

UC Berkeley

UC Berkeley Previously Published Works

Title

Is Overconfidence a Trait? An Adversarial Collaboration

Permalink

<https://escholarship.org/uc/item/20t3p23p>

Journal

Psychological Science, 36(12)

ISSN

0956-7976

Authors

Binnendyk, Jabin
Li, Sophia
Costello, Thomas
[et al.](#)

Publication Date

2025-12-01

DOI

10.1177/09567976251401538

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

Is Overconfidence a Trait? An Adversarial Collaboration

Psychological Science
2025, Vol. 36(12) 941–951
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976251401538
www.psychologicalscience.org/PS



Jabin Binnendyk¹, Sophia Li², Thomas Costello³,
Randall Hale⁴, Don A. Moore², and Gordon Pennycook^{1,5}

¹Department of Psychology, Cornell University; ²Haas School of Business, University of California, Berkeley;

³Department of Psychology, American University; ⁴Department of Psychology, University of California, Los Angeles; and ⁵Hill and Levene Schools of Business, University of Regina

Abstract

A fundamental underlying question about the nature of overconfidence has continued to be subject to scholarly dispute: Is overconfidence a genuine psychological trait? To advance this contested research topic, we engaged in an adversarial collaboration in which two research teams agreed upon a set of critical tests and preregistered their analyses and predictions prior to data collection. Our study ($N = 942$; U.S. adults from CloudConnect) leverages a methodological innovation: To measure trait overconfidence absent task-related confounds, we developed a set of novel tasks in which performance is ostensibly random. When we assess confidence this way, we find robust relationships across tasks as measured by both confirmatory factor analyses and raw correlations. This indicates that some people do believe that they are able to perform relatively well on tasks even when there is little reason for that confidence. Our results support the claim that overconfidence might be a trait.

Keywords

overconfidence, confidence, adversarial collaboration, trait, individual difference

Received 11/15/24; revision accepted 11/4/25

Overconfidence is one of the most consequential and pervasive cognitive biases (Bazerman & Moore, 2013) and has been a major focus of research (Moore & Schatz, 2017). The tendency to be more confident than is justified can exacerbate other biases if people who mistakenly believe that they are good at a particular task see little reason to gather potentially disconfirming information, consider alternative approaches, or invest in improvement (Moore, 2023; Vancouver et al., 2002; Walters et al., 2017). Overconfidence can exacerbate susceptibility to misinformation (Lyons et al., 2021; Peng & Shen, 2024), anti-science attitudes (Lackner et al., 2023; Light et al., 2022), receptivity to bullshit (Littrell et al., 2021), and belief in conspiracies (Pennycook et al., 2025; Vitriol & Marsh, 2018).

Despite consensus around the importance of overconfidence, scholars disagree about whether overconfidence qualifies as a trait. Past work has examined correlations between overconfidence and a variety of other individual differences, such as gender (Barber & Odean, 2001; Exley & Kessler, 2022; Niederle &

Vesterlund, 2007), age (Prims & Moore, 2017), narcissism (Campbell et al., 2004; O'Reilly & Hall, 2021), and occupation (Malmendier & Tate, 2015). However, few of these correlations are robust to replication across different tasks, forms, and measures of overconfidence. For instance, whereas some studies report that men are more overconfident than women, even more studies fail to replicate that relationship (Acker & Duck, 2008; Amirkhanyan et al., 2024; Lundeberg et al., 1994; Nekby et al., 2008; Prims & Moore, 2017).

Furthermore, past work has found that overconfidence within individuals is unstable across situations and measures. For example, the tendency to *overestimate* (i.e., to think you are better than you are at a task) is not as strongly correlated with *overprecision* (i.e., having

Corresponding Authors:

Jabin Binnendyk, Cornell University, Department of Psychology
Email: jdb437@cornell.edu

Sophia Li, University of California, Berkeley, Haas School of Business
Email: sophia-li@berkeley.edu

excessive confidence than you know the truth) as one might expect (Moore & Healy, 2008). Relatedly, overconfidence does not strongly correspond across tasks. For example, one task may involve guessing the winner of 10 football games and then estimating an overall score for the task. If overconfidence on such a task is influenced by expertise and familiarity (Moore & Healy, 2008; Moore et al., 2015), then overconfidence in guessing the winner of football games may not correspond with overconfidence in predicting the weather, stock prices, or how well one scores on a math test. Moreover, there are well-documented effects of task difficulty on overconfidence (Moore & Small, 2007): People believe they are above average on easy tasks but below average on difficult tasks. Thus, although a trait interpretation of overconfidence may be intuitively appealing, the data supporting it has been mixed. Unlike stable traits, overconfidence may be more akin to risk preferences, which vary dramatically depending on the context (Weber & Johnson, 2009). Consistent with this, Li et al. (2024) recently claimed that overconfidence is too inconsistent across forms, time, and measures for it to be justifiably considered a trait.

In contrast, Binnendyk and Pennycook (2023) argued that our ability to determine whether overconfidence is a trait is complicated by the noise inherent in task performance. Overconfidence is typically calculated by subtracting actual performance from some judgment of confidence (such as beliefs about performance); hence, any measure of overconfidence is perturbed by noise in performance (see also Lawson et al., 2023, and Spiller, 2024). Binnendyk and Pennycook (2023) introduced a novel measurement approach to address this problem. Their Generalized Overconfidence Task (GOT) presented participants with grainy images and two options for what each image contained (see Fig. 1). Because people who do better on the task do not know that they are doing better—and, in fact, are only doing better because of chance—Binnendyk and Pennycook argued that systematic variation in confidence ought to be attributable to a more generalized overconfidence trait. And, indeed, they found that scores on the GOT correlate with conceptually related measures, such as overclaiming, bullshit receptivity, actively open-minded thinking, and others.

Although this supports the argument that confidence is a trait, the GOT consists of just a single idiosyncratic task. The results from Binnendyk and Pennycook (2024) merit generalization.

An Adversarial Collaboration

Our experiment represents an adversarial collaboration between two research teams with divergent perspectives.

Team 1 (the believers) predicted that the confidence measured by the GOT represented an underlying psychological trait that would generalize to other tasks. Team 2 (the skeptics) was dubious that overconfidence is a trait, and they predicted that whatever the GOT measured, it would not generalize to other tasks. Following the logic of the GOT, we devised a set of tasks for which participants would have little to no basis for confidence. For example, our survey asked participants to guess whether someone has an “innie” or “outie” belly button based solely on a picture of the person’s face. We then pretested these tasks to confirm the following parameters: (a) The task has a mean score that rarely exceeds chance (indicating that participants are guessing), (b) the task is not skill-based (i.e., there is little correspondence between actual and estimated performance), and (c) the task is novel (i.e., participants would not have a priori reasons for thinking that they would be good at the task). This led us to select eight novel tasks, in addition to the GOT (see the Supplemental Material available online for full details on the pretest). Notably, these tasks were chosen specifically because they do involve novel judgments that do not reflect common real-world cases in which people are overconfident. This ensures that if confidence ratings vary systematically across participants and tasks, we can attribute it to the underlying trait of being overconfident as opposed to prior exposure to the type of decision being made, or to some underlying competency.

If overconfidence is a trait, participants’ judgments of confidence (and estimated performance) should be robustly correlated across the various unique tasks. That is, people who are unjustifiably confident in their ability to discern the objects in fuzzy images (Fig. 1) should also be unjustifiably confident in their ability to guess whether someone has an innie belly button. Alternatively, if overconfidence is not a stable trait, confidence ratings for tasks that involve guessing should either be at floor or (like task performance) driven by randomness. Hence, our study’s preregistration makes opposing predictions for correlations of confidence measures across tasks (outlined below).

Research Transparency Statement

General disclosures

Conflicts of interest: All authors declare no conflicts of interest. **Funding:** This research was funded by the Institute for Humane Studies at George Mason University. We would like to thank Cory Clark and the University of Pennsylvania Adversarial Collaboration Project for their assistance in obtaining funding. **Artificial intelligence:** No artificial-intelligence-assisted technologies

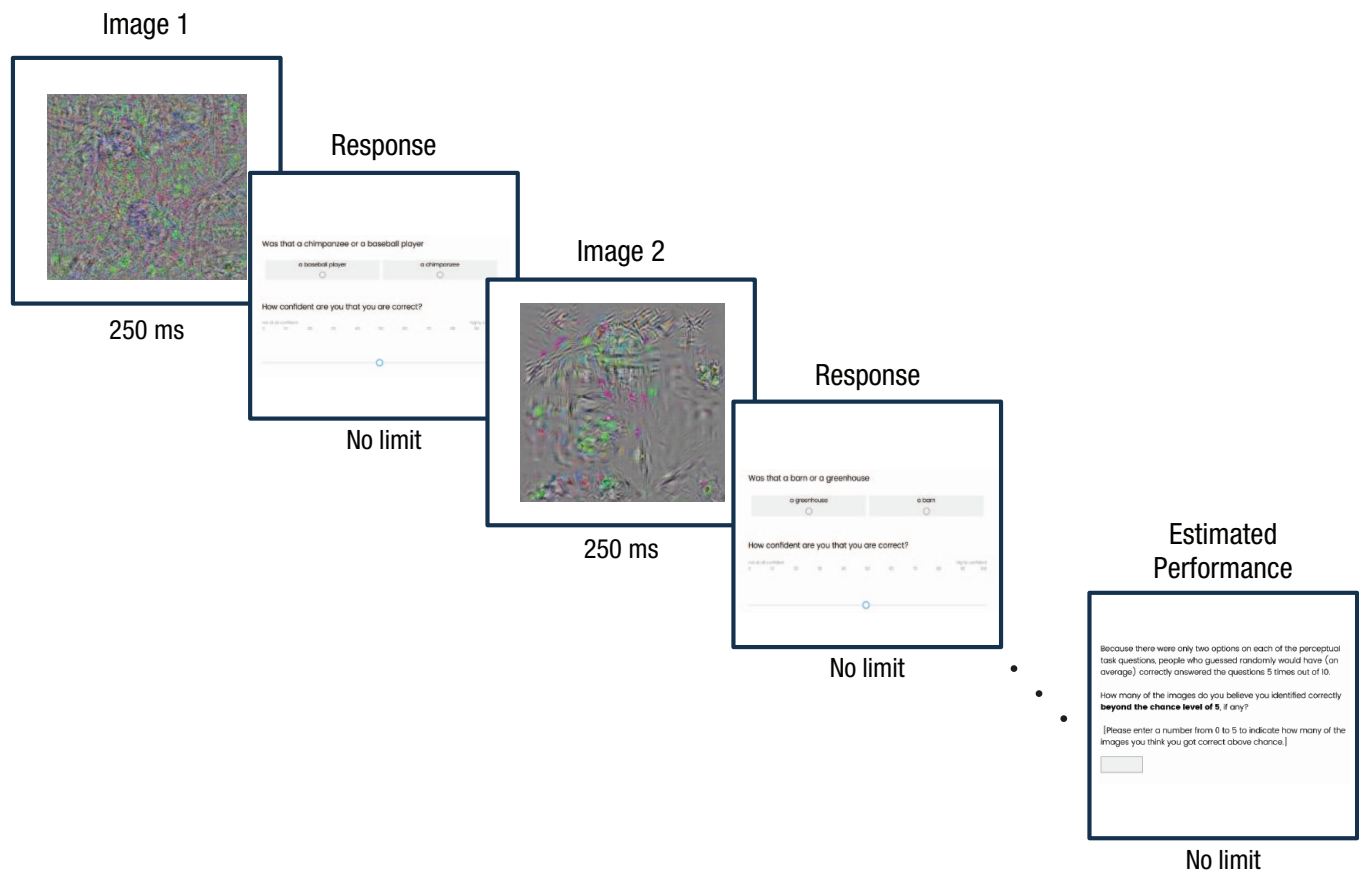


Fig. 1. Generalized Overconfidence Task (GOT) procedure from Binnendyk and Pennycook (2024).

were used in this research or the creation of this article.

Ethics: This research received approval from the ethics board at Cornell University (ID No. 148258). **Open Science Framework (OSF):** To ensure long-term preservation, we registered all OSF files at <https://osf.io/dxkq8>. **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the Supplemental Material) has been independently confirmed by the journal's STAR team.

Study 1 disclosures

Preregistration: The hypotheses, methods, and analysis plan were preregistered (<https://osf.io/cqejr>) on August 14, 2024, prior to data collection, which began on August 21, 2024. Deviations from the preregistration were minor and can be found in Table S1 in the Supplemental Material available online. **Materials:** All study materials are publicly available (<https://osf.io/t3xqn/>). **Data:** All primary data are available (<https://osf.io/w3ugn/>). **Analysis scripts:** All analysis scripts are publicly available (<https://osf.io/8rw59/>).

Method

We preregistered this study at <https://osf.io/cqejr> on August 14, 2024, before data collection started on August 21, 2024. All survey materials, data, and analysis code are available at <https://osf.io/9wg8z/>. This research was approved by the Cornell University Institutional Review Board for Human Participant Research Office (Approval No. IRB0148258).

Participants. We preregistered that we would recruit 1,000 participants pre-exclusions, and we wound up with 1,001 participants from CloudConnect. Consistent with our preregistration, we excluded data from any participants who did not complete the study and any subsequent entries from those who completed it multiple times.

Consistent with our preregistration, we excluded data from 32 participants who did not select “strongly agree” on an instructed-choice (attention-check) multiple-choice question;¹ an additional 1 participant who admitted to “not at all” paying attention; and an additional 26 participants who indicated that they “agree” or “strongly agree” that they are paid biweekly by

leprechauns. These exclusions left us with a final sample of 942.

Our participants were 51.27% male, 47.56% female, 1.17% other gender; 66.56% White, 9.02% Asian, 12.63% Black, 5.10% Hispanic, 6.69% other race; $M_{\text{age}} = 38.42$ years, $SD = 12.07$. The median education level was a bachelor's degree in college.

Procedure. Participants completed five (randomly selected) of the following nine pretested tasks, each of which featured 10 items:

- Age: guessing the age of someone (who was either 32 or 33 years old) on the basis of a picture of his or her face, taken from the Iranian Emotional Face Database from Heydari et al. (2023);
- Birth: guessing someone's birth order (i.e., whether a person was the oldest, middle, or younger child) on the basis of a picture of his or her face;
- Birthday: guessing whether someone's birthday was on an even or odd date on the basis of a picture of his or her face;
- Book: guessing the range containing the number of pages in each of 10 books (0–99, 100–199, 200–299, or 300–399);
- Card suit: guessing the suit of cards (clubs, hearts, diamonds, or spades) drawn from a full deck;
- GOT: distinguishing which of two possible objects was present in a fuzzy image, from Binnendyk and Pennycook (2023);
- Innies: guessing whether someone has an innie or outie belly button on the basis of a picture of his or her face;
- Light bulb: guessing which of four light-bulb switches corresponded to one of four light bulbs; and
- Preference: guessing the percentage of other participants in the study who indicated preferring one thing over another (considered correct if the guess was within 5% of the true answer).

For each question they answered, they also indicated how confident they were in their answer, as a percentage up to 100% (the bottom endpoint of the scale varied depending on how many possible answers there were; e.g., questions that had three possible answers were scaled from 33% to 100%, and questions that had three possible answers were scaled from 25% to 100%).² After completing the 10 items, participants also estimated their score (e.g., "How many questions do you think you answered correctly on this task? In other words, how many people's age[s] did you correctly identify as 32 or 33?") and their placement (e.g., "What percentage of other participants do you think you outscored on this task? In other words, fill in the following sentence:

"I correctly identified more 33-year-olds than ___% of participants").

After participants completed their five tasks, they self-reported their attention ("How much were you paying attention to the survey, in the questions you answered so far?") on a scale ranging from 1, *a great deal*, to 5, *none at all*. Then they filled out two trait-measure scales: actively open-minded thinking (Baron, 1994) and neuroticism (Soto & John, 2017). We included these as secondary measures to investigate the strength of the correlations between confidence judgments and conceptually related traits. We also added a bogus item to the neuroticism scale ("I see myself as someone who is paid by leprechauns biweekly") as an additional attention check. Finally, participants provided their demographic information. Participants earned \$2.00 for participating in the study, which on average took 15.91 min.

Results

Summary statistics

As preregistered, we excluded data from two tasks in which task performance and mean confidence judgments were positively correlated ($p < .005$): guessing birth order and guessing participants' preferences (see Table 1, column 7, for performance–confidence correlations on all tasks). We include results with data from all tasks in the Supplemental Material.

On each task, the lowest and probabilistically correct answer on each item (e.g., 25% on the book task when there were four options, and 50% on the age task when there were two options) was also the modal level of item-level confidence. For each task, we calculated the percentage of item-level responses that were the lowest and probabilistically correct answer; these percentages averaged 25.21% ($SD = 8.85\%$) across the seven tasks, ranging from 9.28% on the book task to 35.58% of responses on the birthday task. Similarly, the modal score estimate was the probabilistically correct answer (e.g., 5/10 on tasks like GOT or Innies—i.e., questions with two alternatives). Participants were essentially equally likely to overestimate (41.50%) or underestimate (40.69%) their scores, with the remaining 17.81% of score estimates being correct; 68 participants (7.22%) overestimated their score on every task, and 61 participants (6.48%) underestimated their score on every task. We include graphical descriptives of both item-level confidence and score estimates on all tasks in the Supplemental Material.

Inter-task correlations

We preregistered two tests: a confirmatory factor analysis, and, secondarily,³ a (simpler) analysis in which we

Table 1. Summary Statistics by Task

Task	<i>n</i>	Score	Average item confidence	Score estimate	Score placement	Performance–confidence correlation
Age	520	4.66 (<i>SD</i> = 1.50)	67.51 (<i>SD</i> = 13.92)	4.82 (<i>SD</i> = 1.66)	42.14 (<i>SD</i> = 21.24)	–0.03 (<i>p</i> = .538)
Birth	518	3.90 (<i>SD</i> = 1.55)	67.57 (<i>SD</i> = 13.30)	4.09 (<i>SD</i> = 1.87)	37.53 (<i>SD</i> = 22.48)	0.15 (<i>p</i> < .001)
Birthday	520	4.82 (<i>SD</i> = 1.46)	60.93 (<i>SD</i> = 12.32)	4.40 (<i>SD</i> = 1.55)	37.51 (<i>SD</i> = 19.51)	0.01 (<i>p</i> = .864)
Book	525	3.04 (<i>SD</i> = 1.45)	57.53 (<i>SD</i> = 17.05)	4.32 (<i>SD</i> = 1.72)	39.34 (<i>SD</i> = 21.26)	0.04 (<i>p</i> = .344)
Card suit	541	2.75 (<i>SD</i> = 1.28)	39.08 (<i>SD</i> = 17.17)	2.90 (<i>SD</i> = 1.73)	28.92 (<i>SD</i> = 21.49)	< 0.01 (<i>p</i> = .951)
GOT	523	5.96 (<i>SD</i> = 1.56)	61.42 (<i>SD</i> = 0.53)	3.62 (<i>SD</i> = 1.87)	32.58 (<i>SD</i> = 22.07)	–0.03 (<i>p</i> = .530)
Innie	520	4.95 (<i>SD</i> = 1.51)	67.03 (<i>SD</i> = 14.02)	5.06 (<i>SD</i> = 2.00)	41.41 (<i>SD</i> = 22.20)	–0.02 (<i>p</i> = .713)
Light bulb	523	2.52 (<i>SD</i> = 1.16)	46.00 (<i>SD</i> = 19.96)	3.42 (<i>SD</i> = 2.20)	31.10 (<i>SD</i> = 21.81)	0.02 (<i>p</i> = .712)
Preference	520	1.46 (<i>SD</i> = 1.09)	73.04 (<i>SD</i> = 10.76)	4.61 (<i>SD</i> = 2.07)	42.32 (<i>SD</i> = 22.33)	–0.14 (<i>p</i> = .001)

Note: Correlations in bold exceeded our preregistered exclusion cutoff for correlations between performance and confidence.
GOT = Generalized Overconfidence Task.

tested the consistency of confidence across tasks by measuring each pairwise correlation of confidence between tasks. We calculated these correlations for both task-level measures of confidence: score estimates and average item-level confidence (Table 2). The average correlation between average item confidence on different tasks was .63, and the average correlation between score estimates on different tasks was .42, both of which exceed our preregistered threshold (.4) as evidentiary support for overconfidence being a trait.

We also note that although actively open-minded thinking (AOT) appears to be negatively correlated with average item confidence, this pattern appears to be less consistent when correlating with score estimates. Neuroticism does not appear to consistently correlate with either task-level measure of confidence. Correlating these traits with averaged participant-level item confidence yielded similar results ($r = -0.29$, $p < .001$ with AOT; $r = -.08$, $p = .013$ with Neuroticism), as does correlating them with averaged participant-level score estimates ($r = -.16$, $p < .001$ with AOT; $r = -.13$, $p < .001$ with Neuroticism).

Factor analyses

We conducted three confirmatory factor analyses (CFAs) to investigate whether item and task scores loaded onto a common latent confidence factor: (a) a one-factor CFA using average confidence ratings across tasks (key test of the hypothesis), (b) a one-factor CFA using item-level confidence ratings, and (c) a hierarchical CFA, in

which a higher-order factor was modeled to account for seven lower-order task-specific factors, each comprising its respective items.

Traditional fit indices assessed model adequacy. However, model characteristics, such as the number of items, factors, and factor reliability (e.g., Hancock & Mueller, 2011; Heene et al., 2011), can influence the suitability of cutoff values. To address these concerns, we calculated dynamic fit indices (DFI; McNeish & Wolf, 2021). The DFI approach leverages simulations to generate cutoff values at which misfit models are detected in 95% of instances. The DFI analysis reports fit statistics across the simulated data sets and provides levels of misspecification: Level 1 indicates that approximately one third of the items exhibit residual correlations of .30 that are unaccounted for in the model, Level 2 applies to two thirds of the items, and Level 3 involves all items.

Factor loadings from each model provide direct evidence of the relationship strength between observed variables and the latent factor of confidence. We employed the robust maximum likelihood estimator, and we handled missing data with full information maximum likelihood. We reran each of the aforementioned CFA models using estimated performance and two measures of overconfidence—one based on item confidence ($\Sigma(\text{item confidence} / 100) - \text{task score}$) and one based on score estimations (estimated score – task score). In addition, to further ensure the robustness of our findings, we tested models with and without tasks

Table 2. Correlations Between Average Item Confidence and Trait Measures (Lower Diagonal) and Task-Level Score Estimates and Trait Measures (Upper Diagonal) for Each Task

Variable	Age	Birthday	Book	Card suit	GOT	Innie	Light bulb	AOT	Neuroticism
Age		0.46***	0.46***	0.50***	0.52***	0.39***	0.35***	−0.21***	−0.06
Birthday	0.68***		0.36***	0.48***	0.55***	0.44***	0.41***	−0.11*	−0.10*
Book	0.59***	0.64***		0.50***	0.39***	0.35***	0.41***	−0.12**	−0.05
Card suit	0.68***	0.77***	0.51***		0.47***	0.25***	0.37***	−0.21***	−0.12**
GOT	0.67***	0.67***	0.52***	0.69***		0.50***	0.34***	−0.09*	−0.09*
Innie	0.70***	0.73***	0.65***	0.61***	0.63***		0.37***	0.03	−0.08
Light bulb	0.62***	0.63***	0.55***	0.65***	0.60***	0.55***		−0.07	−0.11*
AOT	−0.24***	−0.32***	−0.18***	−0.39***	−0.29***	−0.26***	−0.24***		0.02
Neuroticism	−0.11*	−0.01	−0.01	−0.08	−0.09	−0.03	−0.10*	0.02	

Note: GOT = Generalized Overconfidence Task; AOT = actively open-minded thinking.

* $p < .05$. ** $p < .01$. *** $p < .001$.

that significantly correlated with actual and estimated performance scores (i.e., birth order and participant-preference tasks) and again with those who failed attention checks. We report our primary preregistered model (a one-factor CFA using average item-confidence ratings across tasks), but results for all models are available in the Supplemental Material, and the R scripts can be accessed via the Open Science Framework (<https://osf.io/8rw59/>). The other preregistered models included a one-factor CFA on which individual items loaded and a hierarchical CFA with a higher-order factor on which eight lower-order factors (for task) loaded, each comprised of their corresponding items.

The design purposely introduced planned missingness (matrix sampling): Each participant completed five of the nine tasks selected at random. After preregistered exclusion of two tasks, participants provided data on $M = 3.90$ ($SD = 0.65$) of the seven retained tasks; thus, ~44% of the task-level cells were structurally missing. Because task assignment was purely random, the missing data mechanism is missing completely at random (MCAR). We therefore estimated all CFA models with full-information maximum likelihood, which is unbiased under MCAR/MAR (Enders & Bandalos, 2001).

One-factor CFA: average task confidence

A key test of our hypothesis was a CFA that loaded all seven tasks onto a single latent factor (see Fig. 2 for a path diagram). The chi-square test was significant, $\chi^2(14) = 55.28$, $p < .001$, but given its sensitivity to sample sizes, we considered other fit indices. Both the comparative fit index (CFI = .978) and Tucker-Lewis Index (TLI = .967) exceeded the conventional threshold of 0.90. These indices suggest that the model accurately captures the underlying structure of the data.

Additionally, the root-mean-square error of approximation (RMSEA = .056, 90% CI = [.041, .072] and the standardized root-mean-square residual (SRMR = .028) further corroborate suitable fit. Although the RMSEA exceeded the traditional cutoff of .05, small misfits can inflate RMSEA values when using large samples. Notably, the lower bound of the confidence interval indicated good fit, and the upper bound remained within acceptable limits (see Table 3 for complete model statistics).

The standardized factor loadings demonstrated strong relationships with the latent confidence factor (λ s = 0.72–0.88; see Table 4). The birthday-odd task

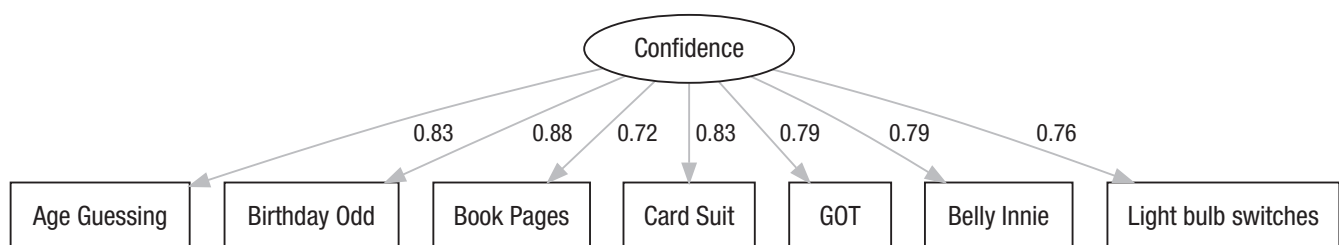


Fig. 2. Path diagram of the one-factor confirmatory factor analysis using average task-confidence scores across tasks. Values along the arrows indicate standardized loadings for each task. GOT = Generalized Overconfidence Task.

Table 3. Fit Statistics for a One-Factor Confirmatory Factor Analysis Using Average Task-Confidence Scores

χ^2	df	χ^2/df	p	CFI	TLI	RMSEA [90% CI]	SRMR	AIC	BIC
55.28	14	3.95	< .001	.98	.97	.06 [.04, .07]	.03	28,369.28	28,471.08

Note: CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion.

exhibited the highest loading ($\lambda = 0.88$), whereas the book-pages task had the lowest loading ($\lambda = 0.72$). The GOT had a relatively low loading ($\lambda = 0.79$), possibly because of familiarity: That is, reduced familiarity may result in floor effects for task confidence, weakening its relation to the latent variable of confidence. Using dynamic fit index simulations, the SRMR (.031) indicated a Level 3 misspecification, whereas the RMSEA (.056) was below the threshold for Level 1 misspecification, and the CFI (.978) fell into Level 2. These results suggest that the model demonstrated adequate fit, even under stringent criteria (see Table 5).

Discussion

Our project began as an adversarial collaboration between two research teams favoring different answers to our research question: Is overconfidence a trait? When our teams achieved agreement about a study worth doing, we preregistered our research plans and our divergent predictions for our results. Those results favor a trait interpretation. Nonetheless, there remain some elements of both agreement and disagreement. Following other adversarial collaborations (Martel et al., 2024), we elaborate on our unique perspectives below.

In an effort to determine a clear outcome of the adversarial collaboration, it appeared necessary to

establish a threshold that would be suggestive of overconfidence being a trait. Although we acknowledge that the distinction between traits and states exists on a continuum, and thus that any specific cutoff we selected would be to some extent arbitrary, the .40 benchmark offers a useful heuristic when interpreted within the broader context of psychological research. A correlation of .40 is generally regarded as a medium effect size (Cohen, 1988); Mischel (1968) argued that correlations below .40 suggest poor cross-situational consistency; and Fleeson and Gallagher (2009) noted that correlations between personality traits (e.g., extraversion and conscientiousness) and their behavioral manifestations hover around .40. Fleeson (2001) also noted that moderately-sized correlations can capture meaningful and stable individual differences in Big Five personality dimensions, such as extraversion. Applied to the present research, a .40 threshold suggests a degree of stability in overconfidence across tasks—an important indicator of trait-like structure—while also recognizing within-person variability.

The case for a confidence trait (by Team 1)

The key evidence we present for whether overconfidence is a trait is twofold. First, correlations between

Table 4. Factor Loadings for a One-Factor Confirmatory Factor Analysis Using Average Task-Confidence Scores

Task	Loadings		SE	z	p
	Unstandardized	Standardized			
Age guessing	1.00	0.83			
Birthday odd	0.95	0.88	0.05	19.90	< .001
Book pages	1.07	0.72	0.05	20.26	< .001
Card suit	1.21	0.83	0.06	20.17	< .001
GOT	0.72	0.79	0.04	17.84	< .001
Belly innie	0.94	0.79	0.04	24.47	< .001
Light-bulb switches	1.31	0.76	0.06	20.67	< .001

Note: GOT = Generalized Overconfidence Task.

Table 5. Dynamic Fit Indices for a One-Factor Confirmatory Factor Analysis Using Average Task-Confidence Scores

Level	SRMR	RMSEA	CFI
Level 1: 95/5	.019	.060	.990
Level 2: 95/5	.025	.090	.977
Level 3: 95/5	.030	.113	.965

Note: SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CFI = comparative fit index.

task-level confidence ($r = .63$) and estimated performance ($r = .42$) exceeded the predetermined threshold of .4. The second piece of evidence stems from confirmatory factor analyses, which show commonality across tasks regardless of the measure. The primary test focused on task-level confidence, which displayed unambiguous evidence via fit statistics (CFI = .978, TLI = .967, RMSEA = .056, SRMR = .031) and factor loadings ($\lambda = .72$ to .88). Further, these findings extend to estimated performance, overconfidence, or overestimation. In fact, all task-level loadings, regardless of the measure, met or exceeded the threshold predetermined to be suggestive of a trait. Given that there was variability across measures (e.g., innie overestimation: $\lambda = .40$), average loadings for overconfidence were .58 and .52 for overestimation scores. Although these values are ostensibly weaker than confidence ratings, they are nevertheless suggestive of an underlying trait.

One possible explanation for the discrepancy between correlational and CFA findings may stem from CFAs' ability to account for noise better than correlations. A CFA separates the latent factor from measurement error, but correlations exclusively focus on observed data, meaning that random noise could weaken the relationship across tasks. Given that the objective was to detect whether there is stability across a variety of tasks, separating signal from noise seems pertinent.

Overall, correlational and CFA results support the generalizability of the GOT and the suspicion that confidence, and plausibly overconfidence, does systematically vary across individuals. This supports our hypothesis that past efforts to measure overconfidence across tasks or measures have been hampered by noise that emerges from task performance. When participants are given tasks in which confidence judgments are uninfluenced by actual performance, individual differences in the tendency to overrate one's abilities emerge. Although this represents a different way of conceptualizing overconfidence, we do believe that it provides evidence that some people do tend to think they are better on a variety of tasks than they actually are.

The case against (by Team 2)

Although we concede that the results demonstrate greater consistency in confidence than we had expected, we remain skeptical that these results provide conclusive evidence that overconfidence qualifies as a durable personality trait. Our skepticism is based on features of both the present results and the published literature.

We employ unconventional operationalizations. Usually, overconfidence is measured as a difference score between confidence and accuracy (e.g., Moore & Healy, 2008). If we had employed more standard operationalizations (subtracting performance), we would have found that the average correlation between overconfidence across tasks in our results falls to .34 for item confidence and .26 for score estimates (full correlation tables are in the Supplemental Material). We recognize that this is due in part to adding in the noisiness of performance measures; nevertheless, it aligns with traditional operationalizations of overconfidence.

We are also keenly aware of how our results fit into the large published literature on overconfidence. There are different forms of overconfidence, and we focus on one: estimations of performance. The overestimation we observe is different from overplacement or overprecision (Moore & Schatz, 2017). Correlations between overestimation and other forms of overconfidence are not always reliable. Indeed, averaging within participants, our item-confidence measure ($r = .40$) is correlated with participants' reported placement (that is, reported percentile rank relative to other study participants). Questions remain about the relationship between overestimation and overprecision. Further, our measure of item confidence also favors finding higher correlations, because response biases (Gao, 2023) may masquerade as consistent responding.

Other evidence casts doubt on the consistency between different measures of a single form of overconfidence. Li et al.'s (2025) Study 3a focused on overprecision, the form of overconfidence that appears, in at least some studies (e.g., Moore & Swift, 2010), to show the most traitlike differences between people. Li et al.'s (2025) innovation was to employ four different approaches to eliciting participants' certainty about the same judgment. The four approaches were Likert-scale reports of confidence, the amount bet on being right, confidence-interval widths, and concentration in reported subjective probability distribution. The average correlation between measures was a measly .13, undercutting the notion that certainty differs between people as a reliable trait. In the present results, the correlation between our two measures of task-level confidence (item confidence and score estimates), averaged within-participants, is .53.

In sum, although we obtain evidence supporting traitlike consistency, it is possible that this consistency is only evident in the particular domains and elicitations we study. Consequently, future research should assess the generalizability (or limitations) of our results.

Reconciliation: what we agree on

Participants' confidence levels on different tasks that are similar to Binnendyk and Pennycook's GOT (2024) are related to each other in both forms of analysis—simple correlations and factor analyses. In other words, some participants are more confident than others across the seemingly random tasks we offered. These relationships provide support for a within-individual consistency and a trait interpretation of confidence and overconfidence.

Several topics are worthy of further discussion and research. First, we note that neither overestimation as calculated when subtracting scores from score estimates (average pairwise r across tasks = .26), nor overconfidence as calculated when subtracting scores from item confidence (average pairwise r across tasks = .34), displayed similar consistency. This discrepancy is due to the inclusion of accuracy, which is ingrained with how overconfidence is traditionally operationalized—an equal blend of confidence and accuracy. The believers on Team 1 suggest that fixing performance at chance provides a more reliable measure of overconfidence, because performance across tasks is entirely unrelated to estimated performance. Some prior work (e.g., Parker & Stone, 2013) has criticized the difference-score approach to measuring overconfidence. The skeptics on Team 2 would note that doing so is more consistent with traditional approaches to measuring overconfidence.

Second, the ecological validity of our paradigm and task set remains an open question. In particular, it is unclear whether the latent trait captured by this framework meaningfully predicts expressions of overconfidence in real-world contexts. A valuable direction for future research could be to evaluate the predictive validity of this paradigm using tasks that more closely approximate real-world decision-making. One such example may be an adaptation of the diagnostic reasoning paradigm developed by Sanchez and Dunning (2018), in which participants are presented with a list of symptoms and asked to determine whether the symptoms correspond to one of two diseases, or to neither. If trait overconfidence, as measured by the current paradigm, predicts performance on such diagnostic tasks, this would strengthen the claim that it is representative of real-world overconfidence. More broadly, we might

consider a similar study design that employs tasks in which confidence is related to accuracy.

Third, it would be worthwhile to conduct similar studies with different forms and measures of overconfidence. Our study relied on item-level percentage and task-level score estimates, which primarily speak to overestimation. Understanding whether overplacement and overprecision follow similar patterns would help us answer the question of whether overconfidence as a whole should be considered a trait. Relatedly, there were a couple of tasks in which we expected no association between task performance and confidence ratings (e.g., guessing birth order from a picture of a face); although we suspect that these results may have been spurious (and we have no compelling explanation for these anomalous results), further research could be devoted to determining whether (and when) task performance and confidence become unlinked.

Conclusion

Our adversarial collaboration has provided evidence supporting a trait-based interpretation of confidence and overconfidence. We found consistent relationships in confidence within individuals across several tasks in both CFAs and raw correlations. Although these relationships weaken when we employed other measures of confidence or when we adjusted for accuracy, the evidence remained relatively robust for an underlying overconfidence trait. This suggests that some people may, in fact, be more confident than others—regardless of their actual competency—across a range of tasks.

Transparency

Action Editor: Yoel Inbar

Editor: Simine Vazire

Author Contributions

Jabin Binnendyk: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing.

Sophia Li: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing.

Thomas Costello: Formal analysis; Methodology.

Randall Hale: Conceptualization; Methodology; Resources; Writing – original draft; Writing – review & editing.

Don A. Moore: Conceptualization; Investigation; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing.

Gordon Pennycook: Conceptualization; Investigation; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing.

Jabin Binnendyk and Sophia Li contributed equally to this work.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was funded by the Institute for Humane Studies at George Mason University. We would like to thank Cory Clark and the University of Pennsylvania Adversarial Collaboration Project for their assistance in obtaining funding.

Artificial Intelligence

No artificial-intelligence-assisted technologies were used in this research or the creation of this article.

Ethics

This research received approval from the ethics board at Cornell University (ID No. 148258).

Open Practices

Open practices for this article are described in the Research Transparency Statement section, which appears at the end of the Introduction section in the main text.

ORCID iDs

Jabin Binnendyk  <https://orcid.org/0000-0002-1491-6373>
 Sophia Li  <https://orcid.org/0000-0002-9165-4493>
 Thomas Costello  <https://orcid.org/0000-0002-5188-3881>
 Gordon Pennycook  <https://orcid.org/0000-0003-1344-6143>

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09567976251401538>

Notes

1. Because of a mistake, the instructed attention check and the leprechaun attention check were not required questions for the first 50 participants. Participants who did not answer either of these questions are counted in these exclusions (22 of the 32 excluded on the basis of the instructed attention check did not answer the question, and 2 of the 26 excluded on the basis of the leprechaun question did not answer it).
2. Because of a mistake, the item-confidence scale on the preference task (not included in our main analyses) was anchored from 50% to 100%, which did not align with chance (10%–100%).
3. Our preregistration designates these correlational results as our secondary test, but we report them first because their simplicity makes them more presentable and easier to understand.

References

- Acker, D., & Duck, N. W. (2008). Cross-cultural overconfidence and biased self-attribution. *Journal of Socio-Economics*, 37(5), 1815–1824.
- Amirkhanyan, H., Krawczyk, M., Wilamowski, M., & Boksztanin, P. (2024). Overconfidence: The roles of gender, public observability and incentives. *Journal of the Economic Science Association*, 10(1), 76–97. <https://doi.org/10.1007/s40881-023-00149-z>
- Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, 116(1), 261–292. <https://doi.org/10.1162/003355301556400>
- Baron, J. (1994). *Thinking and deciding* (2nd ed.). Cambridge University Press.
- Bazerman, M. H., & Moore, D. A. (2013). *Judgment in managerial decision making* (8th ed.). Wiley.
- Binnendyk, J., & Pennycook, G. (2023). *Individual differences in overconfidence: A new measurement approach* (SSRN Scholarly Paper 4563382). <https://doi.org/10.2139/ssrn.4563382>
- Binnendyk, J., & Pennycook, G. (2024). Individual differences in overconfidence: A new measurement approach. *Judgment and Decision Making*, 19, e28.
- Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making*, 17(4), 297–311. <https://doi.org/10.1002/bdm.475>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5
- Exley, C. L., & Kessler, J. B. (2022). The gender gap in self-promotion. *The Quarterly Journal of Economics*, 137(3), 1345–1381. <https://doi.org/10.1093/qje/qjac003>
- Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, 80(6), Article 6. <https://doi.org/10.1037/0022-3514.80.6.1011>
- Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, 97(6), Article 6.
- Gao, R., Wang, L., Nelson, L. (2023, November 17–20). *Extremity bias in survey responses generates strong yet invalid results* [Conference presentation abstract]. Annual Conference of the Society for Judgment and Decision Making, San Francisco, CA, United States. <https://sjdm.org/programs/2023-program.pdf>
- Heydari, F., Sheybani, S., & Yoonessi, A. (2023). Iranian emotional face database: Acquisition and validation of a stimulus set of basic facial expressions. *Behavior Research Methods*, 55(1), 143–150. <https://doi.org/10.3758/s13428-022-01812-9>
- Lackner, S., Francisco, F., Mendonça, C., Mata, A., & Gonçalves-Sá, J. (2023). Intermediate levels of scientific knowledge are associated with overconfidence and negative attitudes towards science. *Nature Human Behaviour*, 7, 1490–1501. <https://doi.org/10.1038/s41562-023-01677-8>
- Lawson, M. A., Larrick, R. P., & Soll, J. B. (2023). Individual differences in overconfidence and their psychological bases (INSEAD Working Paper No. 2025/24/DSC). SSRN. <https://doi.org/10.2139/ssrn.4558486>

- Li, S., Hale, R., & Moore, D. A. (2024). *Is overconfidence an individual difference?* <https://osf.io/tb2me/>
- Li, S., Hale, R., & Moore, D. A. (2025). Is overconfidence an individual difference? *Judgment and Decision Making*, 20, e25.
- Light, N., Fernbach, P. M., Rabb, N., Geana, M. V., & Sloman, S. A. (2022). Knowledge overconfidence is associated with anti-consensus views on controversial scientific issues. *Science Advances*, 8(29), Article eabo0038. <https://doi.org/10.1126/SCIADV.ABO0038>
- Littrell, S., Risko, E. F., & Fugelsang, J. A. (2021). 'You can't bullshit a bullshitter' (or can you?): Bullshitting frequency predicts receptivity to various types of misleading information. *British Journal of Social Psychology*, 60(4), 1484–1505. <https://doi.org/10.1111/bjso.12447>
- Lundeberg, M. A., Fox, P. W., & Punčohář, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86, 114–121. <https://doi.org/10.1037/0022-0663.86.1.114>
- 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 of the United States of America*, 118(23), Article e2019527118. <https://doi.org/10.1073/pnas.2019527118>
- Malmendier, U., & Tate, G. (2015). Behavioral CEOs: The role of managerial overconfidence. *The Journal of Economic Perspectives*, 29(4), Article 4.
- Martel, C., Rathje, S., Clark, C. J., Pennycook, G., Van Bavel, J. J., Rand, D. G., & van der Linden, S. (2024). On the efficacy of accuracy prompts across partisan lines: An adversarial collaboration. *Psychological Science*, 35(4), 435–450. <https://doi.org/10.1177/09567976241232905>
- McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*, 55(3), 1157–1174.
- Mischel, W. (1968). *Personality and assessment*. Wiley.
- Moore, D. A. (2023). Overprecision is a property of thinking systems. *Psychological Review*, 130(5), 1339–1356.
- Moore, D. A., Carter, A., & Yang, H. H. J. (2015). Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organizational Behavior and Human Decision Processes*, 131, 110–120.
- 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), Article e12331. <https://doi.org/10.1111/spc3.12331>
- Moore, D. A., & Small, D. A. (2007). Error and bias in comparative social judgment: On being both better and worse than we think we are. *Journal of Personality and Social Psychology*, 92(6), 972–989. <https://doi.org/10.1037/0022-3514.92.6.972>
- Nekby, L., Thoursie, P. S., & Vahtrik, L. (2008). Gender and self-selection into a competitive environment: Are women more overconfident than men? *Economics Letters*, 100(3), 405–407.
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, 122(3), 1067–1101.
- O'Reilly, C. A., & Hall, N. (2021). Grandiose narcissists and decision making: Impulsive, overconfident, and skeptical of experts—but seldom in doubt. *Personality and Individual Differences*, 168, Article 110280. <https://doi.org/10.1016/j.paid.2020.110280>
- Parker, A. M., & Stone, E. R. (2013). Identifying the effects of unjustified confidence versus overconfidence: Lessons learned from two analytic methods. *Journal of Behavioral Decision Making*, 27(2), 134–145. <https://doi.org/10.1002/bdm.1787>
- Peng, R. X., & Shen, F. (2024). Why fall for misinformation? Role of information processing strategies, health consciousness, and overconfidence in health literacy. *Journal of Health Psychology*, 30(8), 2030–2045. <https://doi.org/10.1177/13591053241273647>
- Pennycook, G., Binnendyk, J., & Rand, D. (2025). Overconfidently conspiratorial: Conspiracy believers are dispositionally overconfident and massively overestimate how much others agree with them. *Personality and Social Psychology Bulletin*. Advance online publication. <https://doi.org/10.1177/01461672251338358>
- Prims, J. P., & Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and Decision Making*, 12(1), 29–41.
- Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, 114(1), 10–28. <https://doi.org/10.1037/pspa0000102>
- Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117–143. <https://doi.org/10.1037/pspp0000096>
- Spiller, S. A. (2024). Widely-used measures of overconfidence are confounded with ability. *SSRN*. <https://ssrn.com/abstract=4468920>
- Vancouver, J. B., Thompson, C. M., Tischner, E. C., & Putka, D. J. (2002). Two studies examining the negative effect of self-efficacy on performance. *Journal of Applied Psychology*, 87(3), 506–516.
- Vitriol, J. A., & Marsh, J. K. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology*, 48(7), 955–969. <https://doi.org/10.1002/ejsp.2504>
- Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2017). Known unknowns: A critical determinant of confidence and calibration. *Management Science*, 63(12), 4298–4307.
- Weber, E. U., & Johnson, E. J. (2009). Decisions under uncertainty: Psychological, economic, and neuroeconomic explanations of risk preference. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics* (pp. 127–144). Elsevier.