Widely-Used Measures of Overconfidence Are Confounded With Ability

Stephen A. Spiller

UCLA Anderson School of Management

March 5, 2024

Author Note

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

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

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

Abstract

The overconfidence concept is one of the great success stories of psychological research, influencing 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 typical measures of overconfidence. First, performance is an imperfect measure of ability; accounting for performance does not sufficiently account for ability. Second, self-evaluations of performance should reflect ability in addition to performance; because performance is ambiguous, people should use 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. I close with recommendations on approaches to recognize and reduce the extent of the problem. This model serves as a stark reminder: when researchers propose that differences in overconfidence are associated with other behaviors, beliefs, or evaluations, they must account for the possibility that differences in ability provide a sufficient explanation.

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

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. She is entitled to acknowledge it. Does this make her overconfident? No. As baseball pitcher Dizzy Dean allegedly said: "It ain't braggin' if you really done it."

Yet 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 the expected value is very close to 17. Through a dogged 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 hard-earned 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. 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 exists and that it 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.[1]

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

[1] Cf., Summers, "Finance and Idiots," as cited in Fox (2009).

& 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 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. The construct of interest, latent overconfidence, refers to latent beliefs about one's own ability (or skill or knowledge) on some dimension that exceed one's true latent ability (or skill or knowledge) on that dimension.[2] I focus on research in overconfidence; there are direct links to the literature on the correlates of positive-self views and self-enhancement as well (Taylor & Brown, 1988; Colvin et al., 1995).

Unfortunately, widely-used methods that are used to assess individual differences in overconfidence confound differences in biased beliefs with differences in actual ability. Although researchers intend to control for latent ability, they instead control for observed performance. The 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, their evidence may instead indicate that ability is associated with those correlates. This confound frequently escapes notice. This may be because when there is an objective performance task and individuals evaluate their performance on that task, it is not obvious (though it is indeed the case) that evaluations will be directly colored by ability. This confound can be particularly pernicious because ability is often explicitly considered and ruled out as an alternative explanation of the results based on how the overconfidence measure is constructed.

[2] If one considers manifest overconfidence (at the level of performance rather than ability), the analogous definition includes both *overestimation*, self-evaluations that overstate absolute performance, and *overplacement*, self-evaluations that overstate relative performance (Moore & Healy, 2008).

I begin by describing a typical paradigm used to assess differences among people in overconfidence, variations on that theme, and why this results in a problem. I next present a mathematical model to formalize and quantify this bias. I 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 with recommendations to recognize and ameliorate the problem, even if eliminating the possibility of the problem in all its forms may be an unattainable target.

Theme and Variations: Measuring Differences in Overconfidence

Research on individual differences in overconfidence has used a dizzying array of measures. I consider cases in which there is a reality criterion against which to compare. When using a single measure of performance and a single self-evaluation, there are at least 20 different ways that overconfidence may be measured, excluding cases in which self-evaluations reflect future expectations. Each 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.[3] 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 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.[4] (I discuss prior critiques of residual and difference scores after further explicating the current problem.)

Self-Evaluate Using the Same vs. Different Metric

The self-evaluation may be assessed in the same metric 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-

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

[4] 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 precisely equivalent to that on self-evaluation when controlling for performance.

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, whether or not the variables have been standardized.

Self-Evaluate Performance vs. Ability

Participants may be asked to evaluate their performance or their ability. The cases above represent self-evaluations of task-specific performance. In other cases, the self-evaluation may be an evaluation of ability rather than an evaluation of performance. For example, after completing a 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 generally about financial matters on a 1 to 7 scale (ability). Researchers residualize this measure of self-evaluated ability on performance (or control for performance in multiple regression) to consider the role of subjective confidence. Although typical examples of self-evaluated ability tend to be in a different metric, it need not be. For example, in principle 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 Evaluation

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 within some specified sample. Self-evaluations of relative standing are often measures of performance, but could instead be measures of evaluations of ability.[5]

Variations on a Theme

These variations may be assembled in any combination as long as it does not involve taking a difference between two measures in different metrics. Evaluations may also be assessed item-by-item to enable assessments of sensitivity or efficiency (e.g., Burson et al., 2006; Fleming & Lau, 2014; Stankov & Crawford 1996). As detailed next, each of these approaches results in a measure of overconfidence that

[5] 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.

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?

I first provide an informal intuitive description of the problem and describe its relation to prior critiques; I then provide a mathematical proof and simulation results. The problem arises from four properties of these performance and self-evaluation measures. First, people typically differ in ability. Second, they typically have at least partial insight into their ability. Third, performance is typically an imperfect measure of ability: it includes some noise and is unlikely to fully and only reflect the construct it is intended to measure. Fourth, performance is typically ambiguous to the target individual: if people were able to unambiguously assess their performance, there would be little potential to show earnest overestimation.

Because performance is ambiguous, self-evaluations of performance ought to regress towards ability under the weak assumption that self-evaluations of ability are correlated with true ability. If two quiz-takers who have some insight into their own ability 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, estimates should be regressive toward people's prior beliefs about their own knowledge. If people have some insight into their own ability, this leads their self-evaluations to be regressive toward their own actual knowledge. Given that self-evaluations are thus a weighted average of knowledge (or ability or skill) and performance, observing self-evaluations exceed performance signals 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; Cohen et al., 2003;

Culpepper & Aguinis, 2011; Gillen et al., 2019; Kahneman, 1965; Westfall & Yarkoni, 2016). Because self-evaluations of performance are regressive towards ability, controlling for performance will leave variation in self-evaluation that is attributable to ability. Although the residuals are uncorrelated with performance by construction, they are still correlated with true ability. So 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 directly assessing ability rather than 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 introduces error and 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 applies 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 again a 13-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 11.6 answers correct a priori, actually got 10 correct (8 of the 9 financial literacy questions and 2 of the 4 trust questions), and, due to the inherent ambiguity, reported that they got 11 correct. A less-literate but more-trusting participant expected to get 5.8 answers correct a priori, actually got 8 correct (4 of the 9 financial literacy questions and all 4 of the trust questions), and so, due to the inherent ambiguity, reported that they got 7 correct. The apparent overconfidence of the first participant and underconfidence of the second participant reflect

true differences in financial literacy, not a surplus nor deficit of confidence.[6] Returning again to the opening example, the student may have decided to only use matches on which he had the most-granular data and therefore 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 adequate coverage of its intended construct (Nunnally & Bernstein, 1994).[7]

A visual depiction of the problem is given in Figure 1. Panel A shows the relationship between skill and performance for 20 individuals. In this example, the performance measure is both noisy and regressive; the 45-degree line is given by the solid line whereas the best-fit regression line is given as the dashed line. The five highest-skilled individuals are depicted as filled circles and the five lowest-skilled individuals are depicted as open circles; the middle ten individuals are depicted as crossed circles.

Panel B depicts self-evaluations as a function of performance, where self-evaluations are regressive toward skill. Both measures of overconfidence are positively confounded with skill ($r_{skill,residual}$ = 0.55, $r_{skill,difference}$ = 0.70). The 9 individuals classified as overconfident by the residual score (i.e., the difference between each point and the best-fit line) and the 8 individuals classified as overconfident by the difference score (i.e., the difference between each point and the 45-degree line) included all 5 of the 5 most-skilled individuals and none of the 5 least-skilled individuals. Conversely, the 11 individuals classified as underconfident by the residual score and the 12 individuals classified as underconfident by the difference score included all 5 of the 5 least-skilled individuals and none of the 5 most-skilled individuals.

Panels C and D plot the correspondence between the residual measure (C) and difference score (D) with an arbitrary correlate of skill. As is evident in this example, these correlates of skill are

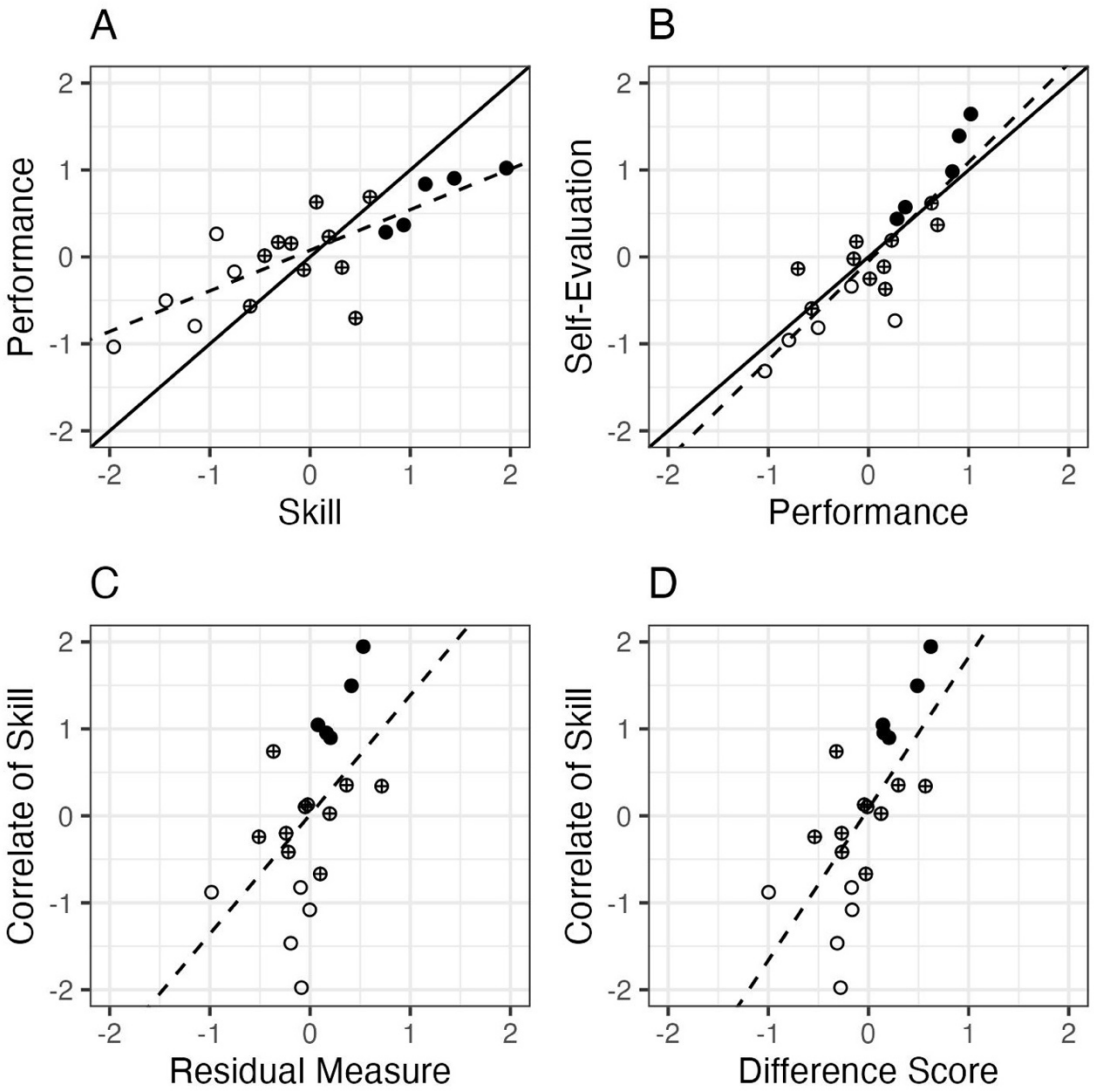
[6] It would be inappropriate to place the 'blame' on the participant: the fact that the participant uses the 'wrong' prior (e.g., financial literacy rather than a linear combination of financial literacy and trust) should not be interpreted as overconfidence if they rely on the very construct the researchers themselves believe they are measuring. I return to this point at the end of paper.

[7] While this example of only sampling matches since 2018 may seem particularly egregious, the problem is quite general: tests are on some topic, but the items necessarily cover only a portion of it, and those items will reliably but idiosyncratically favor some individuals over others.

positively correlated with both measures of overconfidence, despite the fact that both measures account for performance and in this example there is no true overconfidence. As will be derived later, the problem in (C) arises from the simulated measurement error in performance whereas the problem in (D) arises from the simulated regression to the mean in performance.

Figure 1

Visual Depiction of How Measures of Overconfidence Are Confounded With Skill



Note. The five highest-skilled individuals are depicted as filled circles. The five lowest-skilled individuals are depicted as open circles. The remaining ten individuals are depicted as crossed circles. Solid lines depict 45-degree lines; dashed lines depict best-fit regression lines. The parameters used for this example are $\lambda = .4$, $\sigma_v^2 = .16$, $\alpha = .5$, $\sigma_v^2 = .04$.

Because people's private information regarding their own ability affects their evaluations of their own performance but not of others' performance, this same logic leads to the same confound for overplacement. 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. This is attributable to the direct effect of ability on self-evaluations. That is, it is a problem with the position that self-evaluation just measures performance rather than a problem with the use of residual or difference scores per se.

A Note Regarding the Assumption of Accurate Beliefs

A consequence of this line of argument is that researchers will find evidence for correlates of overconfidence even if (a) overconfidence is unrelated to the correlate in question, or, surprisingly, (b) there simply is no overconfidence. The potential to find correlates of overconfidence even when overconfidence does not exist is particularly troubling. For this reason, I base much of my analysis below on the assumption that overconfidence does not exist; that is, I assume that people have perfect knowledge regarding their own skill.

This assumption is almost assuredly wrong. But if the methods we use to assess the presence of overconfidence and its correspondence with other constructs find such evidence even in its absence, we must rethink how we use those results. That is, adopting this assumption enables us to see that our methods reject the null hypothesis of no relationship even when the null hypothesis is known to be true. To relax this assumption, I extend the model in the Appendix to characterize cases in which there is incidental overconfidence. That is, self-evaluations of one's own ability differ from one's own actual ability such that some people overestimate their own ability and others underestimate it, but such self-evaluations remain unrelated to correlates of interest. In such cases, the magnitude and potentially the sign of the confound can vary from those demonstrated under the base case of accurate beliefs, but the problem remains.

Relation to Prior Critiques

The literature on discrepancy scores, including in the context of self-evaluations of performance, is rife with critiques, rebuttals, counterarguments, and comments, yet there remains room to contribute a new perspective. Here I provide a brief discussion of some concerns regarding the distinction between measured and true scores, problems with difference scores, and artifactual accounts of overconfidence and the unskilled-and-unaware effect. A complete characterization of all arguments is out of scope, but a brief discussion helps to contextualize the contribution of the present research.

True vs. Measured Scores

Measurement error can bias coefficient estimates of other variables. Concerns regarding inappropriate inferences about true scores when relying on measured scores are an old problem in measurement (e.g., Birnbaum & Mellers, 1979; Cochran, 1968; Cohen et al., 2003; Cronbach & Furby, 1970; Kahneman, 1965; Lord, 1956, 1958, 1960; McNemar, 1958; Rogosa et al., 1982; Thomson, 1924). This has led to an array of possible approaches to attempt to recover unbiased coefficient estimates (e.g., Cronbach & Furby, 1970; Culpepper & Aguinis, 2011; Kline, 2005; Fuller, 1987). Early discussions considered implications for change scores on tests, recognizing the measurement error inherent in such tests. Why has such a critical determinant not been central in recent discussions of measures of overconfidence? A key contributing factor may be that discrepancy scores are uniquely susceptible to an illusion that the performance measure itself is truly what matters. This is because the performance measure itself is often the target of self-evaluation, diminishing the apparent relevance of latent ability. This diminishment is an illusion. Moreover, unlike in the traditional case, here both the measure (performance) and the true score (ability) are of interest, rather than just the measure (as is typically implicitly implied) or just the true score (as in the traditional treatment of measurement error).

Confounds with Difference Scores

A second overlapping set of critiques have addressed the fact that difference scores are confounded with their component measures: what appears to be a function of the difference may instead reflect a property of one of the components (e.g., Cronbach & Furby 1970; Cohen et al., 2003; Edwards &

Parry 1993; Griffin et al., 1999; Johns 1981; Wall & Payne 1973; Zuckerman & Knee 1996). Response Surface Analysis via polynomial regression (e.g., Edwards 1994; Barranti et al., 2017; Humberg, Dufner, et al. 2019; Humberg, Nestler, & Back 2019) and Condition-based Regression Analysis (e.g., Humberg et al., 2018a, 2019) seek to establish alternative conditions to establish whether a discrepancy vs. a positive self-evaluation is the "active ingredient." The concern I raise regards a confound of self-evaluations with ability, and so is relevant whether one is interested in the discrepancy score or positive self-evaluation. In addition to other concerns regarding condition-based regression analysis in their standard form (Krueger et al., 2017; Fiedler, 2021; cf. Humberg et al., 2018b, 2022), these regression-based approaches do not distinguish between performance and ability, and so are equally susceptible to the concerns I raise here. Thus, the problem I identify is in addition to those previously discussed with respect to change scores. To emphasize there is a distinction, note the typical concern regarding change scores is a *negative* confound between discrepancy scores and baseline. The problem I identify is a *positive* confound between discrepancy scores and baseline.

Overconfidence and the Dunning-Kruger Effect

Within the overconfidence literature, consideration of the role of error in the use of discrepancy scores and imperfect sampling are a repeated theme (e.g., Burson et al., 2006; Erev et al., 1994; Gigerenzer et al., 1991; Juslin, 1993; Klayman et al., 1999). Yet the focus of these critiques has been imperfect calibration and findings regarding aggregate overconfidence rather than the implications for individual-level measures of overconfidence described here.

The present work also follows a longstanding research dialogue and set of critiques regarding the Better-Than-Average effect (e.g., Svenson 1981) and, more recently, whether people who are unskilled are unaware (sometimes referred to as the "Dunning-Kruger Effect," DKE; Kruger & Dunning 1999).

Benoît and Dubra (2011) prove that apparent overconfidence in the aggregate such as a Better-Than-Average effect can come about through Bayesian reasoning regarding a distribution of beliefs. The current work differs in three important ways. First, they consider aggregate levels of overconfidence whereas I consider measured differences in overconfidence, whether or not there is overall

overconfidence; my concern persists even in cases where their concern is ameliorated. Second, their model is based on updating beliefs about one's ability based on one's performance; my model is based on updating beliefs about one's performance based on one's ability. Third, their finding is a function of assessing estimates from distributions which may not aggregate. The model I propose would still predict confounds regarding the use of measures of overconfidence to assess differences across individuals in certain experiments they propose based on mean vs. median assessments. (In a follow-up paper, Benoît et al., 2015, find evidence of aggregate overconfidence in experiments that account for this critique.)

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 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 overconfidence is confounded with ability.

Recent research using alternative approaches further supports the argument that the unskilled are indeed unaware. Feld et al. (2017) 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.

Quantifying the 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, so I adapt their notation where possible.[8] People differ in ability or skill, S_i, distributed with 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 likely to be accurate, but because it presents an important null model to consider: Is there apparent evidence of correlation with overconfidence even when overconfidence does not exist? 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 have at least partial insight into their own abilities. Partial but incomplete insight into their own skill can be modeled as $0 < \rho_{\tilde{S}S} < 1$. In such a case, the residual and difference measures will still be biased, but the magnitude and potentially the sign of the bias will differ; this extension is presented in the Appendix.

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:

$$P_i = \lambda S_i + v_i \quad (1)$$

[8] 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.

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 an unrelated construct) has $\lambda = 0$. Consider a researcher measuring individual differences in intelligence using either (a) a test consisting of three of Raven's progressive matrices, or (b) a 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 $\lambda = 0$.

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 proper Bayesian updaters). For Moore and Healy, these prior beliefs represented beliefs about the simplicity of the task; in the current model, these prior beliefs represent beliefs about one's own Skill, $\tilde{S}_i$. In other words, the key extension here is the relevant and systematic variability in those prior beliefs. As a result, people ought to evaluate their own performance as lying somewhere between their prior beliefs and their true performance[9], plus noise, where the weight on prior beliefs increases with ambiguity. So:

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

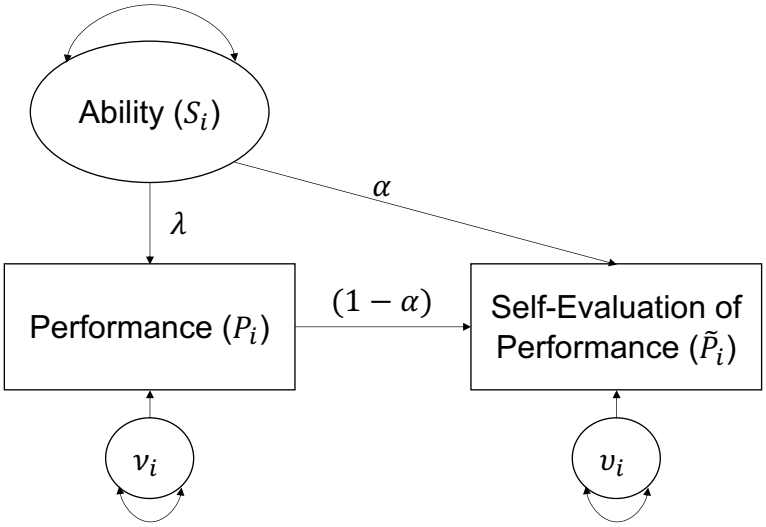
where υ_i has a mean of 0 and variance of $\sigma^2_{\upsilon_i}$. α between 0 and 1 represents the ambiguity of someone assessing their own performance. As ambiguity increases, α gets closer to 1, and self-evaluations of performance reflect their self-evaluations of skill to a greater extent. When self-evaluation 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 2.

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

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, so α and $(1-\alpha)$ may need to be adjusted by a scaling parameter, θ.

Figure 2

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. An extension allowing for $0 < \rho_{\tilde{S}S} < 1$ is presented in Figure A1 in the Appendix.

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

Residual and Multiple Regression Approaches

To calculate overconfidence via residuals, self-evaluations are regressed on performance:

$$\tilde{P}_i = \gamma P_i + \epsilon_i \quad (3)$$

The residuals, $e_i = \hat{\epsilon}_i$, are kept as the measure of overconfidence.[10] Because (a) performance is noisy, and (b) self-evaluations incorporate priors on ability, the expected errors (and thus the residuals) vary with ability (derivation in the Appendix):

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

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

For sufficiently large samples such that $e_i \cong \epsilon_i$, 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 first is simply that there is error not attributable to the construct in the performance measure ($\sigma_v^2 > 0$, making $\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_{v_i}^2}\right) > 0$). The absence of measurement error is the exception, not the rule, so this condition is likely to be met. 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; Cohen et al., 2003; Culpepper & Aguinis 2011; Gillen et al., 2019; Kahneman, 1965; Westfall & Yarkoni, 2016), why does the current paradigm deserve special consideration? First, unlike subjective responses to 7-point scales or preferences as measured by intertemporal choice or risk preference tasks, performance measures contain a veneer of precision and objectivity that may wrongly evoke less concern regarding its status as a noisy measure of ability. Second, without the extension of Moore and Healy's (2008) model, it is not transparent to all researchers that self-evaluations themselves are confounded with ability; this grants a false sense of security regarding the impact of any measurement error in performance.

Difference Score Approach

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

$$\Delta_i = \tilde{P}_i - P_i \quad (5)$$

In expectation, this difference score is also a function of skill (derivation in the Appendix):

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

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 ($\lambda < 1$), and second, if self-evaluation is related to skill conditional on performance ($\alpha > 0$), not just via performance. (In this stylized case, measurement error (σ_v^2) does not affect the bias. However, in practice, P_i often has an upper and lower bound such that $\sigma_v^2 > 0$ would drive effective λ down.)

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

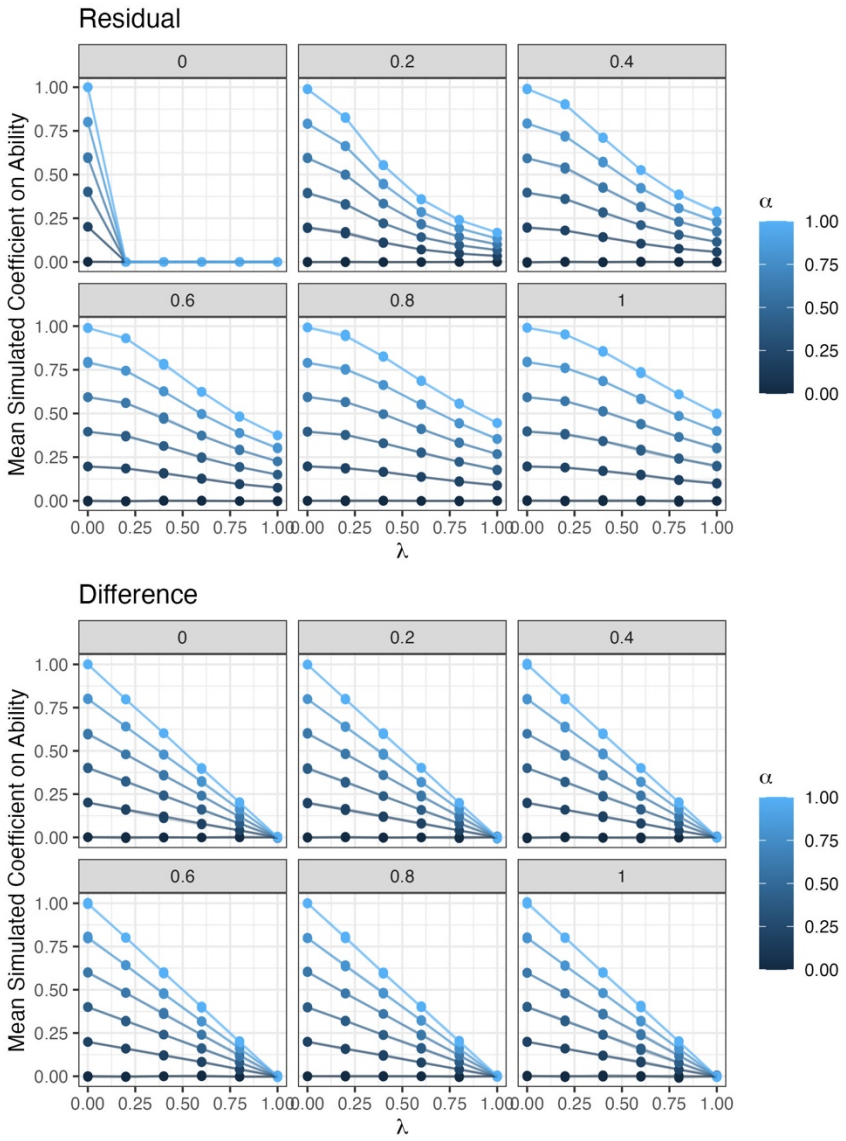
Simulations show that these asymptotic results hold for reasonable sample sizes.[11] For each of 1,296 combinations of parameter values (i.e., all factorial combinations of each of λ, α, σ_v^2, and σ_v^2 taking a value in [0, 0.2, 0.4, 0.6, 0.8, 1.0]), I simulated 1,000 samples of 100 observations each. In each sample, skill was drawn from a standard normal distribution, and error terms were drawn from standard normal distributions scaled by the corresponding variance. Performance and self-evaluations followed from the model. Simulation results are depicted in Figure 3. Figure A2 in the Appendix presents simulation results for correlated inaccurate beliefs.

[11] All code is available at https://researchbox.org/1597&PEER_REVIEW_passcode=ORRDVP.

As shown above, this confound matters in theory. Does it matter in practice? Reanalysis of two datasets indicates a resounding yes.

Figure 3

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 the variance of performance is 0, so the coefficient on performance predicting self-evaluations is not estimable, and so the residuals are merely mean-centered self-evaluations.

Empirical Application I: Overconfidence Predicts Performance

The analysis above indicates that overconfidence measures are biased under appropriate conditions. Is this bias likely to distort inferences? The answer 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 which is a priori likely to be related to skill and unrelated to overconfidence.

Specifically, using data from Moore and Healy (2008),[12] I consider a case where the outcome is a future measure of performance. 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 confounded with skill.

Overconfidence Paradigm

Moore and Healy (2008) collected data from 82 college undergraduates on many measures. I describe the relevant components here and refer the reader to Moore and Healy for 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 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

[12] Moore and Healy do not make the inferential error common in the literature. Rather, the availability (https://osf.io/6tecy/) and richness of their data present a useful opportunity to examine whether the error can affect real inferences.

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: there is a 39% chance that a hard quiz is followed by an easy quiz but only an 11% chance that a hard quiz is followed by another hard quiz. This mechanically generates a negative correlation between performance on one quiz and performance on the subsequent quiz. To address this 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 a modest sample size for between-subject analyses by current standards (82 participants). To exploit the within-subject data, I consider performance on sets of sequential quizzes, clustering errors by subject. For example, to examine skill, I regress block performance on prior block performance; each participant contributes five observations: block 2 performance as a function of block 1 performance, block 3 performance as a function of block 2 performance, etc. The analysis accounts for non-independence through clustered errors using the *lm_robust* function from the *estimatr* package (Blair et al., 2022). Alternative approaches to address these concerns using the within-subject design to maximize statistical power are generally consistent.

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 ($b = 0.298$, $SE = 0.141$, $t(38^{13}) = 2.11$, $p = .042$, 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]).

[13] All degrees of freedom throughout this reanalysis of Moore and Healy (2018) are estimated due to clustering.

This coefficient is necessarily equal to that on the residual, but estimated more precisely as performance accounts for additional variance in the dependent variable in the multiple regression analysis but not the residual analysis. 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]).[14]

One might argue that overconfidence truly improves subsequent trivia quiz ability (e.g., via self-efficacy). Most such stories would suggest the correlation is stronger for adjacent blocks. 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 provides a parsimonious explanation: 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) only partial insight into one's own skill; see Appendix.

To examine whether this theoretical account has teeth for the residual and multiple regression analyses, I examine whether four 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 component is self-evident); and (d) Does a proxy for skill predict self-evaluations beyond performance?

[14] 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 ($0 < \rho_{\tilde{s}S} < 1$). See model extension in the Appendix for details.

Are There Differences in Skill? Yes

If performance is correlated across blocks, there is evidence of systematic differences in trivia quiz skill.[15] I regress performance in block *t* on prior performance in block *t*-1, clustering errors by subject. The coefficient on lagged performance was 0.754 (*SE* = 0.045, $t(31) = 16.59$, $p < .001$; 95% CI: [0.662, 0.847]), indicating high performance on one block is strongly associated with high performance on the next block. When an analogous approach was used with block *t*-2, *t*-3, etc., there was no evidence of a relationship that decays with lag (lag 2: *b* = 0.783, *SE* = 0.057; lag 3: *b* = 0.788, *SE* = 0.076; lag 4: *b* = 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 partial insight into how they will perform.[16] 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 =

[15] Skill includes ability, knowledge, and other necessary inputs that remain stable during the course of the study.
[16] 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 no stronger when considering the third quiz (*b* = 0.177, *SE* = 0.041, $t(45) = 4.33$, $p < .001$).

$0.538 < 1$, indicating that it is not the case that both measures are errorless indicators of the same construct. Performance as a measure of skill contains error.

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

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, but I can use subsequent performance as a noisy proxy. Unlike in other cases in which performance is 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; 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]): controlling for current performance, future performance is predictive of current self-evaluations. 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 Placeholder 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 relevant alternative explanation.

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 a purported correlate of overconfidence may instead merely be a correlate of skill: perhaps this whole endeavor is a statistical curiosity 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.

Overview, Data, and Analysis Reproduction

Parker et al. (2012)[17] study the role of "inappropriate confidence" (what Parker & Stone, 2014, later refer to as "unjustified confidence" and most of the literature simply refers to as overconfidence) 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."[18] 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.

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 quiz and confidence using a single 7-point measure

[17] 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 (https://alpdata.rand.org/), is sufficiently clearly written so as to avoid ambiguity, and is important enough to be well-cited.
[18] 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.

assessing people's subjective understanding of economics.[19] 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.[20] 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 of the original. 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, Ns = 1161, 1106, 988, and 584). The only exception is the correlation between Study 3 performance and Study 4 confidence. I find r = 0.37 and the original paper reports r = 0.26.

A Model Where Overconfidence Does Not Matter

I fit these correlations to the model in Figure 4. Importantly, there is no latent confidence in this model at all. Instead, I model the four performance measures as measures of financial knowledge, each confidence measure as a measure of financial knowledge (Study 1) or financial knowledge and performance (Studies 2-4), and financial planning as a consequence of financial knowledge alone.

[19] This confidence measure $\tilde{P}$ was thus a subjective measure of knowledge (S), not performance (P).
[20] 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.

Table 1

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

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_v^2s (one for each study's performance measure), 3 αs and 4 σ_v^2s (one for each study's confidence measure, except α for Study 1 which was fixed to 1 because that confidence measure assessed ability), and 4 θ scaling factors (one for each study's confidence measure).[21] The model was fit using full information maximum

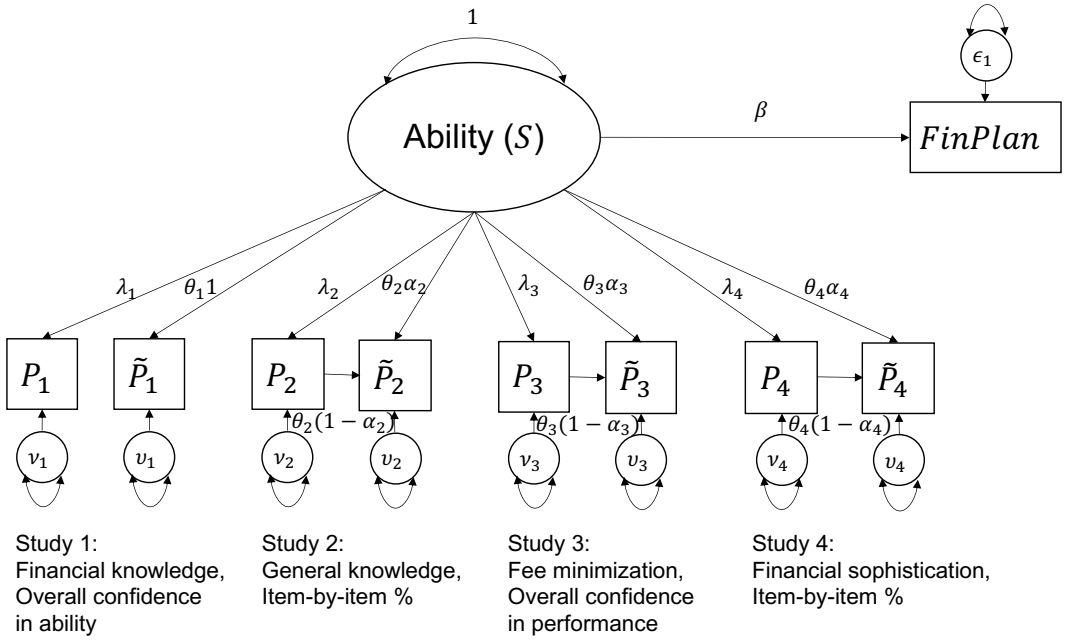
[21] The scaling factors were necessary to account for scale use. To facilitate estimation, rather than estimating θ and α directly, I estimated $\theta\alpha$ and $\theta(1-\alpha)$. θ was then calculated as $\theta\alpha + \theta(1-\alpha)$ and α as $\frac{\theta\alpha}{\theta\alpha+\theta(1-\alpha)}$.

likelihood for missing data using the `lavaan` package v0.6-12 (Rosseel, 2012) in R.[22]

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. Thus there are a priori reasons to expect that the model depicted in Figure 4 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.

Figure 4

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



[22] 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 30 total parameters, leaving 24 degrees of freedom.

Results

Despite these model mismatches, the estimated parameters appear to be 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}	$\sigma_{\epsilon}^{2\,b}$
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		

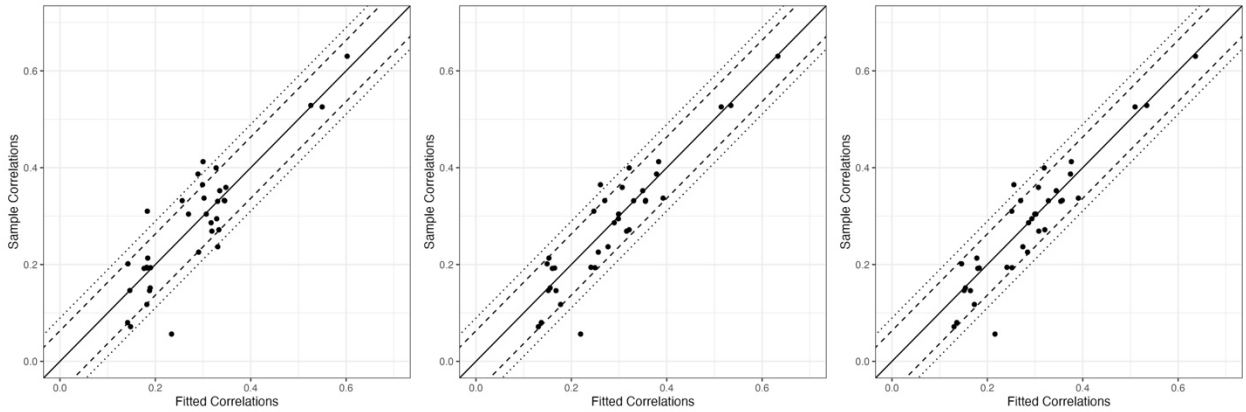
[a] Calculated after rescaling.
[b] Held constant across studies.
[c] Fixed by theory, not estimated.

As shown in Figure 5 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 5

Fitted Correlations and Observed Correlations in the Data in Three Models



Note. The left panel represents the model shown in Figure 4. The center panel allows for the presence of, but no effect of, overconfidence, that is, $\rho_{\bar{S}S} \leq 1$. 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 4 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 allowing $\rho_{\tilde{S}S} \leq 1$. 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 positively but 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 Model 3 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 5).

Taken together, these analyses suggest that a parsimonious representation of the reported data can be derived from a simple model based only on knowledge and without (inappropriate, unjustified, or over-) confidence. Some evidence suggests that the model in Appendix Figure A1 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.

Recommendations

Both in theory and in practice, widely-used measures of overconfidence are problematically confounded with ability. Beyond despair or wishing away the problem, what solutions are available? Although the difference score approach in the first application did not indicate a problem, 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. Furthermore, use of a difference score requires the strong and untestable assumption that the performance measure has a unit loading on the ability construct: even if there were truly no bias in the first application, that would provide no guarantee that there would be no bias in other applications.

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

Use Reliable, Valid Measures

Most importantly, this serves as a call to ensure the use of reliable and valid measures. This recommendation ought to go without saying; after all, no one thinks using an unreliable or invalid measure is a good idea. But given the strong theoretical reasons to believe there ought to be a confound without such measures, it accentuates the importance of using them. This is particularly important given

that performance measures are often moderately reliable at best: Krueger and Mueller (2002) report split-half correlations ranging from 0.17 to 0.56 and Burson et al. (2006) report split-half correlations ranging from -0.24 to 0.52; I note these examples because the data were readily reported, not because they were uniquely low. For the residual version, lack of *reliability* in the performance measure can lead to lack of *validity* in the residual measure. Thus, increasing reliability in the performance measure has the potential to enhance validity of the residual measure.

Using polynomial regression and response surface analysis (e.g., Edwards 1994; Humberg et al. 2019) or condition-based regression analysis (Humberg et al. 2018) are not sufficient by themselves to account for the concern, as neither accounts for measurement error or construct mismatch in its base form. Only in conjunction with a strategy to address measurement error will they address reliability, and even then construct validity remains a concern.

Is $\lambda < 1$ 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 case of a financial literacy scale with items that measure trust instead, is equivalent to construct invalidity, or $\lambda < 1$, which leads to the focal problem for difference scores. (Note that the measure may be highly reliable even with low validity; I return to this point in the discussion of the next recommendation.) A sensible critique is that this is simply a different form of unjustified confidence: people confidently use a prior that should not apply and regress to the wrong belief as a result. I argue we cannot be so quick to attribute such a problem to the participant's updating strategy rather than the researcher's inferential strategy.

Consider again the phrenologist introduced earlier. Both the phrenologist and the participant may earnestly believe that the phrenologist is generating a diagnostic measure of intelligence. If the participant is asked how they perform on this measure of intelligence, but they have substantial ambiguity about their own head measurements ($\alpha = 1$), they will report their 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 skeptic may argue: "That is overconfidence! The participant is regressing their self-evaluation of performance 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 (when the participant earnestly believes the measure is measuring the same construct the researcher earnestly believes it measures), we perhaps ought to be cautious accepting an overconfidence label in the presence of misinformation (when the participant earnestly believes the measure is measuring the construct the researcher tells them it measures; Ehrlinger & Dunning, 2003).

Account for Measurement Error

To provide an unbiased test, it is useful to recall the conditions under which there is no bias for the role of self-evaluation when controlling for performance. 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 concern addresses a classic problem 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. Of course, these approaches only help to the extent that the intended construct is the construct the measure is actually assessing.

Structural Equation Models

Structural equation models (e.g., Kline, 2005) permit the researcher to model relationships among latent variables, unattenuated by measurement error. Indeed, this is the approach taken to model 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. The key assumption is that the common variance underlying the performance measure 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 reflect 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, 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 will be considerably higher than $\frac{\lambda^2}{\lambda^2+\sigma_v^2}$, the relevant quantity, leading us to underestimate the extent of the problem.

A full development of the errors-in-variables approach is beyond the scope of this paper; interested readers are referred to Fuller (1987) for a statistical treatment and Culpepper and Aguinis (2011) for psychology researchers. 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. An adjustment based on the reliability of one predictor may cause the coefficients on other predictors to vary in magnitude or sign. Properly accounting for the measurement error in the performance measure enables the model to control for ability, not just performance, which affects the

coefficient on self-evaluation. The Appendix reports the results from using errors-in-variables methodology in the second empirical application.

These approaches are not a panacea, as they again assume good construct validity. Put simply, if one is able to account for measurement error, one will get an estimate of results using the true score of whatever the measure measures. What the measure measures is not guaranteed to align with the intended construct. Typical indicators of reliability (e.g., test-retest reliability; Cronbach's α) may not be sufficient to determine the unattenuated association using either approach, given potential problems with construct validity. 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. Should one rely on difference scores, one is left with an independent set of concerns (e.g., Cronbach & Furby 1970; Edwards & Parry 1993; Johns 1981).

Bound the Parameter Space

Rather than attempting to rule out this alternative explanation, researchers may instead relax the strength of their claims by acknowledging the conditions under which it may hold. Given the ability to characterize the magnitude of the bias, one can plausibly specify parameter configurations that could and could not account for the observed results. In some cases, there may be parameter configurations which could account for the observed results but are implausible: while they are mathematically plausible, they may be ruled out based on theory.

In other cases, one can rule out the alternative explanation altogether. There are two important cases in which the current proposal regarding overconfidence's confound with ability is 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. But no correlation between *performance* and the outcome measure of interest is not sufficient: such a lack of correspondence could merely indicate that performance is a poor measure of ability even if it is a reliable measure of something else. This would again lead to a biased estimate of the effect of overconfidence.

Second, if the relationship between ability and the outcome measure of interest and the relationship between residualized overconfidence and the outcome measure of interest have opposite signs, the core bias described here could not account for such a pattern of results. 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. (Note this is not guaranteed to hold for the difference score if beliefs are imperfectly correlated with ability due to the patterns displayed in Figure A2.)

Of course, in establishing those bounds, it is important to consider the uncertainty regarding one's estimate, not merely the point estimate itself. Further, these bounds are with respect to this null model. Other null models (e.g., one in which an unskilled-and-unaware effect holds but skill is the only correlate of behavior) may not be so readily ruled out.

Use Alternative Measures

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

Similarly, Lawson et al. (2023) and Binnendyk and Pennycook (2023) have each introduced measures of individual differences in general overconfidence. In the first case, these are based on expected performance on a task for which there is no diagnostic information to go on. One interpretation of this measure in terms of the current model is that beliefs about skill ought to be unrelated to actual skill at the target task. In the second, these are based on estimated performance on a task for which

performance is at chance and difficult to ascertain. One interpretation of this measure in terms of the current model is that skill at this task ought to be unrelated to other correlates of interest. As with the Paulhus et al. (2003) measure, there is reason to be more-optimistic regarding these tasks, and to potentially prefer them over the other methods described here, but they are unable to completely address the problems laid out here. To the extent some people accurately believe themselves to be more generally successful at a variety of tasks than others, the same problems will persist. It is possible that the plausible range of parameters may lead to smaller biases in such cases and so be of negligible concern.

Summary and Conclusion

Research and casual observation suggest that overconfidence is prevalent and varies across people. 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. Given any ambiguity regarding performance, measures of confidence ought to regress towards prior beliefs about ability even when they are intended to be self-evaluations of task performance. Because performance itself is an imperfect measure, the variance of self-evaluation that is attributable to ability is not fully partialed out. The result is that both residual and difference overconfidence measures are confounded with ability. In an idealized model, this bias can be quantified.

These confounds imply that it is possible to observe surprising results in the data: overconfidence predicts subsequent performance even 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 researchers are willing to make strong assumptions regarding construct validity and estimate or assume reliability of each measure, it is possible to address these concerns through structural equation modeling or error-in-variables adjustments. However, these partial solutions are not an automatic panacea, as a number of complications may arise regarding construct validity and unstable estimates. Instead, design-based solutions (e.g., experimental manipulations or using other

measurement approaches) or accepting alternative interpretations of the results (i.e., plausible parameter configurations) may ultimately prove necessary. This work may serve as an impetus and 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.

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

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

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

Binnendyk, J., & Pennycook, G. (2023, August 23). Individual differences in overconfidence: A new measurement approach. https://doi.org/10.31234/osf.io/ugb3s

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

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

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

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

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

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

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

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

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

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

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

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

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

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

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

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

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

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. (1987). *Measurement Error Models*. John Wiley & Sons.

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

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

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

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

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., Küfner, A. C., van Zalk, M. H., Denissen, J. J. A., Nestler, S. & Back, M. D. (2019). Is accurate, positive, or inflated self-perception most advantageous for psychological adjustment? A competitive test of key hypotheses. *Journal of Personality and Social Psychology*, *116*(5), 835-859.

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

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., van Zalk, M. H., Denissen, J. J. A., Nestler, S., & Back, M. D. (2018b). Why Condition-Based Regression Analysis (CRA) is

Indeed a Valid Test of Self-Enhancement Effects: A Response to Krueger et al. *Collabra: Psychology*, *4*(1), 26.

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

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

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

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

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

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

Kahneman, D. (1965). Control of spurious association and the reliability of the controlled variable. *Psychological Bulletin*, *64*(5), 326-329.

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

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

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

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

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

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

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

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

Lawson, A., Larrick, R. P., & Soll, J. B. (2023). Forms of Overconfidence: Reconciling Divergent Levels with Consistent Individual Differences. *Available at SSRN 4558486*.

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

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

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

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

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

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

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

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

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

Moore, D. A., & 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

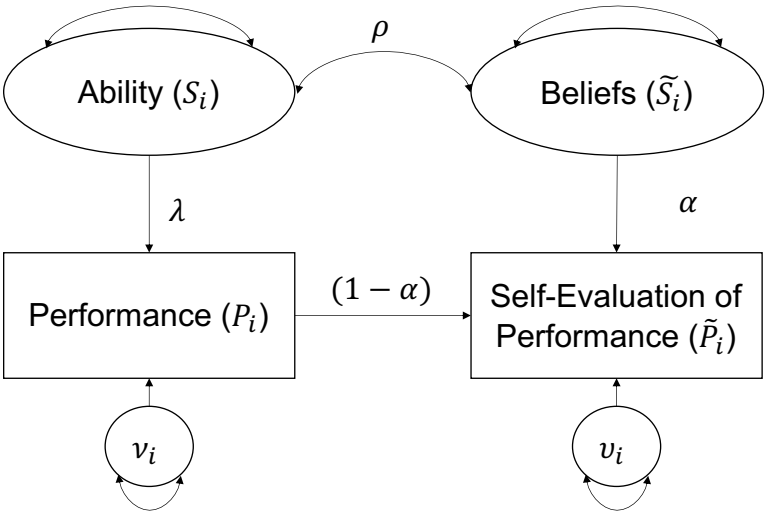
Appendix

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 problem 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 2 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:

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

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

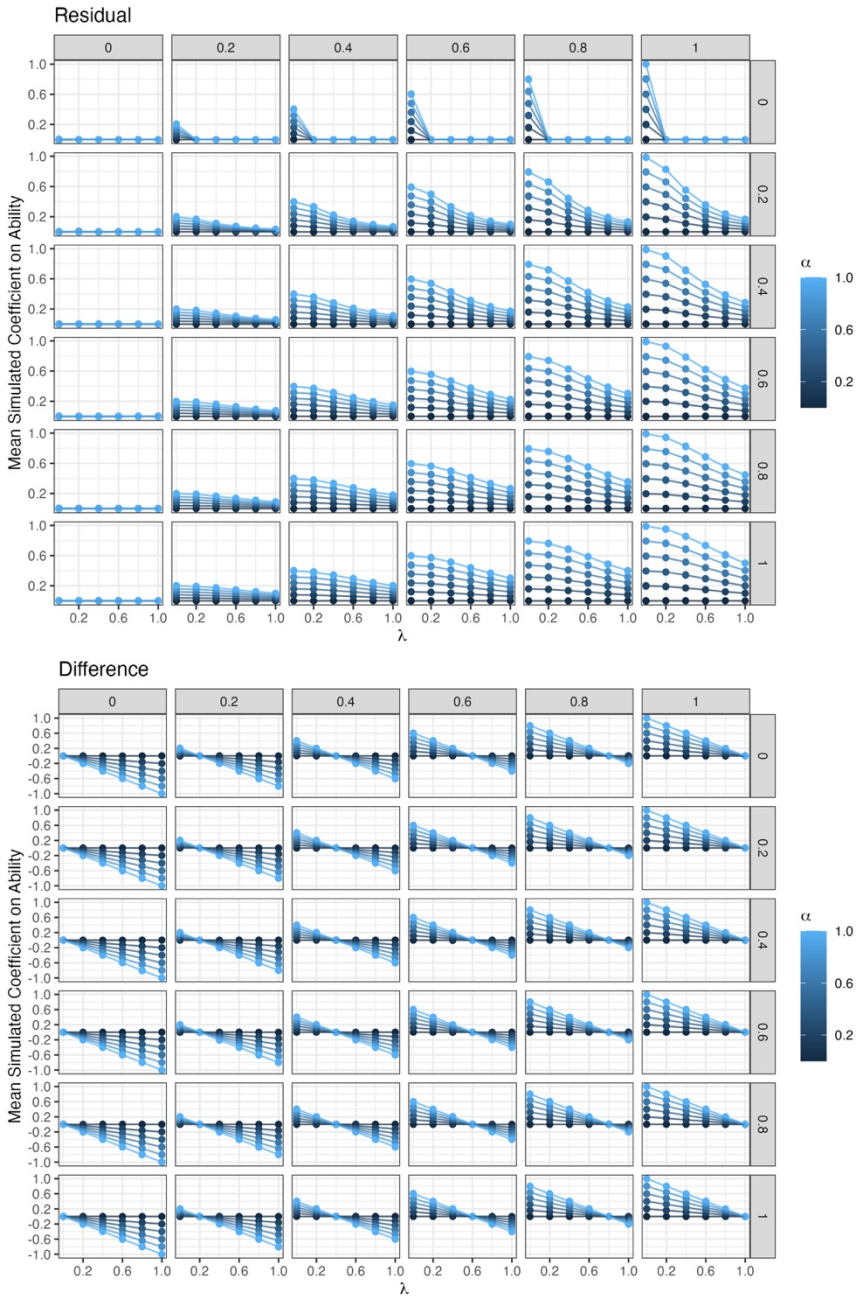
The expectation of the difference score likewise now depends on ρ:

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

And again, if beliefs are equal to ability, then $\rho = 1$, reducing 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. As in Figure 3, for the residual score panel, when $\sigma_v^2 = \lambda = 0$, the coefficient is not 0. This is because the variance of performance is 0, so the coefficient on performance predicting self-evaluations is not estimable, such that the residuals are merely mean-centered self-evaluations

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 $(\rho - \lambda)$, aligning with prior critiques of difference scores discussed in the main text.

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 the presence of overconfidence: if beliefs are imperfectly correlated with skill, then 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.

Derivation of Equations 4 and 6: Confounded Residuals and Difference Scores

Equations (1) through (6) are repeated here as (A1) through (A6) for ease of reference.

$$P_i = \lambda S_i + \nu_i \tag{A1}$$

$$\tilde{P}_i = \alpha\widetilde{S_i} + (1-\alpha)P_i + \upsilon_i \quad \text{(A2)}$$

$$\tilde{P}_i = \gamma P_i + \epsilon_i \quad \text{(A3)}$$

$$E[\epsilon|S] = \left(1 - \frac{\lambda^2}{\lambda^2+\sigma_v^2}\right)\alpha S \quad \text{(A4)}$$

$$\Delta_i = \tilde{P}_i - P_i \quad \text{(A5)}$$

$$E[\Delta|S] = (1-\lambda)\alpha S \quad \text{(A6)}$$

Plugging (A1) into (A2) enables us to rewrite self-evaluations of performance, $\tilde{P}$, in terms of Skill, *S*, in (A7). Given the assumption of accurate beliefs, we can then replace $\tilde{S}_i$ with S_i and rearrange terms to give us (A8):

$$\tilde{P}_i = \alpha\widetilde{S_i} + (1-\alpha)(\lambda S_i + \nu_i) + \upsilon_i \quad \text{(A7)}$$

$$\tilde{P}_i = \lambda S_i - \alpha\lambda S_i + \alpha S_i + \nu_i - \alpha\nu_i + \upsilon_i \quad \text{(A8)}$$

We then use (A3) to rewrite γ in terms of the structural parameters α, λ, and σ_v^2 to solve for ϵ, which the residual will closely approximate for sufficiently large samples. To begin, we decompose γ into two portions: that which relates $\tilde{P}_i$ to *P* directly and independent of *S* (i.e., $(1-\alpha)$), and that which relates $\tilde{P}_i$ to *P* as they each relate to *S*, given by $\lambda\frac{\sigma_S^2}{\sigma_P^2}\alpha$. Although the typical causal interpretation does not align, this logic precisely follows the logic of decomposing a total effect into a direct effect and indirect effect in statistical mediation. We assume for convenience that $\sigma_S^2 = 1$ and reexpress $\sigma_P^2 = \lambda^2 + \sigma_v^2$. This gives us:

$$\tilde{P}_i = \left(1 - \alpha + \lambda\left(\frac{1}{\lambda^2+\sigma_v^2}\right)\alpha\right)P_i + \epsilon_i \quad \text{(A9)}$$

We then replace $\tilde{P}_i$ in (A9) via (A8) and P_i in (A9) via (A1), and simplify and isolate ϵ_i:

$$\epsilon_i = \left(1 - \frac{\lambda^2}{\lambda^2+\sigma_v^2}\right)\alpha S_i + \upsilon_i - \nu_i\lambda\left(\frac{1}{\lambda^2+\sigma_v^2}\right)\alpha \quad \text{(A10)}$$

As ν_i and υ_i are independent and mean 0, they drop out in expectation, providing (A4).

To derive the expected value of the difference score, we use (A1) to reexpress P_i in (A5) terms of S_i and we use (A8) to express $\tilde{P}_i$ in (A5) terms of S_i. Simplifying give us:

$$\Delta_i = (1 - \lambda)\alpha S_i - \alpha v_i + \upsilon_i \quad \text{(A11)}$$

Once again, because v_i and υ_i are independent and mean 0, they drop out in expectation, leaving us with (A6).

An Empirical Application of the Errors-in-Variables Approach

To examine the potential of using the errors-in-variables approach, I use 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. It is important to note that standard measures of reliability (e.g., test-retest; internal reliability given by Cronbach's α) may be optimistic indicators of how reliably the measure measures its *intended* construct. For example, other irrelevant stable constructs that the measure assesses may inflate reliability.

I apply this approach to the American Life Panel application from Parker et al. (2013).[23] Despite the potential concerns noted above, I rely on the reported Cronbach's α where available to 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 A1. All results using *eiv* were quantitatively similar and led to identical statistical conclusions.

[23] I also attempted to use this approach on Moore and Healy's (2008) data. Performance and estimates were extremely strongly correlated across participants within blocks (from 0.87 to 0.96), implying extremely high reliabilities that are inconsistent with other approaches to estimating reliability (e.g., the correlation between performance and lagged performance). This may be attributable to the randomization approach that led to different participants encountering different sets of quizzes in different blocks. Assuming only minimally unreliable measures for both performance and self-evaluations (reliabilities of 0.95 for each), using *eivreg* reveals that lagged performance predicts current performance (b = 0.60, SE = 0.24, $t(407) = 2.47$, $p = .014$) but 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 a small amount of unreliability. However, the results are unstable given even slight differences in estimated reliabilities.

Table A1

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.

Overall these results tell a story that is inconsistent 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. 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. The reliabilities assessed via Cronbach's α may be larger than the proper adjustment would require.