Widely-Used Measures of Overconfidence Are Confounded With Ability

Stephen A. Spiller

UCLA Anderson School of Management

June 20, 2023

Author Note

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

Data and materials from prior investigations are available at https://osf.io/6tecy/ and https://alpdata.rand.org/, respectively. All code is available at https://researchbox.org/1597&PEER_REVIEW_passcode=ORRDVP.

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

Abstract

The overconfidence concept is one of the great success stories of psychological research, influencing discourse in the popular press, business, and public policy. Relative to underconfidence, overconfidence at various tasks is purportedly associated with greater narcissism, lower anxiety regarding those tasks, higher status, greater savings, more planning, and numerous other differences. Yet much of this evidence may merely indicate that there are associations with ability rather than overconfidence. This results from two underappreciated properties of the typical measures of overconfidence. First, performance is an imperfect measure of ability; analyses that account for performance do not sufficiently account for ability. Second, self-evaluations of performance should reflect ability in addition to performance; because performance is ambiguous, people should use their prior beliefs about their own ability. I show these uncontroversial principles imply that commonly-used measures of overconfidence are confounded with ability. I support these analytical results by reanalyzing two previously-published datasets. In the first, overconfidence predicts subsequent performance, consistent with overconfidence as a signal of ability but inconsistent with overconfidence as a bias. In the second, the association between overconfidence and financial planning can be explained by modeling financial knowledge as a common cause of both. Statistical techniques that explicitly address the role of measurement error provide a potential solution if researchers are willing to accept strong untestable assumptions. This research serves as a stark reminder: when researchers propose that differences in overconfidence are associated with other behaviors, beliefs, or evaluations, they must rule out differences in ability as an alternative explanation.

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

Widely-Used Measures of Overconfidence Are Confounded With Ability

By any measure, Serena Williams is one of the greatest athletes of all time. Of 1,014 career tennis matches, she won 858. It would be an understatement to say that she is highly skilled. And, like any other GOAT,[1] she is entitled to acknowledge it. Does this alone make her overconfident? Of course not. As pitcher Dizzy Dean allegedly said: "It ain't braggin' if you really done it."

Now, suppose that after taking a sample of 20 of her matches, an entrepreneurial student manages to snag a few minutes of her time and asks her to report how many of those 20 matches she won. After an odd look, she might respond 17. She would be well-calibrated. The sample might include more than 17 wins or might include fewer, but in expectation it would be very close to 17. Through an incredible effort, this student approaches 99 more retired tennis players, takes samples of 20 of each of their matches, and asks them each to report how many they won. Each and every one of them reports a number consistent with their career record. Armed with this impressive dataset, the student sits down and dutifully calculates each player's overconfidence using established techniques from the literature. This leads him to confidently, but wrongly, claim that Serena Williams is overconfident, despite her perfectly calibrated response. With some artistic license, this is the current state of assessing individual differences in overconfidence. It is a problem.

Overconfidence is widely acknowledged as a ubiquitous bias. It is reliably reproduced in academic research, worthy of chapters in popular business books, and labeled as "the most significant of the cognitive biases" by a founder of the heuristics-and-biases research program (Kahneman, 2011). Casual observation seems to confirm that overconfidence both exists and

[1] Greatest Of All Time

varies across people. Amateur stock traders expect to beat the market, aspiring signers belt out-of-tune solos in auditions for American Idol, and would-be-daredevils confidently instruct their neighbors "hold my beer" as they attempt ill-advised stunts. In other words: There are overconfident people. Look around.[2]

As a result of its apparent importance, pervasiveness, and variability, individual differences in overconfidence have been widely studied. Different flavors of overconfidence have been associated with a wide array of correlates. These include narcissism (Ames & Kamrath, 2004; Campbell et al., 2004; John & Robins, 1994), savings (Avdeenko et al., 2019), advice-seeking (Kramer, 2016), financial planning (Parker et al., 2012), reduced language anxiety (MacIntyre et al., 1997), social status (Anderson et al., 2012), choice of nonlinear incentives (Larkin & Leider, 2012), susceptibility to false news (Lyons et al., 2021), search behavior (Moorman et al., 2004), and more. (See Moore & Schatz, 2017, for a recent review of overconfidence, and Alba & Hutchinson, 2000, and Carlson et al., 2009, for reviews of the correspondence between objective and subjective knowledge.) Such research uses a variety of related terms, including overconfidence, biased self-evaluations or self-assessments, unjustified confidence, inappropriate confidence, subjective knowledge when controlling for objective knowledge, and others. I use overconfidence to refer to the unjustified belief that performance has been or will be better than it actually was or will be, according to some objective standard.[3]

Unfortunately, widely-used methods that are used to assess individual differences in overconfidence confound biased beliefs with actual ability. Although researchers intend to control for latent ability, they instead and insufficiently control for realized performance. The

[2] Cf., Summers, "Finance and Idiots," as cited in Fox (2009).
[3] This includes both *overestimation*, self-evaluations that overstate absolute performance, and *overplacement*, self-evaluations that overstate relative performance (Moore & Healy, 2008).

resulting associations between overconfidence and other constructs are therefore systematically biased. So, although many reports claim to find evidence that overconfidence is associated with various correlates, they may instead find evidence that *ability* is associated with those correlates. This confound is particularly pernicious because ability is often explicitly considered and ruled out as an explanation based on the construction of the overconfidence measures.

I begin by describing a typical paradigm used to assess differences among people in overconfidence, variations on that theme, and why there is a problem. I next present a mathematical model to formalize and quantify this bias. I examine whether these theoretical predictions hold in data and can account for established findings using two previously collected datasets. First, using data from Moore and Healy (2008), I find that measures of overconfidence predict subsequent performance, consistent with an account in which measures of overconfidence are confounded with ability. Second, using four studies from the American Life Panel as documented in Parker et al. (2012), I reexamine the relationship between overconfidence and financial planning. I find that models in which there is no overconfidence or there is overconfidence but it is unrelated to planning except through its relationship with ability are sufficient to explain the overall data pattern. I close by discussing approaches that address the problem using statistical techniques that account for measurement error, as well as limitations of those approaches.

Common Paradigms Used to Measure Differences in Overconfidence

Research on individual differences in overconfidence has used a dizzying array of measures. When using a single measure of performance and a single self-evaluation, there are at least 20 different ways that overconfidence may be measured, not including cases in which self-evaluations are future expectations. Each and every one of them is confounded with ability. Such measures vary in terms of whether the measures assess absolute or relative performance, whether self-evaluations assess performance or ability, whether the self-evaluation measure is in the same metric or a different metric as performance, and whether the measures assess overconfidence by including a control variable, calculating a residual, or calculating a difference score.

Base Case

Begin by considering a plain vanilla version: 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 13-item financial literacy quiz) and then report their self-evaluation of their own performance (e.g., how many of the 13 items do they think that they got correct). The researcher then regresses self-evaluations of performance as the dependent variable on objective performance as the independent variable. The residual of this regression, reflecting how much higher or lower self-evaluations are than is warranted by objective performance, is used as a measure of overconfidence. The researcher then tests whether those residuals are predictive of some other outcome measure (e.g., financial planning) in a new analysis.

Residual vs. Control vs. Difference

There are three broad approaches researchers may take to calculating overconfidence given a measure of performance and a self-evaluation: they may use residualized self-evaluations, they may control for performance in a multiple regression analysis, or they may calculate a difference score.[4] The first two are quite similar to one another. All three are problematic, though for somewhat different reasons, as described in the subsequent sections. If researchers *control for* the performance measure, the partial regression coefficient estimate on

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

self-evaluation is precisely the same as the coefficient estimate on the residualized estimate. The regression approach controlling for performance rather than residualizing self-evaluation has the benefit of reducing error variance in the analysis of the outcome measure, thereby providing a more precise estimate. An alternative approach following the same participant experience uses a difference score. In a typical study designed to assess overconfidence using difference scores, the same performance measures are collected, but the researchers calculate the difference between the self-evaluation of performance and the measure of objective performance.[5]

Self-Assess Using the Same vs. Different Metric

The self-assessment may be taken in the same or a different metric. Above, both performance (on a 13-item quiz) and self-evaluation (out of 13 items) are in the same metric. Alternatively, researchers may assess self-evaluations with a different metric (e.g., a 1-7 scale). If the self-evaluation is in a different metric, overconfidence may be assessed using the residual or covariate method but should not be assessed using a difference score.

Self-Assess Performance vs. Ability

Participants may be asked to evaluate their performance or their ability. The cases above represent self-evaluations of task-specific performance. In other cases, the self-evaluation may be an evaluation of ability rather than an evaluation of performance. For example, after completing a 13-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-assessed ability on performance (or control for performance in multiple regression) to consider the role of subjective

[5] Researchers will occasionally use the difference method and then also control for objective performance. In such a case, the coefficient on the difference score is equivalent to that on self-evaluation when simply controlling for performance. There is no benefit to calculating the difference score first.

confidence. Although typical examples of self-assessed ability tend to be in a different metric, it need not be. For example, researchers could inquire about expected performance on a 13-item test drawn from the same test bank to assess ability in the same metric.

Absolute vs. Relative Assessment

Performance and 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. Self-evaluations of relative standing are often measures of performance, but could instead be measures of evaluations of ability.[6]

Variations on a Theme

These variations may be assembled in any combination as long as it does not involve taking a difference between two measures in different metrics. Evaluations may also be assessed item-by-item to enable assessment sensitivity or efficiency (e.g., Burson et al., 2006; Fleming & Lau, 2014). As detailed next, without proper adjustment, each of these approaches results in a measure of overconfidence that is confounded with ability. As a result, using any of these measures biases measures of the relationship between overconfidence and outcome measures. The confound (and resulting bias) is present whether the residual, covariate, or difference approach is taken, for both overestimation and for overplacement, whether self-evaluations are of performance or of latent ability, and whether they use the same or different metrics.

What's the Problem?

The problem arises from four properties of these performance and self-evaluation measures. First, people typically differ in ability. Second, they typically have some insight into

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

their ability. Third, performance is typically an imperfect measure of ability: it includes some noise and is unlikely to fully and only reflect the construct it is intended to measure. Fourth, self-evaluations of performance are typically ambiguous: if people were able to unambiguously assess their performance, there would be little potential to show earnest overestimation. Such ambiguity appropriately leads people to use their prior beliefs.

Because self-evaluations of performance are ambiguous, they ought to regress towards self-evaluations of ability. If two well-calibrated quiz-takers each scored an 80%, and one believes she scored a 90% and another believes he scored a 70%, there is good reason to suspect that the first test-taker is indeed more-knowledgeable than the second. Given that self-evaluations of performance are ambiguous, each of their estimates should be regressive toward their prior beliefs about their own knowledge. If people hold accurate beliefs, this leads them to be regressive toward their own actual knowledge.[7] If self-evaluations are a weighted average of knowledge (or ability or skill) and performance, seeing that self-evaluations exceed performance indicates that knowledge exceeds performance too.

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

[7] I use this assumption regarding accurate beliefs not because it is necessarily perfectly accurate, but because it is important to observe whether one can observe results that appear to be resulting from overconfidence in the known absence of overconfidence. The extension of the model in the Appendix describes the implications for inaccurate but correlated beliefs. In such cases, the magnitude, and under some conditions potentially the sign, could vary from those demonstrated under the base case of accurate beliefs. While the specific manifestation of the confound depends on whether beliefs are accurate, the presence of the problem does not.

true ability. So overconfidence, as measured via residuals or controlling for performance, is confounded with ability. When the self-evaluation is of ability rather than performance, the confound is even more severe because the measure is of ability rather than merely being contaminated by ability. Returning to the opening example, the student may calculate overconfidence by taking the residuals after regressing evaluations on sample performance. Sampling variability in performance measures attenuates the correlation between performance and evaluations, leading the most-skilled players, like Serena Williams, to consistently have positive residuals despite being well-calibrated.

The concern above applies when self-evaluations are residualized or the analysis controls for performance. A variant arises when difference scores are used. If the measure does not fully and only measure what it is believed to measure, scores will exhibit regression to the mean. People who are very high in ability will perform moderately highly, and people who are very low in ability will perform moderately poorly. The result is that, just like the residual and covariate measures, the difference measure will be confounded with ability. Consider a 10-item quiz designed to measure financial literacy, but, unbeknownst to researchers or participants, four of the items inadvertently assess trust instead. A financially-literate but average-trusting participant expected to get 8.3 answers correct a priori, actually got 7 correct (5 of the 6 financial literacy questions and 2 of the 4 trust questions), and so, due to the inherent ambiguity, reported that they got 8 correct. A less-literate but more-trusting participant expected to get 5 answers correct a priori, actually got 7 correct (3 of the 6 financial literacy questions and all 4 of the trust questions), and so, due to the inherent ambiguity, reported that they got 6 correct. The apparent overconfidence of the first participant and underconfidence of the second participant reflect true differences in financial literacy, not a surplus or lack of confidence.[8] Returning again to the opening example, the student may have made a seemingly-innocent sampling decision to only sample matches since 2018, a period during which Serena Williams had a less-dominant record. In other words, the performance measure did not have full coverage of its intended construct.[9]

Because ambiguity leads to estimates that regress toward one's own ability, but not toward others' ability, this leads to a confound for overplacement too. Using residuals or difference scores as a proxy for overconfidence will inadvertently confound overconfidence with ability, even though that is precisely the construct that often needs to be ruled out.

Relation to Prior Critiques

The present work follows a longstanding research dialogue about whether people who are unskilled are unaware (sometimes referred to as the "Dunning-Kruger Effect," DKE; Kruger & Dunning 1999). That dialogue is not central to the current development and a complete characterization of all arguments is out of scope, but a brief discussion helps to contextualize the contribution of the present research. Impatient readers can skip this section.

The DKE is characterized by the data signature that subjective performance more-closely tracks objective performance for skilled people than it does for unskilled people. An early critique noted that part of the data signature can be accounted for by combining a Better-Than-Average effect with regression to the mean (Krueger & Mueller 2002; see also Nuhfer et al., 2016, 2017). But that critique does not address the correspondence between objective and subjective performance among the skilled versus unskilled. The absolute deviation is a function

[8] It would be inappropriate to place the 'blame' on the participant: the fact that the participant uses the 'wrong' prior (e.g., financial literacy rather than a linear combination of financial literacy and trust) should not be interpreted as overconfidence if they rely on the very construct the researchers themselves believe they are measuring.
[9] While this particular example may seem egregious (sampling since 2018 but implying sampling across one's career), the problem is ubiquitous: tests are on some topic, but the questions necessarily cover only a portion of it.

of item ease or difficulty which can lead to a Better-Than-Average or Worse-Than-Average effect (Burson et al., 2006). As a result, there are circumstances under which the absolute deviation can be larger (i.e., worse) for skilled than unskilled participants, but there is still evidence of reduced correspondence among unskilled participants. Yet these findings do not speak to the present concern regarding how relative overperformance is confounded with ability.

Recent research using alternative approaches further supports the argument that the unskilled are indeed unaware. Feld et al. (2014) use instrumental variables and find evidence for the DKE. However, their model assumes use of a difference score and non-regressive (though noisy) performance measures, assumptions I relax. Jansen et al. (2021) present a Bayesian account of the DKE. They find that much, though not all, of the effect can be accounted for through Bayesian belief updating. But their model does not explore the consequences of well-calibrated beliefs (as they consider responses conditional on potentially miscalibrated beliefs) and does not discuss the broader implications beyond the DKE.

Importantly, the DKE is indicated by a multifaceted data signature. But the claimed association of overconfidence with various correlates often relies solely upon a correlation or regression coefficient. The present work shows such single statistics are insufficient to establish even a correlational association with overconfidence that cannot be accounted for by ability.

Calculating the Extent of Bias in Measures of Overconfidence

It is possible and useful to formalize and quantify the bias qualitatively described above. Specifically, a straightforward extension of Moore and Healy's (2008) model of overconfidence permits a focus on individual differences.[10] People differ in ability or skill S_i, distributed with

[10] This is related to a discussion in Healy and Moore (2007) and footnote 2 in Moore and Healy's (2008) in which they separate out expectations of ability from luck, but the implication for potential bias in the measure of overconfidence is not addressed in those discussions.

mean of 0 and variance of 1. Supposing they have perfect insight into their own skill, self-evaluations of skill, $\tilde{S}_i$, are equal to true skill S_i. This perfect insight assumption is used not because it is necessarily accurate, but because it presents an important baseline to consider: Is there apparent evidence of overconfidence even when there is none?[11] But there is indeed good reason to believe people can and do have meaningful insight into their own ability. The Subjective Numeracy Scale (Fagerlin et al., 2007) was developed to find a way for people to self-report their own numeracy using a less-burdensome task than a math test. Objective financial literacy shows correspondence with subjective financial literacy (Lusardi & Mitchell 2017). Objective knowledge and subjective knowledge are correlated across a range of domains (Carlson et al., 2009). Across multiple domains, there is good reason to expect people to have insight into their own abilities.

Although skill or ability varies across people, it is not directly observable. Instead, people's performance, P_i, is assessed via a proxy task. Performance is the result of skill and luck:

$$P_i = \lambda S_i + \nu_i \tag{1}$$

where luck, ν_i, has a mean of 0 and variance of $\sigma^2_{\nu_i}$. λ represents performance's loading on skill. A perfect measure that fully and only captures the focal skill has $\lambda = 1$ whereas an invalid measure (e.g., a measure of pure noise or a measure of something else) 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 phrenologists' head measurements. Both measures contain noise, but for Raven's matrices we expect $\lambda > 0$ (whether or not $\lambda = 1$) whereas for the phrenologists' head measurements we expect $\lambda = 0$.

[11] If individuals have inaccurate but correlated beliefs about their own skill, the residual and difference measures will still be biased, though the magnitude and potentially sign of the bias will be different. See Appendix for details and simulation.

People's self-evaluations, $\widetilde{P_i}$, are noisy measures of performance, P_i. After complete feedback, performance may be unambiguous. But prior to such feedback, people have uncertainty regarding how they performed; if they did not, their evaluations would simply be reports of actual performance. As Moore and Healy (2008) persuasively argue, the presence of such uncertainty should lead to self-evaluations that incorporate prior beliefs through Bayesian-like reasoning (whether or not people are normative Bayesian updaters). As a result, people ought to evaluate their own performance as lying somewhere between their prior beliefs and their true performance, plus noise, where the weight on prior beliefs increases with ambiguity. So:

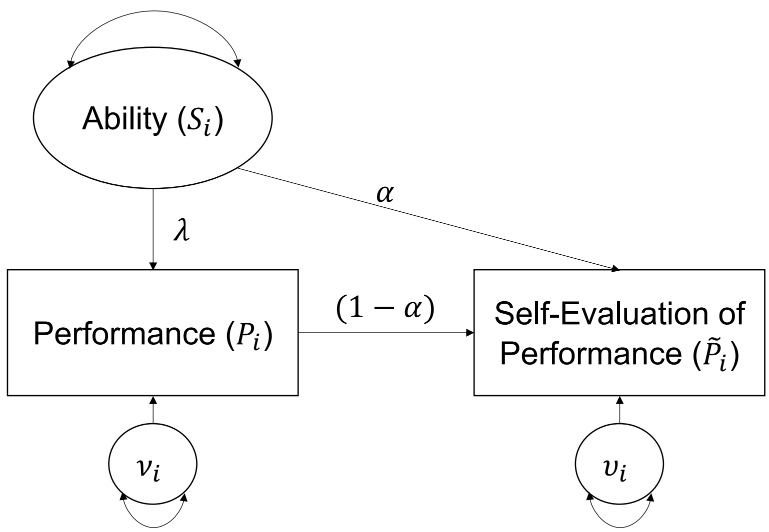
$$\tilde{P}_i = \alpha\widetilde{S_i} + (1-\alpha)P_i + \upsilon_i \tag{2}$$

where υ_i has a mean of 0 and variance of $\sigma^2_{\upsilon_i}$. α between 0 and 1 represents the ambiguity of someone assessing their own performance. As ambiguity increases, α gets closer to 1, and self-evaluations of performance reflect their self-evaluations of skill to a greater extent. When self-evaluation measures are measures of ability rather than measures of performance, $\alpha = 1$, as the measure is only a measure of ability and is not designed to assess performance at all. The measurement model is depicted in Figure 1. Because there is no guarantee that the scale used to measure confidence is the same as, or will be used in the same way as, the performance metric, α and $(1-\alpha)$ may need to be adjusted by a scaling parameter, θ.

The researcher then isolates the role of overestimation by: (a) regressing $\tilde{P}_i$ on P_i and keeping the residual, (b) by including both self-evaluation $\tilde{P}_i$ and performance P_i in a single multiple regression model, or (c) taking the difference between $\widetilde{P_i}$ and P_i. I first address the residual and multiple regression approaches together (as they are equivalent with respect to regression coefficients), and then the difference score. Both are potentially problematic.

Figure 1

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



Note. This depiction assumes beliefs about ability, $\tilde{S}_i$, are equal to ability, S_i.

Residual and Multiple Regression Approaches

To calculate overconfidence using residuals, the researcher regresses self-evaluations on performance:

$$\tilde{P}_i = \gamma P_i + \epsilon_i \tag{3}$$

The residuals, $e_i = \hat{\epsilon}_i$, are kept as the measure of overconfidence.[12] Because (a) performance is noisy, and (b) self-evaluations incorporate priors on ability, the expected errors (and thus the residuals) vary with ability:

$$E[\epsilon|S] = \left(1 - \frac{\lambda^2}{\lambda^2 + \sigma_v^2}\right)\alpha S \tag{4}$$

The residual from regressing self-evaluation on performance is positively confounded with skill if $\left(1 - \frac{\lambda^2}{\lambda^2 + \sigma_{v_i}^2}\right)\alpha > 0$. In other words, there is a confound if two conditions hold. The

[12] Throughout, I exclude intercepts for simplicity; because my focus is on individual differences in overconfidence rather than mean levels of overconfidence, intercepts can be addressed by centering variables as necessary.

first is simply that there is error not attributable to the construct being measured ($\frac{\lambda^2}{\lambda^2+\sigma_\nu^2} < 1$, or $\sigma_\nu^2 > 0$). The absence of measurement error is the exception, not the rule, so this condition is likely to be met. The second is that self-evaluations are related to skill conditional on performance ($\alpha > 0$), not just through performance. Any application, correct or incorrect, of basic Bayesian logic in the presence of uncertainty will lead to a direct effect of skill on self-evaluations, so this condition is likely to be met as well. Multiple regression can be rewritten as a regression of residuals on residuals, so the regression coefficient on evaluations controlling for performance will be precisely the same as the regression coefficient on residualized evaluations, though the multiple regression estimate will be more-precise.

Given the broader literature on measurement error in predictors (e.g., Birnbaum and Mellers, 1979; Culpepper & Aguinis 2011; Gillen et al., 2019; Kahneman, 1965; Yarkoni and Westfall, 2016), why does the current paradigm deserve special consideration? First, unlike subjective responses to 7-point scales or preferences as measured by intertemporal choice or risk preference tasks, performance measures contain a veneer of precision and objectivity that may wrongly evoke less concern regarding its status as a noisy measure of ability. Second, without the extension of Moore and Healy's (2008) model of overconfidence, it may not be obvious that self-evaluations themselves would be confounded with ability, granting a false sense of security regarding the impact of any measurement error in performance.

Difference Score Approach

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

$$\Delta_i = \tilde{P}_i - P_i \tag{5}$$

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

$$E[\Delta|S] = (1 - \lambda)\alpha S \tag{6}$$

The difference is confounded with skill if $(1 - \lambda)\alpha > 0$. Once again, it is confounded if two conditions hold: first, if performance does not perfectly load on skill ($\lambda < 1$),[13] and second, if self-evaluation is related to skill conditional on performance ($\alpha > 0$), not just via performance.

In the idealized case in which the measure of performance fully and only measures the construct that researchers and participants think it measures such that $\lambda = 1$, there would be no association between the difference score and skill. Similarly, and as in the residual case, if self-evaluation only depends on performance and not skill, then $\alpha = 0$ and there would be no relationship between the difference and skill.

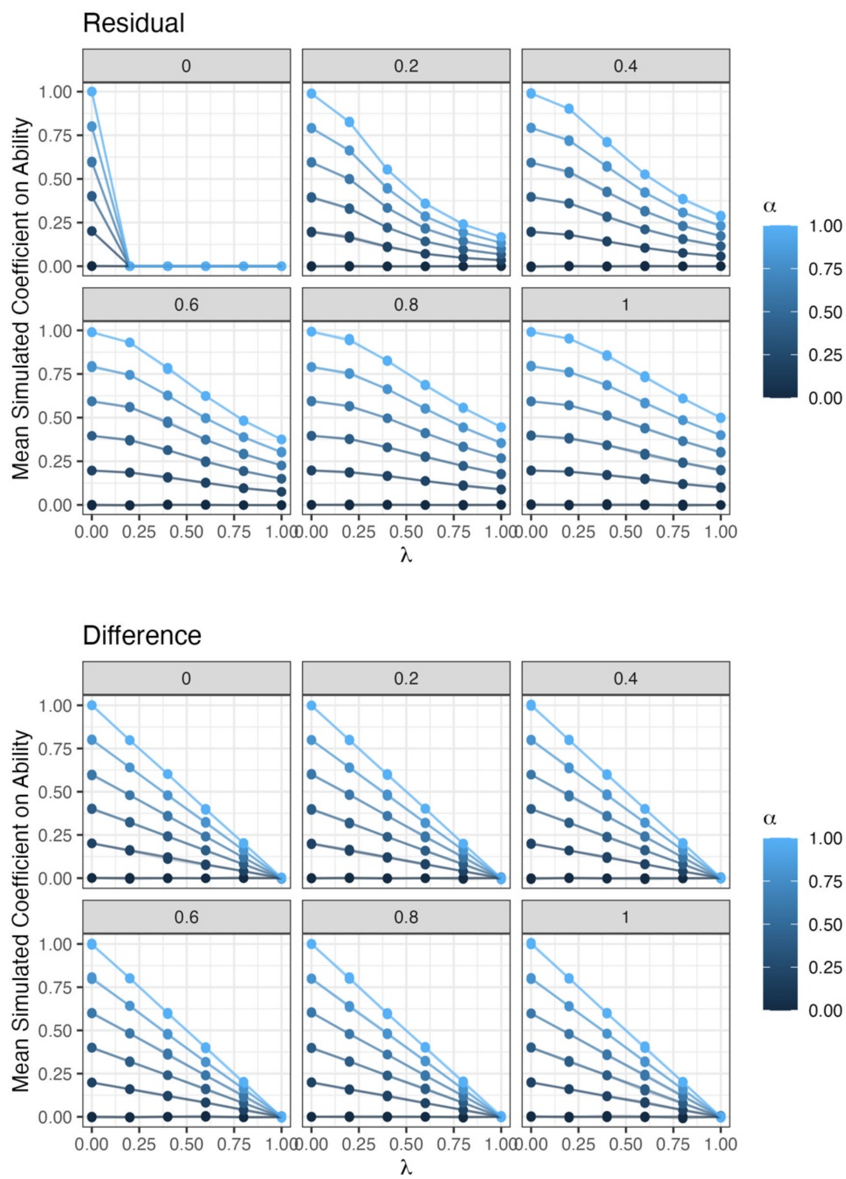
For both the residual and difference measures, the same confound holds for both overestimation of absolute performance and overplacement of relative performance. Although evaluations of one's own performance are regressive to one's own idiosyncratic prior, evaluations of others' performance are not. As a result, the bias in one's absolute performance (overestimation) carries over to one's relative performance as well (overplacement).

Simulations show that these asymptotic results hold for reasonable sample sizes. For each of 1,296 combinations of parameter values (i.e., all factorial combinations of each of λ, α, σ_ν^2, and σ_υ^2 taking a value in [0, 0.2, 0.4, 0.6, 0.8, 1.0]), I simulated 1,000 samples of 100 observations each. In each sample, skill was drawn from a standard normal distribution, and error terms were drawn from standard normal distributions scaled by the corresponding variance. Performance and self-evaluations followed from the model. Simulation results are depicted in Figure 2. The Appendix presents results for correlated inaccurate beliefs.

[13] Measurement error (σ_ν^2) does not matter in this stylized case, although in practice, P_i typically has an upper and lower bound, such that $\sigma_\nu^2 > 0$ would drive effective λ down.

Figure 2

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



Note. Facets represent different values of σ_v^2. σ_v^2 is plotted but not apparent as it does not affect the bias. For the residual score panel, when $\sigma_v^2 = \lambda = 0$, the coefficient is unexpectedly not 0. This is because performance does not vary, so the coefficient on performance predicting self-evaluations is not estimable, and so the residuals are merely mean-centered self-evaluations.

This confound matters in theory. Does it matter in practice? Reanalysis of two datasets, from Moore and Healy (2008) and Parker et al. (2012), respectively, provides a resounding yes.

Empirical Application I: Overconfidence Predicts Performance

The analysis above indicates that such a bias is possible under appropriate conditions. Is it likely to actually distort inferences? This depends on typical parameters, whether people truly exhibit (even imperfect) Bayesian updating, and whether they have sufficient self-insight. To test this, I first examine a case where: (a) there is a measure of performance, (b) there is a self-evaluation of that performance, and (c) there is an outcome measure that is a priori likely to be related to skill and unlikely to be related to overconfidence. Specifically, I consider a case where the outcome is a future measure of performance, using data from Moore and Healy (2008).[14] Future performance cannot cause past overconfidence, and it is unlikely that past overconfidence causes future performance in a way that is not attenuated as the number of intervening tasks increases. As a result, finding that past residuals or difference scores predict future performance can illustrate that the past residuals or difference scores are in fact confounded with skill.

Transparency and Openness

Decisions regarding designs, sample sizes, manipulations, and measures were by the original researchers, as all analyses in this manuscript are of secondary data. All data exclusions are solely based on missing data. Data were analyzed using R, v4.3.0 (R Core Team, 2023) and the following packages: *tidyverse* v2.0.0 (Wickham, et al., 2019), *lavaan* v0.6-15 (Rosseel, 2012), *estimatr* v1.0.0 (Blair, et al. 2022) and *eivtools* v0.1-8 (Lockwood, 2018). These analyses were not preregistered. Data and materials for the first application are available via https://osf.io/6tecy/ and Moore and Healy (2008). Data and materials for the second application are available via https://alpdata.rand.org/ upon registration and Parker et al. (2012). All code is posted at https://researchbox.org/1597&PEER_REVIEW_passcode=ORRDVP.

[14] Moore and Healy do not make the inferential error common in the literature. Rather, the availability and richness of their data present a useful opportunity to examine whether the error can affect real inferences.

Overconfidence Paradigm

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

Analysis of these data requires addressing two key issues. First, quizzes systematically differ in difficulty. Performance on other quizzes provides a proxy for skill, such that if trivia quiz skill 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 skill may be high, but also that difficulty may be low. Indeed, the block-randomized design leads to negative autocorrelation in difficulty between successive quizzes: a hard quiz is more likely to be followed by an easy quiz than it is by another hard quiz. This mechanically generates a negative relationship between performance on one quiz and performance on the subsequent quiz. To address this issue, I consider expectations, estimates, and performance for *blocks* (where each block is a triplet of quizzes) rather than for *quizzes*. A block always consists of one easy, one medium, and one hard quiz, reducing the extent to which performance on one block is negatively correlated with performance on other blocks.

Second, the data provide rich within-subject data but with an arguably modest sample size for between-subject analyses by current standards (82 participants). To exploit the within-subject data, I consider performance on sets of sequential quizzes, clustering errors at the subject level. For example, to examine the presence of skill, I regress block performance on previous block performance. In this analysis, 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). Alternative approaches to addressing these concerns are generally consistent, as long as one uses the within-subject design to maximize statistical power.

A puzzle: Overconfidence predicts subsequent performance

Using the first five quiz blocks to provide measures of performance and self-evaluations, I follow the established approaches from the literature to construct three measures of overestimation: residualized self-evaluations, controlling for performance, and the difference.

Overconfidence as assessed via residualized self-evaluations predicted performance in the next block (b = 0.298, SE = 0.141, t(38[15]) = 2.11, p = .042, 95% CI: [0.012, 0.584]). Overconfidence as assessed via the partial coefficient on self-evaluations controlling for performance also predicted performance in the next block (b = 0.298, SE = 0.088, t(38) = 3.40, p = .002, 95% CI: [0.121, 0.476]).[16] Overconfidence as assessed via the difference did not significantly predict performance in the next block (b = 0.085, SE = 0.121, t(37) = 0.71, p = .485, 95% CI: [-0.160, 0.330]).[17]

[15] All degrees of freedom throughout this reanalysis of Moore and Healy (2018) are estimated due to clustering.
[16] As noted earlier, this coefficient is necessarily equal to that on the residual, but estimated with more precision as the variance shared by the dependent variable and performance is accounted for in the multiple regression analysis but not the residual analysis.
[17] Similar results held for residualized (b = 0.401, SE = 0.103, t(42) = 3.89, p < .001, 95% CI: [0.193, 0.609]) and partial regression coefficent (b = 0.401, SE = 0.056, t(42) = 7.13, p < .001, 95% CI: [0.287, 0.514]) measures of relative performance (overplacement). The difference score measure of overplacement showed a significant negative coefficient (b = -0.188, SE = 0.062, t(52) = -3.05, p = .004, 95% CI: [-0.311, -0.064]). This may be attributable to beliefs that are not perfectly calibrated. See Appendix for details.

One could tell a story about how overconfidence truly improves subsequent trivia quiz ability. Any such story would depend on factors that would presumably dissipate over time. Yet there is no evidence that the coefficients from the residual or multiple regression analyses diminish with lags (residual: lag 2: b = 0.381, SE = 0.164; lag 3: b = 0.323, SE = 0.208; lag 4: b = 0.434, SE = 0.233; lag 5: b = 0.295, SE = 0.403; multiple regression: lag 2: b = 0.381, SE = 0.092; lag 3: b = 0.323, SE = 0.164; lag 4: b = 0.434, SE = 0.155; lag 5: b = 0.295, SE = 0.365).

Instead, the theoretical account given above parsimoniously explains this pattern: performance in the current and future blocks are both driven by skill, and the measure of overconfidence is confounded with skill. The fact that difference scores did not predict future performance may be attributed to (a) the fact that given a sufficiently-high λ in the presence of error, the bias in difference scores, $(1-\lambda)\alpha$, is smaller than the bias in residuals, $\left(1-\frac{\lambda^2}{\lambda^2+\sigma_v^2}\right)\alpha$, or (b) inaccuracy in beliefs that approximately matched λ; see Appendix.

To examine whether this theoretical account has teeth for the residual and multiple regression analyses, I examine whether the necessary components are in place: (a) Are there differences in skill? (b) Do participants have insight into their own skill? (c) Does trivia quiz performance contain error as a measure of trivia quiz skill? (Although arguably this is self-evident); and (d) Does a proxy for skill predict self-evaluations beyond performance?

Are there differences in skill? Yes

If performance is consistent across blocks, there is evidence of systematic differences in trivia quiz skill.[18] I regress performance in block t on prior performance in block t-1, clustering errors by subject. The coefficient on lagged performance was 0.754 (SE = 0.045, $t(31)$ = 16.59, p < .001; 95% CI: [0.662, 0.847]), indicating high performance on one block is associated with

[18] Skill includes ability, knowledge, and other necessary inputs that remain stable during the course of the study.

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 individual differences in skill at trivia quizzes, which are measured with noise by each quiz.

Do participants have insight into their own skill? Yes

If participants can predict how they will perform on a quiz without knowing the specific content of that quiz, it suggests they have some insight into their own trivia quiz skill. I regress pre-quiz expectations on subsequent performance, clustering errors by subject. (At the time of the pre-quiz expectation, subjects had little information on which to base their predictions, as neither the quiz difficulty nor the quiz topic was known yet.) The coefficient on performance was 0.467 (SE = 0.068, $t(31)$ = 6.87, p < .001, 95% CI: [0.329, 0.606]). This suggests participants have insight into how they will perform.[19] Given the limited information available, this is most readily attributable to awareness of their own skill.

Does trivia quiz performance contain error as a measure of trivia quiz skill? Yes

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

Does a proxy for skill predict self-evaluations beyond performance? Yes

[19] One might be concerned that participants are aware of the likely difficulty of the third quiz in each block, thereby artificially inflating this relationship: If the first quiz was a hard quiz and the second quiz was a medium quiz, it could be determined that the third quiz would be an easy quiz. The main result also holds if one only considers the first quiz from each block (adjusted for difficulty), which was completely randomized (b = 0.233, SE = 0.048, $t(31)$ = 4.80, p < .001, 95% CI: [0.134, 0.332]). The coefficient was, if anything, directionally weaker when considering the third quiz (b = 0.177, SE = 0.041, $t(45)$ = 4.33, p < .001).

The last necessary component is that participants provide evaluations that are regressive toward their own skill when reporting their self-evaluations. I cannot observe skill (if I could, I would face many fewer problems), but I can use subsequent performance as a noisy proxy. Unlike in other cases when one uses performance as a proxy for skill, here the only concern is that it contains sufficient signal, not that it excludes sufficient noise. I regress self-evaluations on current performance and subsequent performance, where subsequent performance serves as a noisy proxy for skill. The coefficient on current performance was 0.880 (SE = 0.035, t(49) = 24.94, p < .001; 95% CI: [0.809, 0.951]), indicating that participants indeed have some idea of how well they did on each block (though this coefficient also partially captures the role of skill). Critically, the coefficient on subsequent performance was 0.099 (SE = 0.029, t(49) = 3.43, p = .001, 95% CI: [0.041, 0.156]), indicating that future performance is predictive of current self-evaluations, controlling for current performance. The magnitude of this coefficient did not attenuate as there were more intervening blocks (1 intervening block: b = 0.119, SE = 0.034; 2 intervening blocks: b = 0.101, SE = 0.052; 3 intervening blocks: b = 0.141, SE = 0.044; 4 intervening blocks: b = 0.099, SE = 0.107). This suggests that post-quiz estimates are indeed regressive toward idiosyncratic skill in addition to assessing performance as intended.

Subsequent performance is a substitute for correlates of overconfidence

Although the puzzle suggests that overconfidence predicts future performance, a more parsimonious (and, in the current context, arguably more probable) explanation is that there are differences in skill, people have self-insight, performance is a noisy measure of skill, and self-evaluations pick up skill in addition to performance. As a result, the measure of overconfidence is confounded with skill and skill is what predicts future performance. A key problem is that many findings in the literature of an association between overconfidence and other correlates use an approach equivalent to that in the puzzle above, but then do not sufficiently consider the alternative explanation that followed.

Empirical Application II: Reassessing Correlates of Financial Planning

The analysis above is readily explained by the fact that overconfidence is confounded with skill. But this result does not cast doubt regarding whether any purported correlate of overconfidence may instead merely be a correlate of skill. Perhaps the whole endeavor is a fun statistical game with little connection to substantive claims. Using another example from the literature, I show how the current proposal ought to make us to reconsider our assessments of how individual differences in overconfidence relate to other important constructs and behaviors.

Parker et al. (2012)[20] study the role of "inappropriate confidence" (what Parker & Stone, 2014, later refer to as "unjustified confidence") in retirement planning and pithily summarizes the finding that with respect to retirement planning as "it may be more important to be confident than to be appropriately confident."[21] To draw this conclusion, the authors reported the analysis of four studies conducted with the same panel of participants over time by different research teams using the American Life Panel (ALP; Pollard & Baird, 2017). These four studies used different tasks to assess both performance and confidence. Because they all drew from a common panel of participants, each could be related to a common three-item measure of retirement planning behavior measured in Study 1. Using four separate regressions, one for each study, the authors find that each measure of confidence predicts retirement planning, controlling for the corresponding measure of knowledge along with demographic covariates.

[20] The alternative explanation I propose here is not unique to this particular paper. Rather, this paper provides a clean example that is well-structured for the current purpose, has available data, is sufficiently clearly written so as to avoid ambiguity, and is important enough to be well-cited.
[21] The paper does note the correlational nature of the findings as a caution on drawing causal conclusions. My critique applies to both causal claims and correlational claims.

An exhaustive description of the underlying methods of each of the four studies are beyond the scope of this paper; readers may consult the original paper for more details. In brief, Study 1 (*N* = 1150) assessed financial knowledge using a 13-item true/false quiz and confidence using a single 7-point measure assessing people's subjective understanding of economics.[22] Study 2 (*N* = 1114) assessed general knowledge using a 14-item true/false quiz and confidence using 14 item-by-item measures on a scale ranging from 50% = just guessing to 100% = absolutely sure. Study 3 (*N* = 1005) assessed financial literacy using a binary measure of whether participants minimized fees in an experimental task and confidence using a 5-point measure assessing people's subjective confidence in their task performance. Study 4 (*N* = 566) assessed financial sophistication using a 70-item true/false financial sophistication quiz and confidence using a 100% = surely true to 100% = surely false confidence scale.

In reanalyzing the data, I examined whether it was possible to account for the observed patterns in the data without any role for confidence in financial planning.[23] To do so, I reanalyzed the original data from the four ALP studies. Relevant correlations and descriptive statistics are given in Table 1, both as reported in the original manuscript and in my re-analysis. My calculations closely match those given in the text of the original manuscript. With one exception, all correlations are within 0.03. Such slight differences may be attributable to (a) my use of the full 14-item quiz from Study 1 whereas the original authors used a 13-item version, and (b) slight differences in sample size, presumably due to slight differences in exclusions based on missing values (in my analyses, *N*s = 1161, 1106, 988, and 584). The only exception is

[22] This confidence measure was thus a subjective measure of knowledge, not performance.
[23] Of course, such a test does not rule out a role for confidence. It simply indicates whether it is possible to account for the observed data without any role of confidence.

the correlation between Study 3 performance and Study 4 confidence, where I find $r = 0.37$ and they report $r = 0.26$. This may either be a larger discrepancy given the binary measure or a typo.

Table 1

Reported Zero-Order Correlations Among Performance Measures, Confidence Measures, and Financial Planning from Parker et al. (2012) (top) or Calculated from ALP Data (bottom)

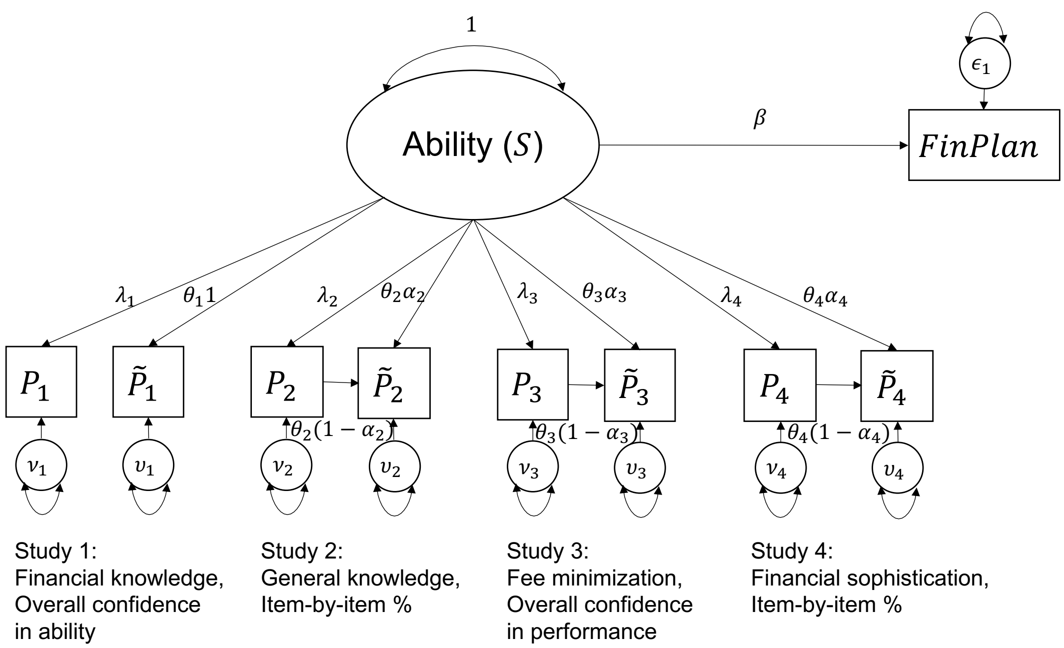
Reported	Perf1	Perf2	Perf3	Perf4	Conf1	Conf2	Conf3	Conf4	Outcome
Perf1									
Perf2	0.29								
Perf3	0.35	0.16							
Perf4	0.63	0.33	0.38						
Conf1	0.37		0.18						
Conf2		0.34	0.15		0.19				
Conf3			0.30		0.31	0.19			
Conf4			0.26	0.53	0.34	0.42	0.38		
Outcome					0.21	0.20	0.19	0.26	
N	1150	1114	1005	566	1150	1114	1005	566	1150
Mean	0.75	0.93	0.33	0.74	4.53	0.89	3.51	0.78	0.46
SD	0.21	0.10		0.10	1.26	0.07	0.89	0.11	0.44

Calculated	Perf1	Perf2	Perf3	Perf4	Conf1	Conf2	Conf3	Conf4	Outcome
Perf1									
Perf2	0.31								
Perf3	0.34	0.16							
Perf4	0.63	0.33	0.39						
Conf1	0.36	0.08	0.19	0.25					
Conf2	0.34	0.33	0.16	0.32	0.20				
Conf3	0.29	0.10	0.31	0.29	0.34	0.21			
Conf4	0.53	0.05	0.37	0.53	0.35	0.41	0.39		
Outcome	0.35	0.21	0.14	0.29	0.22	0.20	0.21	0.25	
N	1161	1106	988	566	1161	1106	988	566	1161
Mean	0.77	0.93	0.36	0.74	4.53	0.89	3.53	0.78	0.47
SD	0.20	0.08	0.48	0.10	1.25	0.07	0.90	0.11	0.44

I fit these correlations to the model depicted in Figure 3. Importantly, there is no latent confidence in this model. Instead, I model the four performance measures as measures of financial knowledge, each confidence measure as a measure of financial knowledge (Study 1) or both financial knowledge and study-specific performance (Studies 2-4), and financial planning as a consequence of financial knowledge alone. To do so, I fit 21 parameters: β (a single coefficient representing the relationship between knowledge and retirement planning), σ_ϵ^2 (the error for retirement planning), 4 λs and 4 σ_ν^2s (one for each study's performance measure), 3 αs and 4 σ_υ^2s (one for each study's confidence measure except for Study 1; α for Study 1 was fixed to 1 because that confidence measure assessed ability), and 4 θ scaling factors (one for each study's confidence measure).[24] The model was fit using full information maximum likelihood for missing data using the `lavaan` package v0.6-12 (Rosseel, 2012) in R.[25]

Figure 3

Model Accounting for Relationships Among Performance, Measures of Confidence, and Financial Planning in the Absence of Overconfidence



[24] The scaling factor was necessary to account for scale use. To facilitate estimation, rather than estimating θ and α directly, I estimated $\theta\alpha$ and $\theta(1-\alpha)$. θ was then calculated as $\theta\alpha + \theta(1-\alpha)$ and α as $\frac{\theta\alpha}{\theta\alpha+\theta(1-\alpha)}$.

[25] Although variables were standardized prior to estimation, in addition to the 9*8/2 = 36 covariances, the model was fit using an additional 9 variances and 9 means. In addition to the 21 parameters noted above the model fit 9 intercepts. Thus, there were 54 observations fit using 29 total parameters, leaving 25 degrees of freedom.

This model is clearly misspecified in several ways unrelated to latent confidence. First, the model makes no allowance for common method bias, but self-evaluations were assessed using item-by-item percentage confidence reports for Studies 2 and 4 and single 7- or 5-point items for Studies 1 and 3. Second, the model makes no allowance for the fact that participants completing the general knowledge scale should show self-evaluations that regress toward their *general* knowledge, not their financial knowledge. There are a priori reasons to expect that the relatively simple model depicted in Figure 3 is insufficient to fully account for patterns in the data because it is known to be wrong in ways unrelated to the addition of overconfidence.

Despite these model mismatches, the estimated parameters appeared reasonable; see Table 2. λs for the general knowledge quiz (0.34) and fee-minimizing task (0.43) were lower than those for the financial literacy quiz (0.80) and financial sophistication quiz (0.76). This is consistent both with theory (e.g., the general knowledge quiz ought to load on financial knowledge less than the financial quizzes should, and the fee-minimizing measure is almost certainly affected by other factors) as well as reported scale reliabilities (Cronbach's α was lower for the general knowledge quiz than either financial quiz). The estimated link from financial knowledge to behavior was moderate (0.42).

Table 2

Standardized Parameter Estimates from Model Excluding the Possibility of Overconfidence

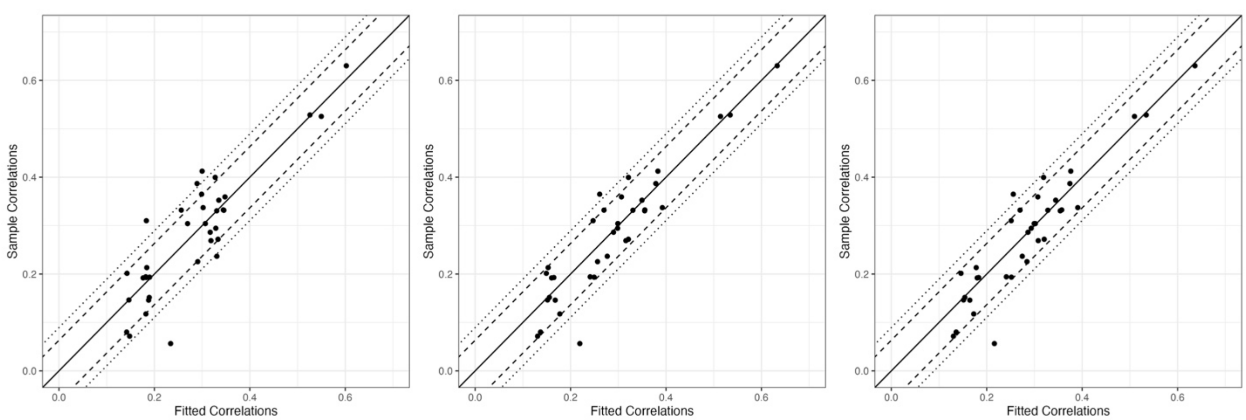
Study	λ	α[a]	θ[a]	σ_ν^2	σ_υ^2	β[b]	σ_ϵ^{2}[b]
Study 1	0.80	1.00[c]	0.44	0.37	0.81	0.42	0.82
Study 2	0.34	0.64	0.57	0.89	0.77		
Study 3	0.43	0.70	0.51	0.81	0.81		
Study 4	0.76	0.99	0.69	0.43	0.52		

[a] Calculated after rescaling.
[b] Held constant across studies.
[c] Fixed by theory, not estimated.

As shown in Figure 4 and Table 3, the set of correlations derived from the fitted parameter estimates fit the observed data moderately well, especially considering the ways in which it is known to be inadequate. The largest absolute deviations are also instructive. First, the model overestimates the correlation between Study 2 performance and Study 4 confidence by 0.18. Notably, Study 2's performance measure is of general knowledge, not financial knowledge, so it may not load on ability equivalently to the other measures. Second, the model underestimates the correlation between Study 1 confidence and Study 3 confidence by 0.13 and the correlation between Study 2 confidence and Study 4 confidence by 0.11. Studies 1 and 3 assessed confidence via 7- or 5-point scales and Studies 2 and 4 assessed confidence via item-by-item percentage confidence. In other words, the model may fail to capture patterns in the correlations due to factors unrelated to the presence or impact of overconfidence.

Figure 4

Fitted Correlations and Observed Correlations in the Data in Three Models



Note. The left panel represents the model shown in Figure 3. The center panel allows for the presence of, but no effect of, overconfidence. The right panel allows for both the presence and effect of overconfidence. The solid line represents a perfect match between the sample correlations and the fitted correlations. The dashed lines represent $\pm \frac{2}{\sqrt{1000}}$, very roughly the 95% confidence band for N = 1000 (largest correlation N = 1161). The dotted lines represent $\pm \frac{2}{\sqrt{500}}$, very roughly the 95% confidence band for N = 500 (smallest correlation N = 500).

Table 3
Model Fit Statistics

Model	df	χ^2	CFI	RMSEA	logLik	AIC	AICc[a]	BIC
1 (Ability)	24	194.09	0.90	0.072	-11735	23530	23607	23687
2 (Correlated Confidence)	23	124.23	0.94	0.056	-11700	23462	23548	23624
3 (Causal Confidence)	22	121.62	0.94	0.057	-11699	23461	23557	23629
Just S1, S4								
4 (Ability)	4	19.08	0.98	0.057	-6113	12258	12394	12339
5 (Correlated Confidence)	3	8.01	0.99	0.038	-6108	12249	12453	12335
6 (Causal Confidence)	2	2.47	1.00	0.014	-6105	12246	12588	12337

[a] Corrected AIC to account for a small number of variances and covariances.

The baseline model in Figure 3 does an excellent job of accounting for qualitative patterns in the data and an adequate job of accounting for specific quantitative patterns in the data; see Table 3 Model 1. I also considered two additional models by freeing implied fixed parameters. In the first (Model 2), I allow the confidence measures to load on a separate correlated confidence construct (i.e., $\tilde{S}_i$) rather than ability directly (i.e., S_i), as in Appendix Figure A1. This is a nesting model, as it is equivalent to the baseline model if the correlation between ability and confidence is fixed to 1. Again, only ability is allowed to affect financial planning. In the second nesting model (Model 3), I free a parameter to allow confidence to independently affect financial planning; this path is fixed to 0 in the first two models. As shown in Table 3, both models somewhat outperform the baseline model. But there is little to no evidence that allowing confidence to impact financial planning in Model 3 improves fit beyond merely allowing confidence to be imperfectly correlated with ability in Model 2 ($\hat{\rho} = 0.72$). The improvement in fit in the model allowing for an influence of confidence (relative to the model for confidence as a correlated construct) is not worth the extra parameter given the very slight improvement in χ^2, log likelihood, and AIC, and decrement in small-sample corrected AIC and BIC. Moreover, in that third model, the estimate of the latent relationship between confidence and financial planning, controlling for ability, is only marginally significantly different from 0 ($\beta_2 = 0.11$, $z = 1.65$, $p = .099$). None of the models adequately account for the correlation between Study 2 performance and Study 4 confidence (i.e., the negative outlier that is apparent in each panel of Figure 4).

Taken together, these analyses suggest that a parsimonious representation of the reported data can be derived from a simple model with just knowledge and without (inappropriate, unjustified, or over-) confidence. Some evidence suggests that the model allowing beliefs about skill to be imperfectly correlated with skill fits better, but there is no evidence to suggest that the model with a causal role for overconfidence improves fit further. Even the fit improved by allowing confidence to be correlated with ability may in part be attributable to differences in the relevant constructs assessed across studies and/or common method bias. If one fits the model using only Study 1 and Study 4 (in which we can be more assured that the measured ability construct is the same, and across which there is reduced common method bias), even enabling confidence to be a separate construct from ability is not favored by all comparison statistics (see corrected AIC), although the models are nearly saturated and leave very few degrees of freedom. These results are given as Models 4, 5, and 6 in Table 3.

This analysis does not indicate that inappropriate confidence plays no role. Instead, it indicates that the reported evidence is not sufficient to indicate that it does play a role (or is even non-causally correlated). Indeed, there may be other evidence even in the same datasets that could bolster the role of inappropriate confidence. This analysis merely indicates that the typical reported evidence does not provide a strong basis on which to draw the conclusion that inappropriate confidence is relevant to financial planning beyond mere ability or knowledge.

Accounting for Measurement Error

Both in theory and in practice, widely-used measures of overconfidence are problematically confounded with ability. What solutions are available beyond either despair or wishing away the problem? The difference score approach in the first application did not indicate a problem—but this provides cold comfort. In part, this may have been due to lack of power. In other cases, using a difference score may be impossible if self-evaluations are in a different metric as performance, and difference scores may introduce additional undesirable mechanical relationships. But perhaps most importantly, use of a difference score requires the strong and untestable assumption that the performance measure has a unit loading on the ability construct. In other words, even if the true bias were 0 in the first application, that provides no guarantee that there would be no bias in other applications.

To provide an unbiased test, it is useful to recall the conditions under which there is no bias for the role of self-evaluation when controlling for performance. The bias is eliminated if either: (a) $\alpha = 0$, meaning there is no ambiguity and participants have no reason to regress their self-evaluations towards their prior beliefs, or (b) $\frac{\lambda^2}{\lambda^2+\sigma_v^2} = 1$, meaning there is no measurement error and performance is at least partially related to ability. This latter case is a traditional case in which measurement error in one independent variable (performance) biases both its own coefficient and the coefficients of correlated variables. Possible solutions to address this include structural equation models and errors-in-variables.

Structural Equation Models

Structural equation models (e.g., Kline, 2005) permit the researcher to model the relationships among latent variables, unattenuated by measurement error. Indeed, this is the approach taken to model the results above using Parker et al.'s (2012) data. This typically relies upon multiple indicators of performance, although as noted by Westfall and Yarkoni (2016), it is feasible to use such models with an estimate of reliability even without multiple indicators. Each measure of self-evaluation is then permitted to load both on ability as well as its corresponding performance indicator(s). The key assumption is that the common variance underlying performance measures reflects the ability that the performance measure is purported to tap into. If the performance indicators share variance not attributable to ability, this may falsely suggest little error, when in fact it could merely be little idiosyncratic error but considerable shared error.

Errors-in-Variables

Even with a single performance measure, established solutions for errors in variables can prove useful given a measure or assumption of reliability of each measure (e.g., Fuller, 1980, 1987; Culpepper & Aguinis, 2011). Once again, a key assumption is that the reliability estimate appropriately captures all components of variance other than ability. For example, if the performance measure reliably picks up a linear combination of both financial knowledge and trust in institutions and we assess reliability via test-retest reliability, our measure of reliability may be considerably higher than $\frac{\lambda^2}{\lambda^2+\sigma_v^2}$, the relevant quantity, leading us to underestimate the extent of the problem.

Culpepper and Aguinis (2011) provide a useful background regarding how Fuller (1980, 1987) helped to develop and disseminate errors-in-variables methods. A full discussion is beyond the scope of this article; the interested reader can consult Culpepper and Aguinis (2011) for a treatment of the issue intended for psychologists. In short, the estimate and standard error of the coefficient on each predictor in a model may be adjusted in accordance with the reliability of that predictor and the other predictors. More-reliable measures lead to less adjustment; less-reliable measures lead to greater adjustment. Adjustments are made not just to the coefficient on the unreliable measure, but also the coefficients on the other predictors in the model. Properly accounting for the measurement error in the performance measure effectively means the model controls for ability, not just performance, which affects the coefficient on self-evaluation. The second application included an example of using structural equation modeling to address the issue of measurement error. I next consider an example using errors-in-variables methodology.

An Empirical Application of the Errors-in-Variables Approach

To examine the potential of this solution, I examine both the *eivreg* function from the *eivtools* package (Lockwood, 2018) as well as the *eiv* function provided by Culpepper and Aguinis (2011), both implemented in R. Errors in variables adjustments require an estimate of the reliabilities of each measure. This is intended to assess the ratio of the variance attributable to the latent construct to the total variance of the measure. Above I noted potential problems with using test-retest reliability. Similarly, the internal reliability (e.g., that given by Cronbach's α) may not necessarily be an appropriate measure for such purposes. For example, other irrelevant stable constructs that the measure picks up on may inflate internal reliability.

I apply this approach to the American Life Panel application from Parker et al. (2013).[26] Despite its potential flaws noted above, I rely on the reported Cronbach's α where available to

[26] I also attempted to use this approach on Moore and Healy's (2008) data. However, performance and estimates were extremely strongly correlated across participants within blocks (ranging from 0.87 to 0.96), which implies extremely high reliabilities that are not consistent with other approaches to estimating reliability (e.g., the correlation between performance and lagged performance). This is likely attributable to the fact that the randomization approach led to different participants encountering different sets of quizzes in different blocks. However, even accounting for only slightly imperfect reliabilities for both performance and self-evaluations (0.95 for each), we still observe using *eivreg* that lagged performance predicts current performance (b = 0.60, SE = 0.24, $t(407) = 2.47$, $p = .014$) but, unlike in the baseline analysis, lagged self-evaluations do not (b = 0.17, SE = 0.24, $t(407) = 0.73$, $p = 0.468$); results were equivalent using *eiv*. This reinforces the importance of accounting for even slight degrees of unreliability. Note that with such high correlations among predictors, the results are rather unstable to even slight differences in estimated reliabilities.

assess reliability (Study 1 performance: 0.77; Study 2 performance: 0.66; Study 2 confidence: 0.78; Study 4 performance: 0.75; and Study 4 confidence: 0.97). For Study 3 performance, I use its single highest correlation with another performance measures (Study 4 performance, 0.34) as an imperfect proxy. For Study 1 confidence and Study 3 confidence, I use their correlations with one another as imperfect proxies (0.31). The results (using standardized variables and *eivreg*) are given in Table 4. All results using *eiv* were quantitatively similar and led to identical statistical conclusions.

Table 4

Coefficients from American Life Panel analysis using Errors in Variables adjustments

Study	Variable	Reliability	Orig. Est.	Adj. Est.	SE	t	p
1	Performance	0.77	0.318	0.296	0.098	3.01	.003
	Confidence	0.31[a]	0.100	0.349	0.182	1.92	.055
2	Performance	0.66	0.156	0.233	0.054	4.34	<.001
	Confidence	0.78	0.141	0.147	0.047	3.16	.002
3	Performance	0.34[b]	0.081	-4.24	17.22	-0.25	.806
	Confidence	0.31[a]	0.176	4.84	17.47	0.28	.782
4	Performance	0.75	0.210	0.321	0.077	4.19	<.001
	Confidence	0.97	0.140	0.084	0.057	1.47	.143

[a] Reliability based on correlation between Study 1 confidence and Study 3 confidence.
[b] Reliability estimate based on highest correlation with another performance measure.

Although these results are not as clearly interpretable as one might like, overall they tell a story that is not consistent with a strong replicable role for confidence in contributing to the understanding of financial planning. In Studies 1 and 4, the coefficient on confidence is not significant, though it is marginally significant in Study 1 and in the expected direction in Study 4. In Study 2, both coefficients are significant, though more weight is given to performance over confidence relative to the unadjusted coefficients. The Study 3 results are effectively uninterpretable because the low estimated reliabilities substantially inflated both coefficients and standard errors. Overall, these results are quite sensitive to the assumptions about reliabilities.

A proponent of the confidence-causes-planning story might focus on the Study 2 results, finding that even after accounting for measurement error, confidence appears to play a role. A detractor from the confidence-causes-planning story might focus on the Study 4 results, given its closer connection to the construct of interest and lack of significance on the confidence coefficient. Study 1 and particularly Study 3 are difficult to interpret given the ad hoc proxies used regarding the reliability of the single item measures. Of course, the reliabilities assessed via Cronbach's α may be larger than the proper adjustment would require.

Recommendations

These initial investigations regarding ways to address measurement error through use of structural equation models or errors-in-variables adjustments suggest a way forward but are not conclusive. First, if a researcher has multiple performance and evaluation measures, one may consider using a structural equation model. Second, if a researcher has information on reliability, one may be able to rely on errors in variables instead, a technique that is perhaps less familiar to many psychology researchers. In either case, this analysis reinforces three key properties. First, use of reliable performance and confidence measures is important even when one is able to adjust for unreliability, as adjustments for severe unreliability may be unstable. Second, each approach benefits from the use of multiple measures to assess the reliability of the measures used. Yet internal reliability may not be sufficient to determine the unattenuated association using either approach, given potential problems with construct validity. Third, in the absence of reliability estimates, sensitivity analyses based on plausible degrees of reliability may still be informative. In the Moore and Healy (2008) data, even very modest deviations from perfect reliability are sufficient to change one's inferences regarding the independent role of confidence; see footnote 26. These attempts to attenuate the problem are appropriate for the residualized or covariate measures of overconfidence, but not the difference score, as the problem for the difference score is not measurement error but rather a less-than-unit loading on ability.

Is This Just a Form of Overconfidence?

Throughout, I have repeatedly returned to the notion that the measure of performance must fully and only measure the target construct to make use of the difference score measure. This matters because the prior to which people are regressing must align with the construct being measured. A mismatch (as in the financial literacy and trust example) is equivalent to $\lambda < 1$, which leads to the problem with using difference scores. A sensible critique is that this is simply a different form of unjustified confidence: people confidently use a prior that should not apply and adjust to the wrong belief as a result.

But there is a problem if researchers "blame" a participant for relying upon the wrong prior when the researchers themselves use the wrong construct. Consider again the phrenologist introduced earlier. Both the phrenologist and the patient may earnestly believe that the phrenologist is generating a good diagnostic measure of intelligence. If the participant is asked how they perform on this measure of intelligence, but they have substantial ambiguity about their own head measurements ($\alpha = 1$), they will report their true intelligence. Of course, their score on the phrenology examination will be unrelated to their intelligence ($\lambda = 0$). As a result, on average, people with a high residual or difference (i.e., those who think they received a higher score from the phrenologist than they truly did) will be more intelligent.

The skeptical reader may argue: "That is overconfidence! The participant is regressing their performance self-evaluation to their beliefs about their own intelligence when they should be regressing to their own beliefs about the shape of their head." In such a case, it would be inappropriate to fault the participant for regressing to the very construct the researcher claims to be measuring with a worthless instrument. Thankfully, most researchers are not phrenologists and are using instruments with greater validity. But greater validity than phrenology is a low bar.

This raises a thorny question regarding whether the effects of using misleading labels for a performance task ought to be considered overconfidence. If we do not accept the overconfidence label in the case described above, we perhaps ought to be cautious accepting the overconfidence label in the presence of a misleading label (Ehrlinger & Dunning, 2003).

When Is There Not A Problem?

There are two important cases in which these results regarding a bias of overconfidence with ability are unlikely to lead to qualitatively mistaken inferences. First, if there truly is no relationship between ability and the candidate correlate, then although the measure is confounded, the confound has no bite to it. Of course, no correlation between *performance* and the outcome measure of interest is not sufficient: such a lack of correspondence could merely indicate that performance is a poor measure of ability even if it is a reliable measure of something else. This would again lead to a biased estimate of the effect of overconfidence.

Second, if the relationship between ability and the outcome measure of interest and the relationship between residualized overconfidence and the outcome measure of interest have opposite signs, the bias described here could not account for such a pattern of results. This does not mean that the bias is inconsequential: indeed, it may suggest that the magnitude of the relationship between overconfidence and the outcome measure of interest is underestimated. As a result, the estimate is still biased, but qualitatively the correct inference. The relationship for difference score is more nuanced. If beliefs are correlated with but not equal to ability, it is possible under certain circumstances for the sign on the difference score to reverse as shown in the Appendix. In such cases, while there may be a correlation with overconfidence (to the extent that beliefs do not match ability are characterized as differences in overconfidence), it may still be entirely attributable to ability alone and not at all by beliefs.

Summary and Conclusion

Research and casual observation suggest that overconfidence is ubiquitous and variable. Yet widely-used measures of individual differences in overconfidence are confounded with the very thing they are designed to rule out: ability. This is because measures of performance are imperfect, so accounting for performance is insufficient to account for ability. If there is any ambiguity regarding performance, as there frequently is, measures of confidence ought to regress towards prior beliefs about ability even when they are intended to be self-evaluations of performance on a particular task. Because performance itself is an imperfect measure, the variance of self-evaluation that is attributable to ability is not fully partialed out. The result is that both residual and difference score measures of overconfidence are confounded with ability. The magnitude of this bias in the idealized case can be quantified analytically.

These confounds mean that it is possible to observe surprising results in the data: overconfidence predicts later performance after several intervening tasks. When reevaluating one set of published results on overconfidence through this lens, I find little evidence for the purported role of overconfidence in financial planning. Instead, the entire pattern of results could be driven through financial knowledge alone. If one is willing to make assumptions regarding construct validity and estimate or assume reliability of each measure, it is possible to resolve these concerns through structural equation modeling or error-in-variables adjustments. However, the empirical applications demonstrate these partial solutions are not an automatic panacea, as a number of complications may arise regarding construct validity, noisy estimates of reliability, and unstable estimates. Instead, design-based solutions may ultimately prove necessary. This work may serve as an impetus and a guide (and perhaps a wake-up call) to further improve our collective attempts to measure individual differences in overconfidence and their true associations with traits, decisions, and behaviors.

References

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

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

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

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

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

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

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60-77.

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

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

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

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.

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

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

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

Fox, J. (2009). *The myth of the rational market: A history of risk, reward, and delusion on Wall Street*. New York: Harper Business.

Fuller, W. A. (1980). Properties of some estimators for the errors-in-variables model. *The Annals of Statistics*, *8*(2), 407-422.

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

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.

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

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

Behaviour, *5*(6), 756-763.

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

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.

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

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

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*(2), 180-188.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134.

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

Lockwood J (2018). eivtools: Measurement Error Modeling Tools. *R package version 0.1-8*, https://CRAN.R-project.org/package=eivtools.

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

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

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

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

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

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

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

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

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

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

Decision Making, *27*(2), 134-145.

Pollard, M. S., & Baird, M. (2017). *The RAND American life panel: Technical description*.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686. https://doi.org/10.21105/joss.01686.

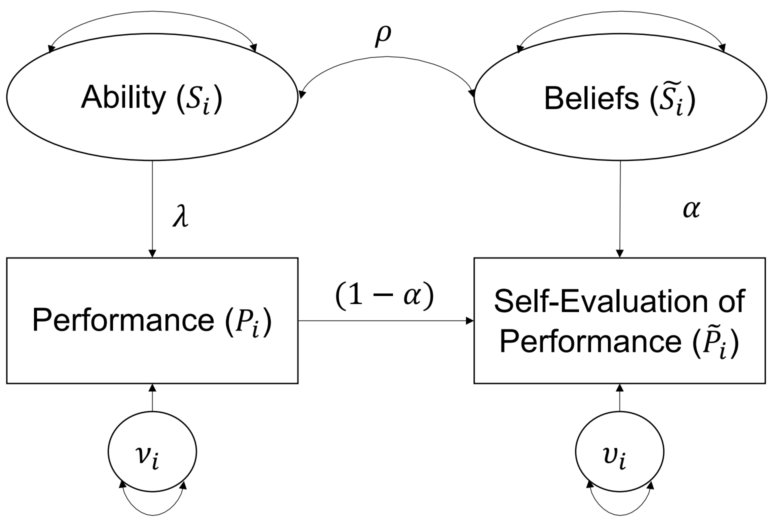
Appendix

Extending the Model to Incorporate Inaccurate but Correlated Beliefs

The main text presents a model in which people's beliefs about their own ability are accurate. Here I address the case in which beliefs may be inaccurate and merely correlated with ability. The problems described in the main text remain, though additional care is required to interpret the bias. This model may be represented via the measurement model in Figure A1.

Figure A1

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



The difference between Figure A1 and Figure 1 is that beliefs are correlated with ability, with correlation ρ, and self-evaluations regress toward beliefs rather than ability. Like ability, I assume beliefs are distributed with mean of 0 and variance of 1. The expectation of the residual is given by equation A1:

$$E[\epsilon|S] = \rho\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_\nu^2}\right)\alpha S \tag{A1}$$

Note that if beliefs are accurate, $\rho = 1$, so we are back to equation 4.

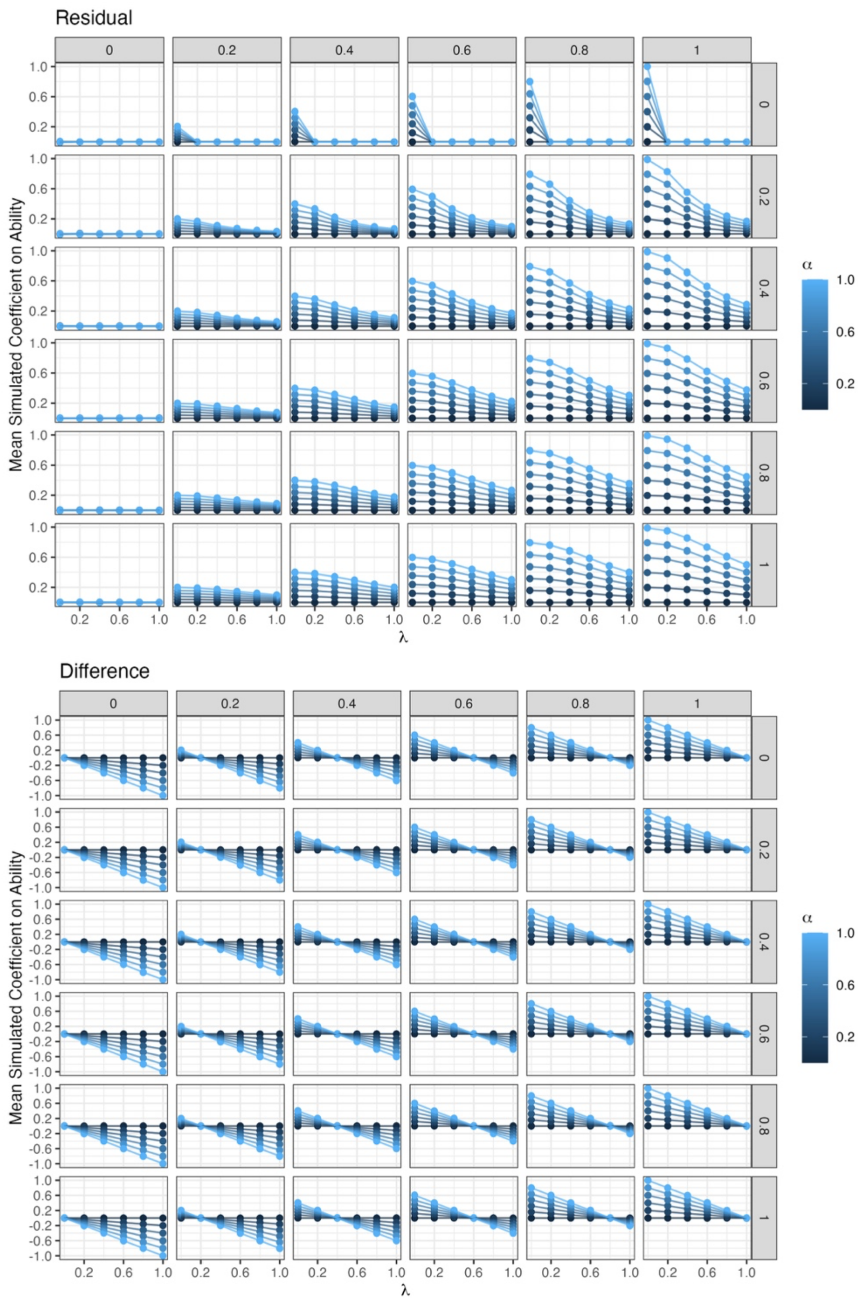
The expectation of the difference score likewise now depends on ρ:

$$E[\Delta|S] = (\rho - \lambda)\alpha S \tag{A2}$$

And again, if beliefs are equal to ability, then $\rho = 1$, so we are back to equation 6. As shown in Figure A2, simulations find these expectations hold given realistic sample sizes.

Figure A2

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



Note. Rows represent σ_v^2, variance of the error in the performance measure. Columns represent ρ, the correlation between Ability and Beliefs, where each has unit variance.

Mathematically, these equations imply that as the correlation between beliefs and skill is reduced in magnitude, the bias moves downwards. For the residual analysis, this works as a multiplier: if beliefs are unrelated to ability, then there is no bias because scores are regressive to something else. For difference scores, however, this has the potential to reverse the sign of the bias: if performance is a good measure of ability and beliefs are weakly related to ability, the difference score may be negatively confounded with ability, due to $(\rho - \lambda)$.

Note there are two distinct ways that the link between beliefs and performance may be severed. First, performance may not be indicative of ability ($\lambda = 0$). In such a case, self-evaluation remains a measure of ability. Second, beliefs may not be indicative of ability ($\rho = 0$). In such a case, self-evaluation is not correlated with ability beyond its relationship with performance. In other words, in the former case, self-evaluations controlling for performance are confounded with ability, whereas in the latter case, they are not. Of course, much of the time beliefs may be inaccurate but correlated with ability ($0 < \rho < 1$). In such cases, the measures of overconfidence remain confounded with ability.

Unlike the base case in the text, this model allows for a form of overconfidence: if beliefs are imperfectly correlated with skill, then that suggests there are individuals who differ in their degree of overconfidence. Such differences could range from underconfident to properly confident, properly confident to overconfident, or underconfident to overconfident. But the concern regarding the confound with ability remains. An outcome measure may be affected by ability and unrelated to beliefs except for how beliefs relate to ability. Yet researchers may believe they have accounted for the role of ability and find a relationship with a measure of overconfidence, when all it is reflecting is how the measure of overconfidence is confounded with ability.