., the causal impact of ability on perceived competence) and β_S (i.e., the causal impact of ability on rated status). For simplicity, I constrained error variances such that each variable had a variance of 1. It is possible to recover nearly-identical correlations to those observed in the data with $\rho = 1$, $\lambda = .7$, $\alpha = .8$, $\beta_{PC} = .55$ and $\beta_S = .45$. Given the number of parameters, multiple configurations fit similarly well (e.g., $\rho = .75, \lambda = .55, \alpha = .75,\ \beta_{PC} = .75, \beta_S = .6$).[24] The observed correlations and correlations implied by these parameters are given in Table 1.

Table 1

Empirical Application IIA: Observed and Modeled Correlations

Measure	Perceived Competence	Status
Actual Performance	.39	.33
Residual	.36	.26
$\rho = 1, \lambda = .7, \alpha = .8$	$\beta_{PC} = .55$	$\beta_S = .45$
Actual Performance	.39	.32
Residual	.35	.28
$\rho = .75, \lambda = .55, \alpha = .75$	$\beta_{PC} = .75$	$\beta_S = .6$
Actual Performance	.41	.33
Residual	.35	.28

The present model has enough free parameters to readily fit the data so it is not a stringent test. But if one accepts the core logic of the current argument (performance is an imperfect measure of ability,

[24] 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 recovers .56 ($\lambda\rho\alpha + (1 - \alpha) = .55 * .75 * .75 + .25$).

people have partial insight into their own ability, and people are partially sensitive to their priors), then those parameters should be accommodated. If one does so, the original analysis would require additional free parameters to represent paths from beliefs to outcomes.

IIB: Overplacement and Social Class

Whereas Anderson et al. (2012) propose overconfidence causes status, Belmi et al. (2020) propose social class causes overconfidence due to the pursuit of status. (Each proposal implies a set of correlations, no matter the direction of the causal arrow, and in both cases the target studies assessed correlations.) Could social class result in greater test-taking ability instead (i.e., mere ability to perform on tests, perhaps due to tests biased by social factors or differential experience with test-preparation)? This example provides a somewhat more-stringent test of the model's explanatory ability given a larger set of correlations to be fit, including both residual measures and difference score measures of overconfidence.

Across four studies including three different tasks, the paper uses four measures of social class (self-report, income, education, and parental education), and multiple ways of measuring overplacement on each task. 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 paper also uses the difference score measure: the difference between self-evaluated and actual percentiles.[25]

Across studies, three different tasks were used to assess overplacement. In Study 1, participants assessed whether two successive images matched; after a few practice trials, the task consisted of 20 trials. In Study 2, the task was a test intended to assess general cognitive ability (20 timed items drawn from the Wonderlic Personnel Test). In Studies 3 and 4, the tasks were 15-item trivia quizzes. In every study, self-evaluated placement was assessed as how well participants thought they did on that task relative to others after having completed the task. In Study 2, this was averaged with an item on general mental abilities relative to others. Successful performance relies on a combination of inputs, including

[25] The paper also uses Edwards' (1995) proposed approach of calculating Wilks' lambda to compare coefficients across two regressions: one predicting self-evaluated percentile from social class and the other predicting actual percentile from social class. While I do not model that analysis here, I do examine whether the model parameters successfully capture the correlations between social class and both self-evaluated and actual percentiles.

mere test-taking ability which may be biased by social factors (e.g., how the test is constructed, differential access to test-preparation resources, or experience with such assessments in the past).

The paper reports meta-analytic estimates in its Tables 14 and 15. Each measure of social class was more-correlated with perceived performance than actual performance, each measure of social class correlated with residualized self-evaluations, and self-reported social class (but not income, education, nor parental education) correlated with the difference between perceived and actual percentiles. I examine whether plausible parameter values are consistent with the reported estimates (without covariates). These correlations, along with their 95% confidence intervals, are reproduced in the top panel of Table 2.[26]

Table 2

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 variance of 1. The correlations resulting from two such sets of parameters are reported in the lower two panels of Table 2. The middle

[26] There are substantive differences across studies, and a single set of parameters may not be able to account for every study's results. 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 explicit. λ, α, and ρ need not be consistent across studies.

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 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.

There are four key takeaways. First, the 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). 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 paper's theory, which would add 1 to 4 more parameters (i.e., to relate latent overconfidence to each measure of social class) if one accepts the model's core logic. Each correlation is within the 95% confidence interval, and more than two-thirds are within the implied 67% confidence interval (half the 95% confidence interval). This is despite both the residual score (for every measure of social class) and the difference score (for the self-report measure of social class) implicating a relationship between overplacement and social class.

Second, this neither rules out the authors' hypothesis nor indicates an impact of social class on test-taking ability. It merely presents an alternative interpretation of the data in which social class may not be related to overplacement. Importantly, even under this interpretation, ability could refer to mere test-taking ability (e.g., access to test-preparation resources or prior experience with similar tests).

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 5 presents a set of compatible values of ρ, λ, α, and β for self-reported social class. Similar values may be constructed for income, education, and parental education.

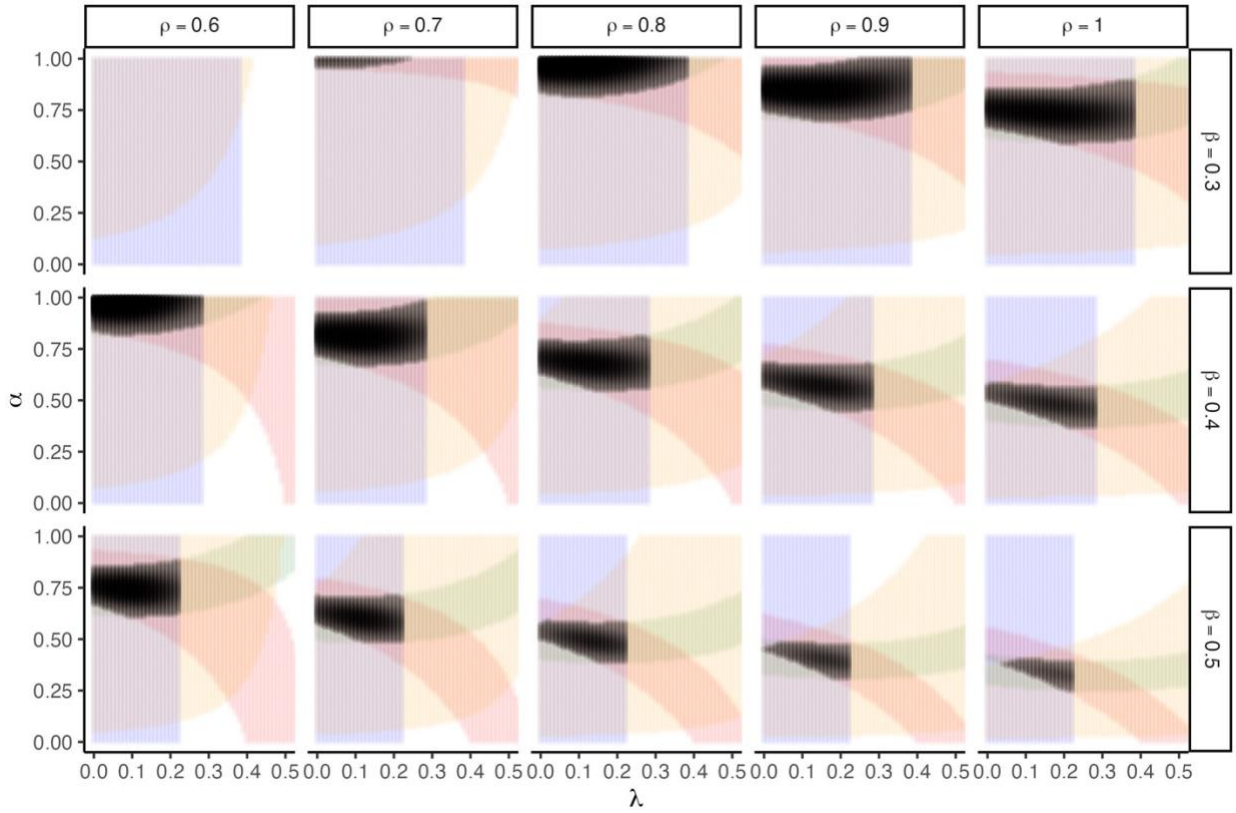
Finally, one should evaluate whether these parameter values are plausible. For low values of β, these data require high ρ, low λ, and high α. Whether these are themselves plausible depends on outside

knowledge and cannot be resolved using these data alone. The paper also proposes a rich network of associations with desire for status and perceived competence, which may address the model's plausibility.

Table A3 in the Supplement provides analogous parameter configurations consistent with findings for 16 different articles. Some of the parameter configurations are plausible, but others are not.

Figure 5

Empirical Application IIB: Compatible Parameter Values for Self-Reported Social Class



Note. Dark shaded regions indicate regions where each of the four target correlations implied by the set of parameters falls within the 95% confidence interval of the correlation for self-reported social class. Cases where each value is closer to the center of the confidence interval are shaded darkest. Lighter colors indicate regions where only a subset of the target correlations with self-reported social class fall within the 95% confidence interval (blue: performance; red: self-evaluation; green: residual; orange: difference.)

General Discussion

Both in theory and in practice, widely-used measures assessing differences in overconfidence are confounded with ability. This applies to both overestimation and overplacement. While the biases in the residuals and difference scores are driven by different properties (the first by measurement error, the

second by imperfect construct validity), both are attributable to the fallacy of equating performance with ability. Under many reasonable parameter specifications, the positive bias is likely to be greater for residuals, whereas the difference score also has the potential for negative biases. Below I discuss additional nuances and caveats of the model and end with some recommendations.

Model Nuances and Caveats

Model Specifications

First, I model the measure of performance as a function of ability (S_i, weighted by λ) and luck (v_i). I have discussed luck as though it is mere measurement error. But one might decompose luck into temporally inconsistent random luck (e.g., momentary distraction) and task-specific stable luck (e.g., idiosyncratic advantages on non-relevant features). Once luck is decomposed in this way, it is apparent that addressing measurement error is unlikely to completely address the problem for the residual. Reducing measurement error to 0 may not reduce σ_v^2 to 0.

Second, if luck is only measurement error, lack of *reliability* in the performance measure leads to lack of *validity* in residual measure. Increasing reliability of the performance measure could increase validity of the residual measure. I address this in the second recommendation.

Finally, this model represents but one functional form which is unlikely to be perfectly specified. For example, it posits a simple linear relationship between ability and beliefs. Yet people who are less-skilled may be less-calibrated regarding their own skill (Kruger & Dunning 1999). Despite statistical critiques (Krueger & Mueller 2002; see Supplement), there is evidence for the unskilled-and-unaware effect that addresses those critiques (Feld et al. 2017; Jansen et al. 2021). Such an effect will affect the magnitude and form of the confound, but it will not alleviate the problem. This reinforces that the model here represents one 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. There remains debate about the relative merits of the two (see e.g., Belmi et al. 2020; Lyons et al. 2021; Parker & Stone 2014 for recent discussions of the two

approaches in the context of substantive questions). The current model reinforces that neither can be considered right and the other wrong; their relative biases depend on the relative contributions of ability and luck to performance. Under many reasonable sets of parameters, the bias is more-positive for residuals than difference scores. Based on the empirical applications, one might be more optimistic about difference scores than residuals. Yet the relative biases depend on the parameter values, such that difference scores may have a negative bias and the observed patterns may not necessarily generalize.

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 to address the concerns regarding difference scores and residuals (e.g., Edwards 1994; Humberg et al. 2018a, 2019). In their base form, they do not account for measurement error or construct mismatch. Only when combined with a strategy to address measurement error will they address reliability for the residual score measure. Even then, construct validity remains a concern for both the residual score and the difference score measures.

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

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 people regress to must align with the construct performance assesses. A mismatch, as in the case of a financial literacy scale with items that measure trust instead, is equivalent to construct invalidity ($\lambda < 1$). This leads to the core problem for difference scores. A potential critique is this simply represents a different form of improper confidence: people confidently rely upon a prior that should not apply and therefore regress to the wrong belief. I argue we cannot be so quick to attribute such a problem to the participant's updating strategy (and therefore label it a variant of overconfidence) rather than the researcher's inferential strategy.

Consider again the phrenologist introduced earlier. Both the phrenologist and the participant may believe 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 no idea 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 counter: "That is overconfidence! The participant is inappropriately regressing their self-evaluation of performance on a nondiagnostic test to their beliefs about their own intelligence." But in such a case, it would be inappropriate to fault the participant for relying on the very construct the researcher purports to measure. Thankfully, most researchers are not phrenologists and use instruments with greater validity. But this provides only cold comfort.

Alternatively, the participant may recognize the invalidity of the test. They again may have substantial ambiguity about their test score and rely on their priors instead. In this case, their priors would not represent their beliefs about their intelligence but about a relevant target, e.g., their hat size. Thus, the relevant target ability is that viewed through the eyes of the participant. In many cases this may be straightforward, but there may be meaningful disconnects in edge cases like that above.

This discussion raises a thorny question regarding whether the effects of using misleading labels for a performance task can be considered overconfidence. Overconfidence should not depend on what the researcher believes a task measures. Above I argued we should not accept the overconfidence label when the participant believes the measure is measuring the same construct the researcher earnestly believes it measures. Given that, we ought to be cautious in accepting the overconfidence label when the participant believes the measure is measuring the construct the researcher merely tells them it measures (Ehrlinger & Dunning, 2003).

Remaining Ever-Vigilant

Some methodological problems must be addressed once and then are considered resolved. Others require vigilance in each usage. Measures of overconfidence fall into the latter case. One of the larger literatures excluded from Table A1 in the Supplement is that on CEO overconfidence (e.g., Malmendier and Tate 2005, 2008). In this literature, CEOs' holding of options is used as a measure of CEO overconfidence, which is then associated with numerous characteristics. The early papers that developed the measures took great care to rule out the possibility that such measures represent use of private

information. In the context of the present model, they recognized the potential for ability to act as a confound and through triangulation persuasively argued that there are several indications that the proxy for self-evaluations is not problematically confounded with latent ability. This is not guaranteed to always hold. When there is a reason to expect a measure of overconfidence may be confounded with ability, observing that it does not suffer from the confound in one sample does not guarantee it will not suffer from the confound in another.

Beyond Overconfidence

These concerns regarding residual and difference scores represent a broader concern. They apply anytime there are two noisy measures of correlated constructs 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). In a meta-analysis of the relationship between self-enhancement and narcissism, Grijalva and Zhang (2016) find that the reported relationship is stronger when researchers use residual scores than difference scores. This finding is consistent with a larger bias for residual scores than difference scores across a wide range of parameter values. It may also speak to other varied applications with structural similarities. This model addresses overconfidence, but 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 present 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 is a call to use reliable, valid measures. No one touts the use of an unreliable or invalid measure, but given a strong rationale to believe in the potential for a confound, this reinforces the importance of good measures. This is particularly important given that performance measures are often reasonably brief and moderately reliable at best. Anderson et al. (2012) 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. There is little reason to believe these values are uniquely low.

All else equal, assessments using larger random samples of items will tend to have better coverage and reliability than smaller samples. Consider the infinite population of items that properly capture the target construct. A measure constructed using a larger random sample of items from that population will have both better construct coverage (as there will be fewer untapped facets) and less measurement error than a measure constructed using a smaller sample of items. But properly specifying the population of items from which to sample is difficult. While it may be a valuable theoretical principle, simply relying on a larger sample of items may not always be feasible in practice.

2. Account for Measurement Error

To provide an unbiased test when controlling for performance (i.e., when using the residual or multiple regression measure), it is useful to recall the conditions under which there is no bias. The bias is eliminated if any of three conditions hold: (a) $\alpha = 0$, meaning there is no ambiguity and participants do not regress their self-evaluations towards their prior beliefs, (b) $\rho = 0$, meaning latent beliefs are unrelated to latent ability, or (c) $\frac{\lambda^2}{\lambda^2+\sigma_v^2} = 1$, meaning there is no error in performance and performance is at least partially related to ability.

This latter concern is analogous to 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 (e.g., Culpepper and Aguinis 2011). These approaches help to the extent that σ_v^2 only represents measurement error and not other constructs (as in the case of stable luck described earlier). The Supplement presents additional details regarding both approaches. In a third empirical application presented in the Supplement (Parker et al.'s 2012 study of inappropriate confidence predicting a three-item measure of self-reported retirement planning), I reanalyze the original data, first using a structural equation model and then using an adjustment for errors-in-variables. In this application,

the standard analysis finds an apparent correlation between self-reported retirement planning and overconfidence. Accounting for measurement error largely attenuates or 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 the null may hold. Given the ability to characterize the magnitude of the bias, one can qualitatively specify parameter configurations that could or could not account for the results. In some cases, no parameter values may be able to account for the set of correlations without overconfidence. In others, there may be feasible but implausible parameter configurations: while they are mathematically plausible, they may be ruled out based on theory.

Consider Figure 5 in application IIB, depicting the parameter values that are compatible with the null model. As can be seen by the crosscutting lightly shaded areas, while many parameters may be compatible with any single correlation, a much smaller set is compatible with the full set of correlations. In this case, a set of parameter values remains. In some cases, the remaining values may be implausible. One implication is that considering both residuals and difference scores may more-effectively limit the set of plausible parameters. In Figure 5, fewer parameters are compatible with both residuals and difference scores than either one alone.

In other cases, one can rule out the alternative explanation altogether. First, if there 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 candidate correlate 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.

Second, if the relationship between ability and the candidate correlate and the relationship between residualized overconfidence and the candidate correlate have opposite signs, the bias described here could not account for the results. For example, as shown in Table A3 in the supplement, while the results of Lyons et al. (2021) regarding overconfidence and false news could be accounted for with the reported parameters, it would require an unlikely (perhaps implausible) sign on β, suggesting it is

unlikely to account for the results. The bias may still be consequential: it may imply the magnitude of the relationship between overconfidence and the outcome measure of interest is underestimated. This is not guaranteed to hold for difference scores due to the potential negative confound shown in Figure 3.

To examine this, one might consider whether overconfidence on one set of items predicts performance on a separate set of items, as in application I. 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; cf. Burson et al. 2006). But several notes are in order. First, the items used to calculate performance must be separate from those used to calculate overconfidence. Second, this analysis can provide evidence of a problem, but cannot provide evidence of absence of a problem. Third, one must avoid inadvertently accepting the null hypothesis, particularly given uncertainty around the estimates. In establishing such bounds, it is important to consider the estimate's uncertainty, not merely the point estimate itself. Finally, these bounds are with respect to this particular 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) might not be so readily ruled out.

4. Use Alternative Measurement Approaches

Finally, one may opt to use a different measurement approach altogether. Multiple measures have been introduced which may be less susceptible to the focal confound. Direct measures of overclaiming (i.e., claiming one recognizes people, objects, or events that do not exist; Paulhus et al. 2003) may reduce 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. If there is any ambiguity about the answer to a particular item, people likely incorporate their prior beliefs. Relying on priors in the presence of ambiguity could lead high-ability people (with defensibly higher-priors) to be more likely to overclaim than low-ability people on items with sufficient ambiguity.

Similarly, Binnendyk et al. (2025), Binnendyk and Pennycook (2024) and Lawson et al. (2024) each propose measures of individual differences in overconfidence. These are based on estimated

performance on a task for which performance is at or near chance and difficult to ascertain (Binnendyk et al. 2025; Binnendyk and Pennycook 2024) or expectations of performance on specific tasks for which one has little contextual basis to expect superior performance (Lawson et al. 2024). One interpretation of these measures in terms of the current model is that ability at these specific tasks ought to be unrelated to other correlates of interest. As with the Paulhus et al. (2003) measure, there is reason to be optimistic regarding these tasks relative to the other methods described here. But they may not completely address the problems laid out here. If some people accurately believe themselves to be generally more successful than others at a wide variety of tasks, the same problem could persist. To interpret ambiguous evidence, people may rely on a prior regarding ability that is calibrated relative to other people but coarse and relatively undifferentiated across domains. In such a case, the tasks could operate as poor measures of the underlying construct that participants rely on (see "Is $\lambda < 1$ Just a Form of Overconfidence?"). Although the proposed confound could hold in principle, it would provide a less-satisfying explanation for these measures, given the additional tenuous conditions that must be met to support it. These approaches suggest potentially promising directions to minimize the plausibility of the proposed confound.

Summary and Conclusion

Measures of overconfidence on particular tasks vary 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. Because performance measures are imperfect, 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 when accounting for performance. The result is that both residual and difference overconfidence measures are confounded with ability. In an idealized null model, this bias can be quantified.

These confounds imply that it is possible to observe surprising results. In one reanalysis, 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 results could be driven by test-taking ability alone. If researchers are willing to make strong assumptions regarding construct validity and estimate or assume the 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: complications may arise regarding construct validity and unstable estimates. Instead, design-based solutions (e.g., experimental manipulations or other measurement approaches) or accepting alternative interpretations of the results (i.e., plausible parameter configurations) may ultimately prove necessary. This work serves as a call to continue to improve our collective attempts to measure differences in overconfidence (whether stable or transient) and their true associations with traits, decisions, and behaviors.

References

Alba JW, Hutchinson JW (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 DT (2017) Precautionary savings, retirement planning and misperceptions of financial literacy. *Journal of Financial Economics*, *126*(2), 383-398.

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

Ames DR, Kammrath LK (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*.

Barranti M, Carlson EN, 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 MA, 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.

Binnendyk J, Li S, Costello T, Hale R, Moore DA, Pennycook G (2025) Is overconfidence a trait? An adversarial collaboration. *Psychological Science*, *36*(12), 941-951.

Binnendyk J, Pennycook G (2024) Individual differences in overconfidence: A new measurement approach. *Judgment and Decision Making*, *19*(e28), 1-24. https://doi.org/10.1017/jdm.2024.22

Birnbaum MH, Mellers BA (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 KA, Larrick RP, 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 WK, Goodie AS, Foster JD (2004) Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making*, *17*(4), 297-311.

Carlson JP, Vincent LH, Hardesty DM, Bearden WO (2009) Objective and subjective knowledge relationships: A quantitative analysis of consumer research findings. *Journal of Consumer Research*, *35*(5), 864-876.

Cheng JT, Anderson C, Tenney ER, Brion S, Moore DA, Logg JM (2021) The social transmission of overconfidence. *Journal of Experimental Psychology: General*, *150*(1), 157.

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

Cohen J, Cohen P, West SG, Aiken LS (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

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

Cronbach LJ, Furby L (1970) How we should measure "change": Or should we?. *Psychological Bulletin*,

74(1), 68.

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

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

Edwards JR (1995) Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes*, *64*(3), 307.

Edwards JR, Parry ME (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 TS, Budescu DV (1994) Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519.

Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM (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 SM, Lau HC (2014) How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(443).

Fuller WA (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.

Grabman, JH, Dodson CS (2024) Unskilled, underperforming, or unaware? Testing three accounts of individual differences in metacognitive monitoring. *Cognition*, *242*, 105659.

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

Grijalva E, Zhang L (2016) Narcissism and self-insight: A review and meta-analysis of narcissists' self-enhancement tendencies. *Personality and Social Psychology Bulletin*, *42*(1), 3-24.

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

Healy PJ, Moore DA (2007) Bayesian overconfidence. *Available at SSRN 1001820*.

Humberg S, Dufner M, Schönbrodt FD, Geukes K, Hutteman R, Küfner AC, van Zalk MH, Denissen JJA, Nestler S, Back MD (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 FD, Geukes K, Hutteman R, van Zalk MH, Denissen JJA, Nestler S, Back MD (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 FD, Geukes K, Hutteman R, van Zalk MH, Denissen JJA, Nestler S, Back MD (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 FD, Geukes K, Hutteman R, van Zalk MH, ..., Back MD (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 MD (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 RA, Rafferty AN, Griffiths TL (2021) A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756-763.

John OP, Robins RW (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).

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.

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 JB, 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.

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

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

Krueger JI, Heck PR, Asendorpf JB (2017) Self-enhancement: Conceptualization and assessment. *Collabra: Psychology*, *3*(1), 28.

Krueger J, Mueller RA (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 MA, Larrick RP, Soll JB (2024) The Overconfidence Test: A parsimonious measure of trait-level overconfidence. http://dx.doi.org/10.2139/ssrn.4893481

Lawson MA, Larrick RP, Soll JB (2025) Individual differences in overconfidence and their psychological bases. http://dx.doi.org/10.2139/ssrn.4558486

Li S, Hale R, Moore DA (2025) Is overconfidence an individual difference? *Judgment and Decision Making, 20*(e25), 1-30. https://doi.org/10.1017/jdm.2025.11

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

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

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

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

Lyons BA, Montgomery JM, Guess AM, 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 PD, Noels KA, 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.

Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psychological Review*, *115*(2).

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

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

Moore D, Swift SA (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 SC, von Hippel W, Dubbs SL, Angilletta Jr MJ, Wilson RS, Trivers R, Barlow FK (2015) The role of overconfidence in romantic desirability and competition. *Personality and Social Psychology Bulletin*, *41*(8), 1036-1052.

Parker AM, 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).

Parker AM, Stone ER (2014) Identifying the effects of unjustified confidence versus overconfidence: Lessons learned from two analytic methods. *Journal of Behavioral Decision Making*, *27*(2), 134.

Paulhus DL, Harms PD, Bruce MN, Lysy DC (2003) The over-claiming technique: measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890.

R Core Team (2025) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rahnev D (2025) A comprehensive assessment of current methods for measuring metacognition. *Nature Communications*, *16*(701).

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

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

Taylor SE, Brown JD (1988) Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193-210.

Thomson GH (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.

Van Marcke H, Denmat PL, Verguts T, Desender K (2024) Manipulating prior beliefs causally induces under-and overconfidence. *Psychological Science*, *35*(4), 358-375.

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

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

Zuckerman M, Knee CR (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.