Supplement to "Widely-Used Measures of Overconfidence Are Confounded With Ability"

Stephen A. Spiller

Frequency of Usage of Different Measures

How is overconfidence measured by researchers in the literature? The main text addresses both difference scores and residual (or equivalent multiple-regression) based measures. But if one of those measures was used only rarely, the corresponding argument may not be widely applicable. I systematically (but not necessarily representatively) construct a sample to examine how researchers assess overconfidence. I collapse across overestimation and overplacement and do not consider overprecision.

On July 17, 2024, I queried Google Scholar for articles since 2000 using: source:"[journal name]" ("measure overconfidence" OR "measuring overconfidence" OR "measure of overconfidence" OR "measured overconfidence") ("correlate with" OR "correlates with" OR "correlated with"). The next day I repeated the searches, replacing "overconfidence" with "overestimation" and "overplacement." This search was intentionally restrictive. The term "overconfidence" matches more than 75,000 results in Google Scholar since 2000. Including variations of "measure" increased the likelihood that overconfidence appeared in the analysis itself (vs. just introduction). Including variations on "correlated" increased the likelihood that articles did not exclusively consider overconfidence as an outcome of an experiment. Limiting to articles since 2000 reduced the likelihood I primarily characterized an approach that has long-since been abandoned. This approach surely leads to the exclusion of relevant articles.

The journals were: *Journal of Personality and Social Psychology*; *Journal of Experimental Psychology: General*; *Psychological Science*; *Psychological Review*; *Psychological Bulletin*; *Cognition*; *Journal of Applied Psychology*; *Management Science*; *Journal of Behavioral Decision Making*; *Organizational Behavior and Human Decision Processes*; *Judgment and Decision Making*; *Proceedings of the National Academy of Sciences*; *Nature: Human Behavior*; *Nature*; *Science*; *American Economic Review*; *Quarterly Journal of Economics*; *Review of Economic Studies*; *Journal of Political Economy*; *Econometrica*; *Review of Economics and Statistics*; *Journal of Finance*; *Journal of Financial Economics*; *Review of Financial Studies*[1]; *Journal of Consumer Research*; *Journal of Marketing Research*; *Marketing*

[1] I queried *Review of Financial Studies* on August 22, 2024 due to an inadvertent omission in my initial query.

Science; *Journal of Marketing*; *Quantitative Marketing and Economics*; and *Journal of Consumer Psychology*. This search led to 60 unique papers, of which 31 used a residual or difference score measure or a measure that did not account for performance. Those 31 articles are listed in Table A1.

Table A1

Sample of Journal Articles Since 2000 Using Residual Measure, Difference Score Measure, Both Measures, or Equivalent Measure

Article	Measure	Notes
Agranov & Buyalskaya (2022), *MS*	Difference*	Control for performance; equivalent to residual
Anderson et al. (2017), *JFE*	Difference*	Control for performance; equivalent to residual
Anderson et al. (2012), *JPSP*	Residual	
Anderson & Lu (2017), *MS*	Difference	
Åstebro et al. (2007), *JBDM*	Equivalent	Subtracts sample performance, not individual performance
Avery et al. (2022), *REStat*	Difference	
Belmi et al. (2020), *JPSP*	Both	Analysis prioritizes residual
Cavalan et al. (2023), *JEP:G*	Difference*	Some control for accuracy; equivalent to residual
Chen et al. (2007), *JBDM*	Equivalent	Assumes no role of ability through use of market returns
Dean & Ortoleva (2019), *PNAS*	Difference*	Define overplacement as predicted own – predicted mean; equivalent as assumes no role of ability Some control for performance; equivalent to residual
Drummond Otten & Fischhoff (2020), *JBDM*	Both	
Eyal et al. (2018)	Difference	
Fast et al. (2012)	Equivalent	One study does not adjust for individual performance Another study uses difference
Gillen et al. (2019)	Difference	Focuses on addressing measurement error
Grinblatt & Keloharju (2009), *JF*	Residual	
Hilton et al. (2011), *JBDM*	Equivalent	Defined self-placement as predicted own – predicted mean; so equivalent as assumes no role of ability Also used differences
Huffman et al. (2022), *AER*	Difference	Differences calculated vs. model-predicted-performance
Ke (2021), *JF*	Difference*	Some control for performance, similar to residual
Krawczyk & Wilamowski (2017), *JBDM*	Difference	
Landier & Thesmar (2008), *RFS*	Difference	
Larrick et al. (2007)	Difference	
Liu et al. (2022), *JFE*	Difference*	Some control for performance; equivalent to residual
Lyons et al. (2021), *PNAS*	Both	Analysis prioritizes difference.
Moore & Healy (2008)	Difference	
Prims & Moore (2017)	Difference	
Ren & Croson (2013)	Equivalent	Focus on overprecision; overplacement used as covariate Overplacement measure described as ignoring performance
Reuben et al. (2024), *JF*	Difference*	Some control for performance; equivalent to residual
Sanchez & Dunning (2021), *JPSP*	Difference	
Van Zant (2022)	Residual	
Varma et al. (2023), *JMR*	Equivalent	Does not adjust for individual performance
Walters & Fernbach (2021), *PNAS*	Equivalent	Assumes no role of ability through use of market returns

Note. "Difference*" indicates a publication in which a difference score is used, but at least some analyses control for performance, meaning that the interpretation of the role of the difference score is equivalent to that of the residual score in those analyses. Residual includes cases of multiple regressions examining the role of confidence when controlling for performance as the coefficient is equivalent.

The remaining 29 articles were not directly relevant. Many of the excluded articles were analyses of CEO overconfidence using Malmendier and Tate's (2005, 2008) options-based measures overconfidence; Malmendier and Tate indicate the measures are not assessing private information. Other excluded articles focused on overprecision or developed an analytical model without data.

The 31 included articles are necessarily a limited subset. They only include articles since 2000 at a narrow, albeit influential, subset of journals. Even within those outlets, there are surely manuscripts not captured by the limiting search terms. For example, Parker et al. (2012), detailed below, is not identified by this search. This list is not intended to be exhaustive, but rather to provide a systematic sample of examples of articles that use each measure. Inclusion in this table does not imply a mistaken inference; in some cases, the overconfidence measure is not even of primary interest. In total, of the 31 articles, 28 used a difference score or an equivalent measure (including no adjustment for performance) and 20 used a residual score or an equivalent measure (including difference score controlling for performance or no adjustment for performance). Sometimes multiple measures are used with unequal focus. Due to inherent edge-cases, it is possible that different researchers may come to different conclusions about some of these codes. But this sample suggests both measures are used frequently-enough to be of interest.

Broad Applicability of Model

The main text presents three detailed re-examinations of articles from Table A1. Using Moore and Healy's (2008) data, I find that overconfidence predicts subsequent performance, consistent with the model. Using Anderson et al. (2012) and Belmi et al. (2020), I find that plausible parameter configurations could be consistent with the data arising from correlations with test-taking ability rather than overconfidence. Later in the supplement I also present a detailed re-examination of Parker et al. (2012), again finding plausible parameter configurations that could be consistent with data arising from correlations with ability, not overconfidence. Below I consider a broader set of findings in less detail.

Using the list of articles gathered in Table A1, I aimed to examine whether key findings are compatible with parameter values from the current model. Not all papers were amenable to this analysis. The 16 papers that were amenable and were not included in the main text are included in Table A2 with

potential parameters. The remaining 15 papers are listed in Table A3 with a brief rationale for their exclusion. Those 15 include the 3 papers discussed in detail in the text; near-exclusive use of overconfidence as a dependent variable in an experiment; a primary focus on overprecision, even when overplacement and/or overestimation are included; or a primary focus on the internal structure of overconfidence (e.g., correlations of overconfidence with performance rather than with a third variable). The current discussion is not necessarily irrelevant to all of those excluded papers, but they do not lend themselves to a straightforward analysis of the key model implications.

Using the analyzable subset, I attempted to identify an analysis that aligned with the main argument of the authors' paper (rather than e.g., a covariate in a minor robustness check). When there were multiple such analyses, I focus on one of the primary ones. I considered four types of reported correlations (between the outcome measure and each of residual, difference, performance, and/or self-evaluation) and used whichever were reported (typically 1-2, sometimes 3). In many cases the regression analyses include other covariates, so these often represent partial correlations, not zero-order correlations.

The focal results were sometimes presented as regression coefficients, sometimes as correlation coefficients, sometimes as group means and standard deviations, and sometimes as t-tests without coefficient magnitudes. When the correlation was reported, I rely on the correlation. When the correlation was not reported, I approximated it using eq. 16-17 from Rosenthal (1994). This equation enables calculation of r from Z and N. When N was not explicitly reported, I approximated it. I approximated Z using t or the ratio of the reported regression coefficient and standard error. I used this equation (rather than one using t and degrees of freedom) for simplicity because it was not always clear how many degrees of freedom were included in regression tables with unenumerated controls, and the differences among them will be relatively minor for reasonable sample sizes. Nevertheless, I emphasize these are crude approximations. But given the model's simplifications, little qualitative insight would likely be gained from more-precise inputs. The most likely qualitative issue to arise in these correlation calculations is from analyses using cluster-robust standard errors. In such cases the reported correlations are directionally correct but may be of the wrong magnitude.

After inferring these correlations, I sought a set of parameters that would generated the identified correlation(s). When possible, I constrained $\rho = 1$ to assume perfectly-calibrated beliefs. I considered other values when $\rho = 1$ could not reproduce the finding. I constrained the error variances to lead to unit variances of performance ($\sigma_v^2 = 1 - \lambda^2$), evaluation ($\sigma_v^2 = 1 - \alpha^2 - (1 - \alpha)^2$), and outcome ($\sigma_\epsilon^2 = 1 - \beta^2$).[2] Even when $\rho = 1$, this still leaves three free parameters (λ, α, β), and no case included more than three correlations. The model is underidentified and there were always multiple sets of compatible parameters. Table A2 reports one such set, along with the implied correlations.

Where possible, the reported parameters aim to represent a plausible set that generate the target correlations. But these parameter values warrant scrutiny. Consider for example, the parameters for Lyons et al. (2021). These would be consistent with the results for a *positive* relationship between false news exposure and news discernment ability. But based on common sense and the multiple regression coefficient in their Table F7, this association is likely *negative*. To reconcile Krawcyk and Wilamowski (2017) requires $\lambda < 1$, yet unless this marathon has poor construct validity as a measure of marathon time, one might expect $\lambda = 1$. For Sanchez and Dunning (2021), Varma et al. (2023), and Walters and Fernbach (2021), the ability confound is a less-compelling explanation than is overconfidence. Many of the remaining cases seem broadly plausible, particularly if one considers parameter configurations in which $\rho < 1$. Even in those cases, some of the reported analyses represent a subset of those reported in the paper. For example, Drummond Otten & Fischhoff (2023) report multiple studies; the table reports results from their Study 3. I include Reuben et al. (2024), but their primary focus is on the interaction between competitiveness and overconfidence. While there are implications for that analysis, the quantitative implications do not fall out of the model in a straightforward fashion. In many finance articles, the use of the overconfidence measures are intertwined with foundational assumptions about the (im)plausibility of retail investors systematically outperforming the market on a risk-adjusted basis.

[2] A critical implication of this is that the proper interpretation of $\lambda < 1$ is $\lambda < 1$ *and* $\sigma_v^2 > 0$. For difference scores, λ is the active ingredient whereas for residuals, λ relative to σ_v^2 matters. Analyses involving residual scores could be recharacterized with a different λ (possibly $\lambda = 1$) and an appropriately-scaled σ_v^2.

Together, these results help to characterize the potential applicability of the model across a wide range of literatures. But an important caveat is that these should not be interpreted as a representative sample, the analyses do not attempt to explain the full set of results, and in some cases there may be additional evidence that could help to address this potential counterexplanation.

Table A2
Sample Parameter Values That Recreate Sample of Correlations

Article and Brief Analysis Summary	Calculated	Source	Parameters	Implied
Anderson, Baker, & Robinson (2017), *JFE*	$r_{r,o} = .123$	Table 7 (2): b = .069, se = .008, N = 4896	$\beta = .25$	$r_{r,o} = .121$
Overconfidence: Financial literacy	$r_{p,o} = .167$	Table 7 (1): b = .082, se = .007, N = 4896	$\rho = 1$	$r_{p,o} = .163$
Outcome: Retirement planning			$\lambda = .65$	
Difference score with performance control			$\alpha = .45$	
Åstebro, Jeffrey, & Adomdza (2007), *JBDM*	$r_{\tilde{p},o} = .060$	Table 3: t = 1.97, p259: N = 780 + 300	$\beta = .15$	$r_{\tilde{p},o} = .060$
Confidence: City size quiz			$\rho = 1$	
Comparison: Inventor vs. not			$\lambda = .40$	
Confidence without performance control			$\alpha = 0$	
Dean & Ortoleva (2019), *PNAS*	$r_{\tilde{p},o} = .15$	Table 1: r = -0.15, N = .92*(190-10) = 165	$\beta = .20$	$r_{\tilde{p},o} = .191$
Overplacement: Self-evaluated relative	$r_{p,o} = .23$	Table 1: r = -0.23, N = .92*(190-10) = 165	$\rho = 1$	$r_{p,o} = .190$
performance on Raven's matrices (not adj.)		Flip sign to ensure positive β	$\lambda = .95$	
Correlate: Discount rate			$\alpha = .10$	
Drummond Otten & Fischhoff (2020), *JBDM*	$r_{p,o} = .33$	Table 4: r = .33, N = 332	$\beta = .35$	$r_{p,o} = .33$
Overconfidence: Scientific Reasoning Scale	$r_{\tilde{p},o} = .14$	Table 4: r = .14, N = 332	$\rho = .2$	$r_{\tilde{p},o} = .14$
Correlate: Science education	$r_{\Delta,o} = -.18$	Table 4: r = -.18, N = 332	$\lambda = .95$	$r_{\Delta,o} = -.18$
Raw correlation with difference score			$\alpha = .75$	
Grinblatt & Keloharju (2009), *JF*	$r_{r,o} = .034$	Table III: ln(Num Trades) t = 2.93, N = 7359	$\beta = .05$	$r_{r,o} = .033$
Overconfidence: Self-confidence rating	$r_{\tilde{p},o} = .058$	Table I ln(Num Trades) r = .058	$\rho = 1$	$r_{\tilde{p},o} = .050$
Outcome: Trading activity			$\lambda = .75$	
Confidence with performance control			$\alpha = 1$	
Hilton et al. (2011), *JBDM*	$r_{\tilde{p},o} = .397$	Table 4: r's = .34, .45, N = 97-4 = 93	$\beta = .45$	$r_{\tilde{p},o} = .396$
Overconfidence: Self-evaluated relative		Use Fisher r to z' transform	$\rho = 1$	
performance on quiz (not adj.)			$\lambda = .80$	
Correlate: Unrealistic optimism			$\alpha = .40$	
Huffman, Raymond, & Shvets (2022), *AER*	$r_{\Delta,o} = .231$	Table 3 (3): b = 0.20, se = 0.10, N = 75	$\beta = .55$	$r_{\Delta,o} = .230$
Overconfidence: Prediction vs. benchmark			$\rho = 1$	
Predictor: Flattering memory with controls			$\lambda = .65$	
Difference score predicted by memory			$\alpha = 1$	
Ke (2021), *JF*	$r_{r,o} = .027$	Table V (2): b = .042, se = .007, N = 50449	$\beta = .05$	$r_{r,o} = .027$
Overconfidence: Memory recall task			$\rho = 1$	
Outcome: Stock ownership			$\lambda = .75$	
Difference score with performance control			$\alpha = .70$	

Krawczyk & Wilamowski (2017), *JBDM*	$r_{\Delta,o} = .191$	Table 1: b = 624.97, se = 175.81, N = 345	$\beta = .60$	$r_{\Delta,o} = .190$
Overconfidence: Forecast marathon time			$\rho = 1$	
Correlate: Gender			$\lambda = .80$	
Gender predicts forecast error			$\alpha = 1$	
Landier & Thesmar (2008), *RFS*	$r_{\Delta,o} = .041$	Table 9: b = .03, se = .01, N = 5474	$\beta = .10$	$r_{\Delta,o} = .042$
Overconfidence: Expect development			$\rho = 1$	
Correlate: Short-term debt			$\lambda = .65$	
Use forecast errors as difference score			$\alpha = 1$	
Liu, Peng, Xiong, & Xiong (2022), *JFE*	$r_{r,o} = .040$	Table 10: b = 15.695, t = 2.760, N = 4648	$\beta = .05$	$r_{r,o} = .042$
Overplacement: Forecast mkt prfmnc			$\rho = 1$	
Correlate: Turnover			$\lambda = .55$	
Control for actual performance			$\alpha = 1$	
Lyons et al. (2021), *PNAS*	$r_{r,o} = .041$	Table F2: b = .0615, se = .0300, N = 2547	$\beta = .15$	$r_{r,o} = .060$
Overconfidence: News discernment task	$r_{\Delta,o} = .061$	Table 1: b = .0569, se = .0186, N = 2547	$\rho = 1$	$r_{\Delta,o} = .043$
Outcome: False news exposure			$\lambda = .35$	
Both difference and residual approaches			$\alpha = .25$	
Reuben, Sapienza, & Zingales (2024), *JF*	$r_{p,o} = .302$	Table II (2): b = .165, se = .027, N = 409	$\beta = .55$	$r_{p,o} = .303$
Overconfidence: Relative task placement	$r_{r,o} = .304$	Table II (3): b = .203, se = .033, N = 409	$\rho = 1$	$r_{r,o} = .304$
Outcome: Compete			$\lambda = .55$	
Control for performance, so use residual			$\alpha = .50$	
Sanchez & Dunning (2021), *JPSP*	$r_{\tilde{p},o} = .08$	Table 2: rs = -.11, -.05, N = 289+350	$\beta = .40$	$r_{\tilde{p},o} = .058$
Overconfidence: Multiple-choice quiz	$r_{p,o} = .355$	Table 2: rs = -.35, -.36, N = 289+350	$\rho = .05$	$r_{p,o} = .400$
Correlate: Jumping to Conclusions	$r_{\Delta,o} = -.325$	Table 2: rs = .32, .33, N = 289+350	$\lambda = 1$	$r_{\Delta,o} = -.262$
Uses difference score assessed per item		Use Fisher r to z' transform; flip signs	$\alpha = .90$	
Varma, Bommaraju, & Singh (2023), *JMR*	$r_{\tilde{p},o} = .120$	Table 6: M = .366, SD = .482 vs.	$\beta = .15$	$r_{\tilde{p},o} = .120$
Overconfidence: Are better leader		M = .488, SD = .500, N = 279+451=730	$\rho = 1$	
Correlate: Gender			$\lambda = .80$	
No control for actual performance			$\alpha = 0$	
Walters & Fernbach (2021), *PNAS*	$r_{\tilde{p},o} = .110$	Table 2: b = .197, se = .062, N = 822	$\beta = .2$	$r_{\tilde{p},o} = .110$
Overconfidence: Forecast mkt performance			$\rho = 1$	
Correlate: Positively biased memory			$\lambda = .55$	
Assumes better than average is bias			$\alpha = 0$	

Table A3

Rationale for Articles in Table A1 Excluded from Table A2

Article	Reason for Exclusion From Table A2
Agranov & Buyalskaya (2022), *MS*	Used as unreported controls
Anderson et al. (2012), *JPSP*	Included in main text
Anderson & Lu (2017), *MS*	Primarily as outcome of experimental treatment
Avery et al. (2022), *REStat*	Correlations with components
Belmi et al. (2020), *JPSP*	Included in main text
Cavalan et al. (2023), *JEP:G*	Primarily as outcome of experimental treatment
Chen et al. (2007), *JBDM*	Use trading as proxy, rather than direct measure
Eyal et al. (2018)	Primarily as outcome of experimental treatment
Fast et al. (2012)	Primarily as outcome of experimental treatment
Gillen et al. (2019)	Methodological paper forcused on measurement error
Larrick et al. (2007)	Internal structure of overconfidence
Moore & Healy (2008)	Included in main text
Prims & Moore (2017)	Focus on overprecision; $\beta = 0$ explains overestimation, overplacement nulls
Ren & Croson (2013)	Focus on overprecision; $\beta = 0$ explains null on overplacement as control
Van Zant (2022)	Primarily as outcome of experimental treatment

Relation to Prior Critiques of the Better-Than-Average and Dunning-Kruger Effects

The present work follows a longstanding research dialogue and set of critiques regarding the Better-Than-Average effect (e.g., Svenson 1981) and whether people who are unskilled are unaware (sometimes referred to as the "Dunning-Kruger Effect," DKE; Kruger & Dunning 1999). A complete characterization of all arguments is out of scope, but a brief discussion helps to contextualize the contribution of the present research.

Benoît and Dubra (2011) prove that apparent overconfidence in the aggregate—like the Better-Than-Average effect—can come about through Bayesian reasoning regarding a distribution of beliefs.[3] The current work differs in three important ways. First, they consider aggregate levels of overconfidence whereas I consider measured differences in overconfidence. The present concern persists even when theirs is addressed. Second, their model is based on updating beliefs about one's ability based on one's performance. The present model is based on evaluating one's performance based on beliefs about one's ability. Third, their finding is due to using estimates from distributions which may not aggregate. The

[3] In a follow-up paper, Benoît et al. (2015) find that although the model can produce such an effect through Bayesian reasoning, there is evidence of aggregate overconfidence once one accounts for this critique.

present model implies the confound I describe would persist in the experiments they propose distinguishing mean from median assessments.

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 does not address the difference in correspondence between objective and subjective performance among the skilled versus unskilled. Because the absolute deviation is a function of item ease or difficulty, it is possible for a larger absolute deviation for skilled participants to coexist with reduced correspondence for unskilled participants (Burson et al., 2006). 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 in certain cases the unskilled are indeed unaware. Feld et al. (2017) use instrumental variables and find evidence for the DKE. By assuming use of a difference score and noisy but non-regressive performance measures, that model does not generate the confound I express concern about in the present research. Jansen et al. (2021) present a Bayesian account of the DKE, finding that much but 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 and most-importantly does not address the focus of the current paper: the broader implications for the measurement of overconfidence beyond the DKE.

The DKE is implicated by a multifaceted data signature. 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.

Derivation of Equations 6 and 8: Confounded Residuals and Difference Scores

Equations (1) through (8) are repeated here as (A1) through (A8) for ease of reference.

$$S_i \sim D(0,1) \tag{A1}$$

$$\tilde{S}_i = \rho S_i + \zeta_i \tag{A2}$$

$$P_i = \lambda S_i + \nu_i \tag{A3}$$

$$\tilde{P}_i = \alpha \tilde{S}_i + (1-\alpha) P_i + \upsilon_i \tag{A4}$$

$$\tilde{P}_i = \gamma P_i + \epsilon_i \tag{A5}$$

$$E[\epsilon|S] = \rho\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_\nu^2}\right)\alpha S \tag{A6}$$

$$\Delta_i = \tilde{P}_i - P_i \tag{A7}$$

$$E[\Delta|S] = (\rho - \lambda)\alpha S \tag{A8}$$

Plugging (A2) and (A3) into (A4), we can rewrite self-evaluations of performance, $\tilde{P}$, in terms of Skill, *S*:

$$\tilde{P}_i = \alpha(\rho S_i + \zeta_i) + (1-\alpha)(\lambda S_i + \nu_i) + \upsilon_i \tag{A9}$$

We then use (A5) to rewrite γ in terms of the structural parameters α, λ, ρ, and σ_ν^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 $(\lambda\frac{\sigma_S^2}{\sigma_P^2})(\rho\frac{\sigma_{\tilde{S}}}{\sigma_S})\alpha$. 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. Recall $\sigma_{\tilde{S}}^2 = \sigma_S^2 = 1$. We reexpress $\sigma_P^2 = \lambda^2 + \sigma_\nu^2$. This gives us:

$$\tilde{P}_i = (1-\alpha)P_i + \left(\frac{1}{\lambda^2+\sigma_\nu^2}\right)\lambda\rho\alpha P_i + \epsilon_i \tag{A10}$$

We then reexpress $\tilde{P}_i$ in (A10) in terms of S_i via (A9) and P_i in (A10) in terms of S_i via (A3), and simplify and isolate ϵ_i:

$$\epsilon_i = \rho\left(1 - \frac{\lambda^2}{\lambda^2+\sigma_\nu^2}\right)\alpha S_i + \upsilon_i + \alpha\zeta_i - \rho\left(\frac{\lambda}{\lambda^2+\sigma_\nu^2}\right)\alpha\nu_i \tag{A11}$$

As ν_i, υ_i, and ζ_i are independent and mean 0, they drop out in expectation, providing (A6).

To derive the expected value of the difference score, we use (A3) to reexpress P_i in (A7) in terms of S_i and we use (A9) to express $\tilde{P}_i$ in (A7) terms of S_i. Simplifying gives us:

$$\Delta_i = (\rho - \lambda)\alpha S_i + \upsilon_i + \alpha(\zeta_i - \nu_i) \quad \text{(A12)}$$

Once again, because ν_i, υ_i, and ζ_i are independent and mean 0, they drop out in expectation, leaving us with (A8).

Empirical Application III: Reassessing Correlates of Financial Planning

Overview, Data, and Analysis Reproduction

Parker et al. (2012) study the role of "inappropriate confidence" (which Parker & Stone, 2014, later refer to as "unjustified confidence" and much 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."[4] 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 re-analysis; 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 assessing people's subjective understanding of economics.[5] Study 2 (N = 1114) assessed general

[4] 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. This paper does note the correlational nature of the findings as a caution on drawing causal conclusions. My critique applies to both causal and correlational claims.

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

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.[6] To do so, I reanalyzed the original data from the four ALP studies. Relevant correlations and descriptive statistics are given in Table A4, 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.[7]

A Model Where Overconfidence Does Not Matter

I first fit these correlations to the model in Figure A1. 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. 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

[6] 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.

[7] In email correspondence with the first author, we attempted to determine the cause of the discrepancy, but were unable to. I am grateful for the first author's time and effort digging into more-than-10-year-old data and code.

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).[8] The model was fit using full information maximum likelihood for missing data using the `lavaan` package v0.6-12 (Rosseel, 2012) in R.[9]

Table A4

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

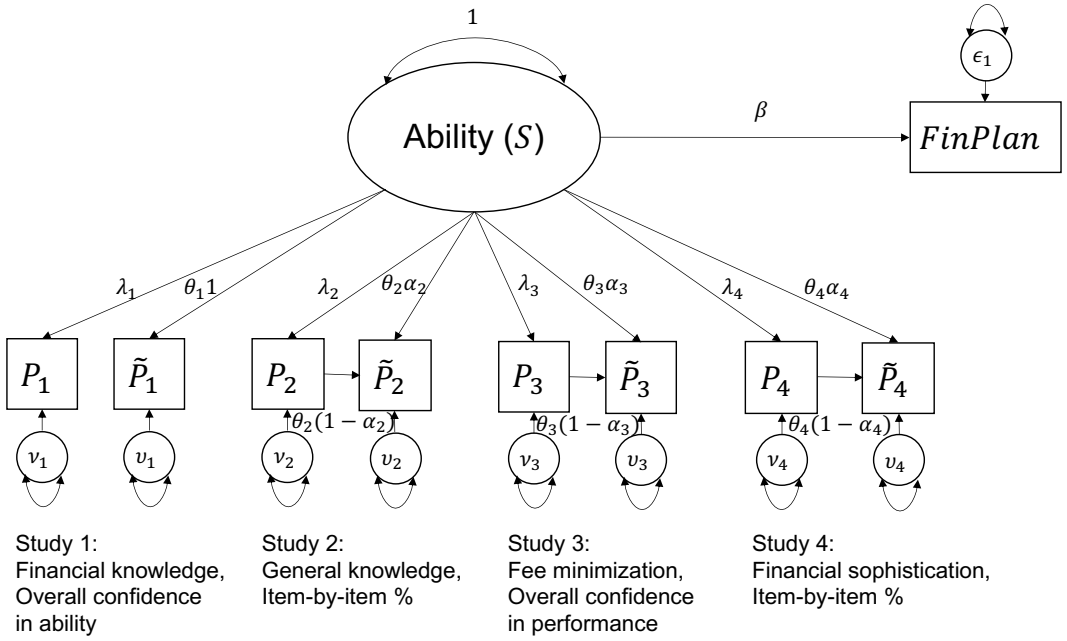
Calculated	Perf1	Perf2	Perf3	Perf4	Conf1	Conf2	Conf3	Conf4	Outcome
Perf1									
Perf2	0.31								
Perf3	0.34	0.16							
Perf4	0.63	0.33	0.39						
Conf1	0.36	0.08	0.19	0.25					
Conf2	0.34	0.33	0.16	0.32	0.20				
Conf3	0.29	0.10	0.31	0.29	0.34	0.21			
Conf4	0.53	0.05	0.37	0.53	0.35	0.41	0.39		
Outcome	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

[8] 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 $\frac{\theta\alpha}{\theta\alpha+\theta(1-\alpha)}$.

[9] 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.

Figure A1

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



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 reasons to expect that the model depicted in Figure A1 is insufficient to fully account for the data, because it is known to be wrong in ways unrelated to the addition of overconfidence.

Results

Despite these model misspecifications, the estimated parameters appear to be reasonable; see Table A5. 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 A5

Standardized Parameter Estimates from Model Excluding the Possibility of Overconfidence

Study	λ	α^{a}	θ^{a}	σ_v^2	σ_v^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.

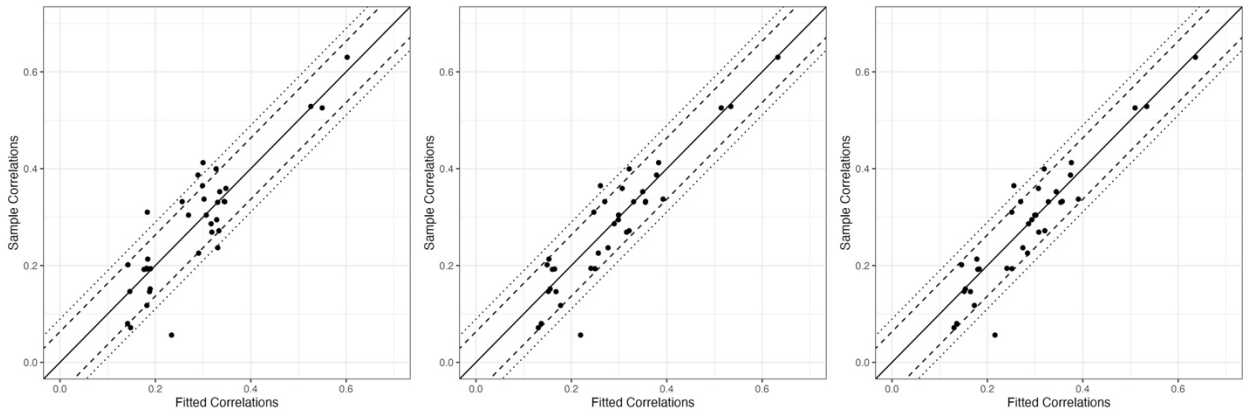
As shown in Figure A2 and Table A6, the set of correlations derived from the fitted parameter estimates fit the observed data moderately well, especially considering the ways in which it is known to be inadequate. The largest absolute deviations are also instructive. First, the model overestimates the 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.

The baseline model in Figure A1 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 A6 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 Figure 2A allowing $\rho \leq 1$. This is a nesting model, as it is equivalent to the baseline model if the correlation between ability and confidence is fixed to 1. But again, only

ability is allowed to affect financial planning. This represents a case in which people may have imperfect self-insight, but ability is still the only causal force regarding 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 A6, 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 A2).

Figure A2

Fitted Correlations and Observed Correlations in the Data in Three Models



Note. The left panel represents the model shown in Figure A1. The center panel allows for the presence of, but no effect of, overconfidence, that is, $\rho \leq 1$, as in Figure 2A. 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 A6

Model Fit Statistics

Model	df	χ^2	CFI	RMSEA	logLik	AIC	AICc[a]	BIC
1 (Perfect Self-Insight)	24	194.09	0.90	0.072	-11735	23530	23607	23687
2 (Partial Self-Insight)	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 (Perfect Self-Insight)	4	19.08	0.98	0.057	-6113	12258	12394	12339
5 (Partial Self-Insight)	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.

Taken together, these analyses suggest that a parsimonious representation of the reported data can be derived from a simple model based only on ability and without (inappropriate, unjustified, or over-) confidence. Some evidence suggests that the model in Figure 2A allowing beliefs about ability to be imperfectly correlated with ability 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 A6.

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.

Approaches to Accounting for Measurement Error

Recommendation 2 in the main text is to account for measurement error. This has the potential to help in the set of cases in which (a) the residual or multiple regression approach is being used rather than the difference score, and (b) σ_v^2 is not driven by stable luck: that is, error represents noise. Two frequently-used techniques for addressing measurement error are structural equation models and errors-in-variables.

Structural Equation Models

Structural equation models (e.g., Kline, 2005) permit the researcher to model relationships among latent variables, unattenuated by measurement error. This typically relies upon multiple indicators of performance, although as noted by Westfall and Yarkoni (2016), it is possible 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 systematic error. If the performance measure includes systematic error in addition to measurement error, eliminating measurement error will not solve the problem. The reanalysis of Parker et al. (2012) presented above is an example of such an approach, and an illustration of how conclusions may change when accounting for measurement 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 there is no systematic error in the error term, only noise. 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 higher than the true reliability of the measure as a measure of financial knowledge, because the error includes both measurement error and systematic error (i.e., trust in institutions). This will lead us to underestimate the extent of the problem.

A full development of the errors-in-variables approach is beyond the scope of the current investigation; interested readers are referred to Fuller (1987) for a statistical treatment and Culpepper and Aguinis (2011) for an application 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 (given its assumptions) to control for ability, not just performance, which affects the coefficient on self-evaluation. In the next section I report the results from using errors-in-variables methodology in the Parker et al. (2012) example.

These approaches are not a panacea: they still assume perfect construct validity. If one is able to adequately account for measurement error, one will get an estimate of results using the true score of whatever the measure measures. But whatever the measure measures is not guaranteed to align with the intended construct (e.g., one might use height as a highly reliable but completely invalid measure of ability). Thus, these can help to address the reliability concern for the residual measure, but do not address the validity concern for either the residual measure or the difference measure. If one relies on difference scores, one is still left with an independent set of concerns (e.g., Cronbach & Furby 1970; Edwards & Parry 1993; Johns 1981).

An Empirical Application of the Errors-in-Variables Approach

To examine the potential of the errors-in-variables approach, I use the *eivreg* function from the *eivtools* package (Lockwood, 2018) and the *eiv* function provided by Culpepper and Aguinis (2011), both implemented in R. Errors in variables adjustments require an estimate of the reliability 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. 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 Parker et al. (2012).[10] 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 A7. All results using *eiv* were quantitatively similar and led to identical statistical conclusions.

Table A7

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.

[10] 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.

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 greater 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 (rather fraught, in this case) assumptions about reliabilities.

A proponent of the confidence-causes-planning story might focus on the Study 2 results: even after accounting for measurement error, confidence appears to play a role. A detractor 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.

References

References marked with an asterisk indicate studies included in Tables A1-A3.

*Agranov, M., & Buyalskaya, A. (2022). Deterrence effects of enforcement schemes: An experimental study. *Management Science*, *68*(5), 3573-3589.

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

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

*Anderson, M. L., & Lu, F. (2017). Learning to manage and managing to learn: The effects of student leadership service. *Management Science*, *63*(10), 3246-3261.

*Åstebro, T., Jeffrey, S. A., & Adomdza, G. K. (2007). Inventor perseverance after being told to quit: The role of cognitive biases. *Journal of Behavioral Decision Making*, *20*(3), 253-272.

*Avery, M., Giuntella, O., & Jiao, P. (2022). Why don't we sleep enough? A field experiment among college students. *Review of Economics and Statistics*, 1-45.

*Belmi, P., Neale, M. A., Reiff, D., & Ulfe, R. (2020). The social advantage of miscalibrated individuals: The relationship between social class and overconfidence and its implications for class-based inequality. *Journal of Personality and Social Psychology*, *118*(2), 254-282.

*Cavalan, Q., Vergnaud, J. C., & De Gardelle, V. (2023). From local to global estimations of confidence in perceptual decisions. *Journal of Experimental Psychology: General*, *152*(9), 2544-2558.

*Chen, G., Kim, K. A., Nofsinger, J. R., & Rui, O. M. (2007). Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors. *Journal of Behavioral Decision Making*, *20*(4), 425-451.

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

*Dean, M., & Ortoleva, P. (2019). The empirical relationship between nonstandard economic behaviors. *Proceedings of the National Academy of Sciences*, *116*(33), 16262-16267.

*Drummond Otten, C., & Fischhoff, B. (2023). Calibration of scientific reasoning ability. *Journal of Behavioral Decision Making*, *36*(3), e2306.

*Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of Personality and Social Psychology*, *114*(4), 547-571.

*Fast, N. J., Sivanathan, N., Mayer, N. D., & Galinsky, A. D. (2012). Power and overconfident decision-making. *Organizational Behavior and Human Decision Processes*, *117*(2), 249-260.

*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.

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

*Hilton, D., Regner, I., Cabantous, L., Charalambides, L., & Vautier, S. (2011). Do positive illusions predict overconfidence in judgment? A test using interval production and probability evaluation measures of miscalibration. *Journal of Behavioral Decision Making*, *24*(2), 117-139.

*Huffman, D., Raymond, C., & Shvets, J. (2022). Persistent overconfidence and biased memory: Evidence from managers. *American Economic Review*, *112*(10), 3141-3175.

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

*Krawczyk, M., & Wilamowski, M. (2017). Are we all overconfident in the long run? Evidence from one million marathon participants. *Journal of Behavioral Decision Making*, *30*(3), 719-730.

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

*Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76-94.

*Liu, H., Peng, C., Xiong, W. A., & Xiong, W. (2022). Taming the bias zoo. *Journal of Financial*

Economics, *143*(2), 716-741.

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

*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.

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. *The Journal of Finance*, *60*(6), 2661-2700.

Malmendier, U., & Tate, G. (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. *Journal of Financial Economics*, *89*(1), 20-43.

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

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

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

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

*Prims, J. P., & Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and Decision Making*, *12*(1), 29-41.

*Ren, Y., & Croson, R. (2013). Overconfidence in newsvendor orders: An experimental study. *Management Science*, *59*(11), 2502-2517.

*Reuben, E., Sapienza, P., & Zingales, L. (2024). Overconfidence and preferences for competition. *The Journal of Finance*, *79*(2), 1087-1121.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L V. Hedges (Eds.), *The*

handbook of research synthesis (pp. 231-244). Russell Sage Foundation.

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

*Sanchez, C., & Dunning, D. (2021). Jumping to conclusions: Implications for reasoning errors, false belief, knowledge corruption, and impeded learning. *Journal of Personality and Social Psychology*, *120*(3), 789-815.

*Van Zant, A. B. (2022). Strategically overconfident (to a fault): How self-promotion motivates advisor confidence. *Journal of Applied Psychology*, *107*(1), 109-129.

*Varma, R., Bommaraju, R., & Singh, S. S. (2023). Female Chief Marketing Officers: When and Why Do Their Marketing Decisions Differ from Their Male Counterparts'?. *Journal of Marketing Research*, *60*(6), 1154-1176.

*Walters, D. J., & Fernbach, P. M. (2021). Investor memory of past performance is positively biased and predicts overconfidence. *Proceedings of the National Academy of Sciences*, *118*(36).