Judgment and Decision Making (2024), **vol. 19**:e28 1–24
doi:10.1017/jdm.2024.22

EMPIRICAL ARTICLE

Individual differences in overconfidence: A new measurement approach

Jabin Binnendyk[1,2] and Gordon Pennycook[1,2,3]

[1]Department of Psychology, Cornell University, Ithaca, NY, USA; [2]Department of Psychology, University of Regina, Regina, SK, Canada and [3]Hill/Levene Schools of Business, University of Regina, Regina, SK, Canada

Corresponding author: Jabin Binnendyk; Email: jdbinnendyk@gmail.com

Received: 16 January 2024; **Revised:** 17 June 2024; **Accepted:** 17 June 2024

Keywords: overconfidence; general overconfidence; scale validation; calibration

Abstract
Overconfidence plays a role in a large number of individual decision biases and has been considered a 'meta-bias' for this reason. However, since overconfidence is measured behaviorally with respect to particular tasks (in which performance varies across individuals), it is unclear whether people generally vary in terms of their general overconfidence. We investigated this issue using a novel measure: the Generalized Overconfidence Task (GOT). The GOT is a difficult perception test that asks participants to identify objects in fuzzy ('adversarial') images. Critically, participants' estimated performance on the task is not related to their actual performance. Instead, variation in estimated performance, we argue, arises from generalized overconfidence, that is, people indicating a cognitive skill for which they have no basis. In a series of studies (total N = 1,293), the GOT was more predictive when looking at a broad range of behavioral outcomes than two other overestimation tasks (cognitive and numeracy) and did not display substantial overlap with conceptually related measures (Studies 1a and 1b). In Studies 2a and 2b, the GOT showed superior reliability in a test–retest design compared to the other overconfidence measures (i.e., cognitive and numeracy measures), particularly when collecting confidence ratings after each image and an estimated performance score. Finally, the GOT is a strong predictor of a host of behavioral outcomes, including conspiracy beliefs, bullshit receptivity, overclaiming, and the ability to discern news headlines.

1. Introduction

Overconfidence is one of the most pervasive problems in human judgment and decision-making (Moore and Schatz, 2017; Ortoleva and Snowberg, 2015; Russo and Schoemaker, 1992; Skala, 2008). The tendency to feel more confident than is justified by one's knowledge, expertise, or experience can be thought of as a meta-bias that underlies many other decision-making biases. Indeed, overconfidence can lead to flawed reasoning (Chen et al., 2015; Deaves et al., 2010; Kahneman, 2011; Ortoleva and Snowberg, 2015; Russo and Schoemaker, 1992; Soll and Klayman, 2004) or even preempt reasoning entirely (Ackerman and Thompson, 2017; Thompson et al., 2011). Applied to broader societal problems, overconfidence may play an important role when discerning the quality of information; for example, overconfidence has been linked to belief in conspiracy theories (Pennycook et al., 2022; Vitriol and Marsh, 2018), anti-scientific stances (Light et al., 2022), and susceptibility to misinformation more broadly (Lyons et al., 2021).

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Interestingly, although there is broad agreement that overconfidence poses a problem for judgment and decision-making, there is far less agreement when it comes to *individual differences* in overconfidence. That is, it is unclear whether overconfidence is primarily bound to specific domains or is a general trait that varies across people and is relatively stable over time. While the focus of the current work is primarily on overestimation—a type of overconfidence that looks at differences between estimated and actual scores—overconfidence can be captured with measures focused on overplacement and overprecision (see Moore and Schatz, 2017, for a full discussion). Regardless, each of these approaches is reliant on performance for a given task. For instance, overconfidence within an individual should be most prominent in domains where competence is the lowest and vice versa with high competency domains (Kruger and Dunning, 1999; Pennycook et al., 2017). The question remains, however: Are some people more overconfident *in general*—across different tasks and regardless of their baseline levels of competency?

1.1. Criticisms of generalized overconfidence

The notion that some individuals are more overconfident than others is intuitively appealing. Anecdotally, there seem to be some people who, regardless of the situation, appear unjustifiably confident. The claim of a general overconfidence, however, has been recently criticized. Specifically, it has been aptly stated that 'the empirical record presents particular individual differences associated with particular measures of overconfidence in particular contexts and settings and studies' (Moore and Dev, 2017, p. 3). Put differently, there is an absence of evidence supporting the claim of a general overconfidence as the field has primarily documented specific instances in which overconfidence arises. Further, situational factors have been shown to influence confidence, meaning that an individual could be overconfident in one task but underconfident in another (Glaser et al., 2005; Kelemen et al., 2000; Moore and Dev, 2017). Findings like these appear to suggest that confidence is more malleable than many intuitively believe, possibly undercutting the idea of a general overconfidence.

An example of the apparent malleability of overconfidence can be seen with task difficulty. People tend to overestimate their performance when tasks are difficult; however, when presented with an easy task or instances where success is likely, people tend to underestimate (i.e., the hard-easy effect; Burson et al., 2006; Lichtenstein and Fischhoff, 1977; Liu and Tan, 2021). For instance, people tend to underestimate the likelihood of high-probability events, such as surviving a bout of influenza (Slovic et al., 1984). Conversely, unlikely events, such as being injured in a terrorist attack, are often overestimated (Lerner et al., 2003). These findings extend to finance (e.g., stock markets; Kirchler and Maciejovsky, 2002; Liu and Tan, 2021), general knowledge questions (Burson et al., 2006), and completion times (Boltz et al., 1998; Burt and Kemp, 1994), suggesting that overconfidence is highly task specific.

Additionally, training paradigms, such as providing timely and informative feedback, have been shown to improve performance and calibration (Lichtenstein and Fischhoff, 1977). One study found that feedback on a forecasting task decreased over-forecasting and overconfidence, with effects persisting even after feedback had stopped (Niu and Harvey, 2022). Another example demonstrating the effectiveness of feedback can be seen with the weather forecasters (Phillips, 1987) and excellent calibration, likely owing to the availability of feedback. That is, the ability to examine the correctness of your predictions in a timely fashion (e.g., did it rain yesterday or not?) can greatly increase calibration suggesting overconfidence can be assuaged by competence within a given domain. Unfortunately, findings like these suggest that features such as task difficulty, *a priori* beliefs, and familiarity hinder our ability to parse out a general overconfidence.

1.2. Improving the measurement of generalized overconfidence

Despite these findings, it may nonetheless be possible that prior work has lacked the appropriate tools required to appropriately measure the general predisposition toward being overconfident. After

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

all, overconfidence appears to be a ubiquitous aspect of human behavior (Dunning, 2011; Kruger and Dunning, 1999; Pennycook et al., 2017)—it would be surprising if this general tendency to overestimate was not also subject to individual differences. Indeed, as noted, the speculation of a general overconfidence is nothing new (Klayman et al., 1999; Stanovich and West, 1997). Additionally, domains such as overclaiming (Paulhus et al., 2003)—that is, claiming knowledge of something made-up—appear to be subject to relatively stable individual differences. Although, whether overclaiming is reflective of a general overconfidence is debatable, as it may be more a matter of self-presentation (Paulhus, 2012).

In support of generalized overconfidence, recent work by Lawson et al. (2023) found that different types of overconfidence (namely, overestimation, overplacement, and overprecision) were all positively correlated (*rs* = .22 − .59), suggesting that individuals appear to be reliably overconfident. Further, it is noteworthy that Lawson et al.'s overconfidence measures were derived from several domains (i.e., geography, history, literature, and science)—all tasks that are reliant on performance scores. Additionally, Pennycook et al. (2022) found that overconfidence is robustly correlated with conspiracy beliefs and, in fact, this association was strongest for fringe claims. Overconfidence also predicted the tendency for conspiracy believers to overestimate how much others agree with them. Interestingly, across three distinct domains (i.e., cognitive, numeracy, and perceptual tasks), overconfidence was significantly correlated, albeit modestly (*rs* ranged from .18 − .39). Granted, as with the Lawson et al. (2023) work, it is unclear whether they are measuring the same underlying construct (likely owing to the reasons already discussed), it contributes to the growing evidence that people are generally overconfident across tasks and that this has consequences for outcomes of interest. While the aforementioned works hint at the possibility of a general overconfidence, a reliable approach to assess it is currently lacking.

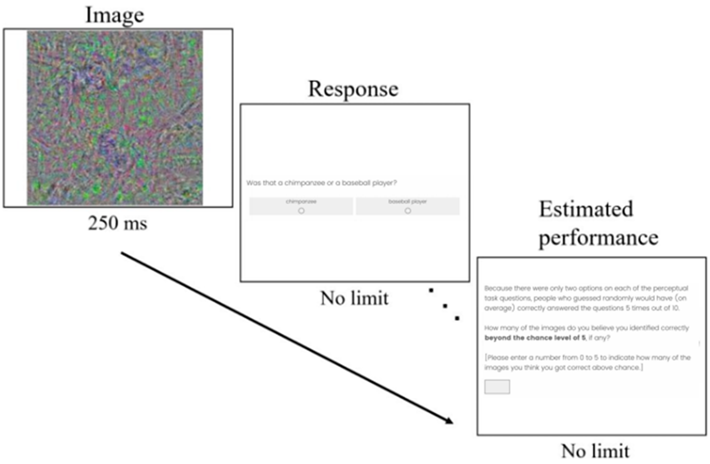
A traditional approach to measuring overconfidence would be to representatively sample items across domains (e.g., Gigerenzer, 1991); however, given the argument that task performance undermines the measurement of a general overconfidence, we advocate for an alternative approach. Specifically, we set out to measure overconfidence using a task where confidence judgments and performance are unconfounded; namely, tasks in which the participant *should* have no reasonable expectation to perform well. Or, in other words, a task where confidence is not influenced by any signals that may come from the participants monitoring their actual performance, or based on prior familiarity with the task that may lead them to form coherent reasons why they might be justifiably confident. For such a task, in theory, if some people did consistently report performing well, it would suggest a general tendency, regardless of situational factors, toward overconfidence.

To test this, we created the generalized overconfidence task (GOT), which uses novel perceptual stimuli (i.e., adversarial images; Zhou and Firestone, 2019) that are very difficult to distinguish visually. During the task, each image is displayed for a fraction of a second followed by a question that asks participants to identify what was contained within the image (e.g., a 'chimpanzee' or a 'baseball player'). Critically, after the presentation of all images, participants are asked to provide an estimated performance score above chance,[1] allowing for an overestimation score to be calculated. Due to the difficulty of the task, it is expected that participants may perform just above chance levels (Zhou and Firestone, 2019); however, more importantly, their estimated performance will be uncorrelated with their actual performance (i.e., participants are guessing). See Figure 1 for the GOT procedures.

In a series of studies, we assessed the predictive capabilities and reliability of the GOT as a measure of general overconfidence. To do this, we first investigate whether the GOT was associated with behavioral outcomes of overconfidence (Study 1a), with a specific focus on epistemically suspect beliefs. It is expected that due to the novel nature of the task, it will tap into a general overconfidence allowing us to uniquely predict these behavioral outcomes beyond any one domain-specific measure.

[1] Previous work (Pennycook, 2022) asked participants for an estimate for the number of correct answers. Most people respond zero even though chance performance on the task is five, making it difficult to know if someone who gives a high estimate is being overconfident or is aware that they have performed at chance levels. We therefore fixed all response to be estimated performance above chance. This does not appear to have an influence on the predictive validity of the measure.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press



Was that a chimpanzee or a baseball player?

chimpanzee ○ baseball player ○

Because there were only two options on each of the perceptual task questions, people who guessed randomly would have (on average) correctly answered the questions 5 times out of 10.

How many of the images do you believe you identified correctly **beyond the chance level of 5**, if any?

[Please enter a number from 0 to 5 to indicate how many of the images you think you got correct above chance.]

Figure 1. *GOT procedure. Participants were shown 10 images. Image order was randomized and estimated performance came after all 10 images/responses.*

Next, we explore whether the GOT is different from several conceptually related measures (but that are not measures of overconfidence, *per se*), such as overclaiming, that may offer a non-confidence explanation of what is being measured (Study 1b). Lastly, the reliability of the GOT is investigated using a test–retest design (Studies 2a and 2b). If overconfidence is indeed a stable individual difference, the GOT should be a reliable measure that can predict various behavioral outcomes.

2. Study 1

Studies 1a and 1b compared the GOT with two other overestimation tasks and a number of additional measures to ensure that the GOT is tapping into a unique construct. In Study 1a, we use outcomes associated with overconfidence to show that higher estimates on the GOT are associated with other overconfidence outcomes. We used three overconfidence measures—that is, the GOT, an analytic thinking task, and a numeracy task—to contrast overconfidence across measures (Pennycook et al., 2022). Notably, these overconfidence measures were focused on overestimation, that is, saying you are performing better at a task than you actually are.

In addition to testing for correlations among the overconfidence measures (i.e., the GOT, cognitive, and numeracy tasks), we set out to test whether the GOT was able to predict outcomes of interest. Specifically, we assessed correlations between the GOT and behavioral outcomes related to so-called epistemically suspect beliefs (Pennycook, 2022) that have previously been associated with overconfidence, namely, an increased susceptibility to false conspiracies (Pennycook et al., 2022; Vitriol and Marsh, 2018), misinformation (Lyons et al., 2021), and bullshit receptivity (BSR; Cavojova et al., 2022; Littrell and Fugelsang, 2024). A commonality across these outcomes is the ability to discern between low- and high-quality information, variously defined; a key benchmark for a generalized measure of overconfidence. To assess not just overall significance but also effect sizes, we also contrasted the GOT against overconfidence on the analytic thinking and numeracy tasks (as well as performance on these tasks).

Since overconfidence is considered a type of positivity bias where participants desire to portray themselves in a positive light (Bensch et al., 2019), we measured the tendency to claim familiarity about nonexistent items (i.e., 'overclaiming'; Bensch et al., 2019; Paulhus et al., 2003). We also looked at intellectual humility, which relates to whether individuals acknowledge that their view may be wrong (Bowes et al., 2023; Leman et al., 2023)—presumably a trait less common among those who are higher in generalized overconfidence (but see Costello et al., 2023).

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Next, we looked at exploratory measures of interest. Specifically, we measured intuitive-analytic thinking styles, narcissism, dogmatism (DOG), optimism, personality dimensions, self-efficacy, wishful thinking, and the illusion of control (IOC) to ensure that there was not substantial overlap between the GOT and these domains.

3. Method

3.1. Participants

Sample characteristics for Studies 1a and 1b can be found in Table 1. Individuals with missing values for any of the overconfidence measures were removed. In total, 323 and 526 participants (Study 1a and Study 1b, respectively) were recruited from Prolific, an online recruitment platform. In both studies, participants were paid commensurate to the recommended £9 per hour (~ 15 pence per minute).

3.2. Materials

3.2.1. Overconfidence measures

Three overconfidence measures were used: a general measure, a cognitive test, and a numeracy test. After the completion of each task, participants were asked to provide an estimated performance score allowing for an overestimation score to be calculated (i.e., overconfidence = estimate – actual score)[2]. The closer an overconfidence score is to 0, the more calibrated an individual is, with scores greater than 0 indicating overconfidence.

3.2.1.1. Generalized overconfidence task (GOT)

Generalized overconfidence was measured using the GOT (See Figure 1). In this task, a very difficult-to-discern image is momentarily flashed on the screen followed by a binary choice asking the participant what was depicted in the image (e.g., a chimpanzee or a baseball player; see Appendix A for all images used). This was repeated with 10 unique images (displayed in a random order), after which participants were asked to estimate their performance above chance. Importantly, the correlation between GOT estimated and actual performance indicates participants were not calibrated (i.e., guessing; $rs = -.17$ and $-.02$ for Studies 1a and 1b), even though performance on the task was above chance ($M_{accuracies} = 5.93$ and 5.92). Although the negative association between estimated and actual performance met significance in Study 1a, this correlation was not found in subsequent studies, including Study 1b.

3.2.1.2. Cognitive Reflection Test (CRT)

We administered a 6-item CRT that included reworded versions (Shenhav et al., 2012) of the original 3-item CRT (Frederick, 2005) and the 4-item non-numeric CRT problems (excluding the 'Emily' problem due to it requiring a non-numeric response; Thomson and Oppenheimer, 2016). CRT questions are intended to produce incorrect intuitive answers. For example, one question used is 'If you're

Table 1. *Sample characteristics for Studies 1a and 1b.*

Study	Source	Initial N	DNF	Removed	Final N	$\bar{x}_{\text{age}}$	N Female	N Male
1a	Prolific	323	15	6	302	33	170	115
1b	Prolific	526	21	6	499	33	249	234

Note: DNF, did not finish (including participants who opened the survey and quit immediately). Participants were removed for skipping the estimate questions. Those who failed an initial attention check at the beginning of the study were removed immediately. N for female; male does not equal the final N because some individuals indicated something other than male/female or did not answer the question.

[2] The midpoint (five) was added to the overconfidence score because we asked for an estimated score above chance.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

running a race and you pass the person in second place, what place are you in?'. In this case, the incorrect answer is first with the correct answer being second. The CRT has been argued to measure one's willingness to engage in analytic thought (Frederick, 2005; Pennycook et al., 2016), and incorrect answers are held with high confidence (Mata, 2023). Participants overestimated performance on the CRT in Study 1a (M_{actual} = 2.8, $M_{estimate}$ = 4.4) and appeared somewhat calibrated as estimated and actual scores significantly correlated (r = .35, p < .001).

3.2.1.3. Numeracy test

A 6-item numeracy and risk literacy test were used including three items from the Berlin Numeracy Test (Cokely et al., 2012) and three Lipkus numeracy task items (Lipkus et al., 2001). These questions measure basic probabilistic and mathematical competency. An example question is 'Imagine that we roll a fair, six-sided die 1000 times. Out of 1000 rolls, how many times do you think the die would come up as an even number?'. Performance on the numeracy task was modest and closely mirrored estimated performance scores (M_{actual} = 3.3, $M_{estimate}$ = 3.3), suggesting that participants were calibrated on the numeracy task. Estimated and actual performance were strongly correlated (r = .57, p < .001).

3.2.2. Outcome measures

Five outcomes associated with overconfidence, including belief in conspiracies, bullshit receptivity (BSR), news headline accuracy task, overclaiming, and intellectual humility, were used to explore predictive capabilities.

3.2.2.1. Belief in Conspiracy Theories Inventory (BCTI)

Following the recommendations of Swami et al. (2017), we measured conspiracy beliefs by asking questions about specific historical (e.g., 'The Apollo moon landings never happened and were staged in a Hollywood film studio') and contemporary (e.g., 'COVID-19 is probably a hoax'.) conspiracies. Participants were shown 12 conspiracies in total. Responses were collected on a 9-point scale ranging from '1—completely false' to '9—completely true' with '5—unsure' being the midpoint. The BCTI displayed good reliability (α = .87).

3.2.2.2. Bullshit Receptivity (BSR) scale (Pennycook et al., 2015)

Participants were asked to rate how profound five randomly generated sentences were on a 5-point scale ranging from '1—not at all profound' to '5—very profound'. An example statement is 'As you self-actualize, you will enter into infinite empathy that transcends understanding'. The BSR measures receptivity to pseudo-profound claims with higher scores meaning increased receptivity. The BSR displayed good reliability (α = .84).

3.2.2.3. News headline sharing and accuracy task

Participants were presented with 12 political or COVID-19 headlines (half true with the other half being false, in random order). The headlines were selected via pretesting (see procedure outlined in Pennycook et al., 2021). The accuracy of headlines was rated on a 4-point scale from 'not at all accurate' to 'very accurate'. Responses were dichotomized (correct or incorrect) with an accuracy score being calculated ($M_{accuracy}$ = 9.24).

3.2.2.4. The Overclaiming Technique (Paulhus et al., 2003)

Overclaiming is a type of self-enhancement technique associated with positivity bias. On this task, participants are asked to rate their familiarity with 15 items on a scale ranging from '0—never heard of it' to '6—very familiar'. However, each list contains three foils, that is, made-up items. An example foil is 'Queen Alberta' when asking about historical names and events. Accuracy was then calculated following simple signal detection theory (Accuracy = p(hits) – p(false alarms)) as suggested by its creators. For simplicity, scores were reversed so higher overclaiming scores indicate more overclaiming. We also looked at analyses using the false-alarm rate alone (i.e., instances where

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

someone indicated knowledge of a foil). Responses were coded in a binary fashion and summed to create a false-alarm score.

3.2.2.5. Intellectual Humility (IH) scale (Leary et al., 2017)

Intellectual humility refers to the ability or willingness to recognize that one's views may be wrong. The IH scale contains 6 statements, such as 'I question my own opinions, positions, and viewpoints because they could be wrong'. Each item is rated on a 5-point scale ranging from '1—not at all like me' to '5—very much like me' with higher scores indicating more IH. The IH scale displayed good scale reliability ($\alpha = .85$).

3.2.3. Additional exploratory measures

In total, eight measures were used (thinking styles, personality traits, narcissism, DOG, optimism, self-efficacy, wishful thinking, and IOC) to determine whether the GOT was adequately different from similar, non-confidence-based, domains.

3.2.3.1. 4-Component Thinking Styles Questionnaire (4-CTSQ; Newton et al., 2021)

The 4-CTSQ contains 24 items across four subscales: Actively Open-minded Thinking (AOT), Close-minded Thinking (CMT), Preference for Intuitive Thinking (PIT), and Preference for Effortful Thinking (PET). An example question from the AOT subscale is 'Whether something feels true is more important than evidence'. The 4-CTSQ uses a 6-point scale that ranges from 'strongly disagree' to 'strongly agree'. AOT and PET subscales are reverse scored allowing for an overall 4-CTSQ score to be calculated, although it is suggested to look at individual subscales since each one corresponds to a discrete thinking style. The 4-CTSQ subscales displayed good to excellent reliability ($\alpha\text{s} = .81 - .92$).

3.2.3.2. Big Five Inventory (BFI-10; Rammstedt and John, 2007)

A 10-item version of the BFI was used to measure common personality traits. The BFI-10 provides a personality measure that can be administered in a minute or less and has been shown to strongly correlate with the longer 44-item version ($r = .83$; Rammstedt and John, 2007). The BFI-10 has two items per personality dimension (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism).

3.2.3.3. Narcissistic Personality Inventory (NPI-16; Ames et al., 2006)

The NPI-16 is a shorter version of the original 40-item inventory that has demonstrated internal reliability ($\alpha = .68 - .78$), stability ($r = .85$ over 5 weeks), and discriminant validity. Further, it has been shown to strongly correlate with the original 40-item version ($r = .90$). The NPI-16 presents two statements and asks participants to choose the statement they most identify with. An example set of items is 'I am more capable than other people' and 'There is a lot that I can learn from other people', with the former response being the narcissistic one. A higher score indicates narcissistic tendencies.

3.2.3.4. Dogmatism (DOG; Altemeyer, 2002)

DOG can be described as having relatively unchangeable beliefs or being unjustifiably certain. The DOG scale contains 20 items with half being pro-trait and the other half being con-trait. An example pro-trait item is 'I am so sure I am right about the important things in life, there is no evidence that could convince me otherwise'. Items are presented on a 9-point scale ranging from 'strongly disagree' to 'strongly agree'. The DOG displayed excellent reliability ($\alpha = .90$) in the present study. Con-trait items are reverse-scored meaning that higher scores indicate dogmatic beliefs.

3.2.3.5. Life Orientation Test—Revised (LOT-R; Scheier et al., 1994)

The LOT-R is commonly used to assess dispositional optimism; however, due to its bidimensional structure, it also provides insights into pessimism. The LOT-R contains 10 items (3 optimistic, 3 pessimistic, and 4 filler statements) presented on a 5-point scale ranging from 'strongly disagree'

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

to ‘strongly agree’. Pessimism items are reverse-scored, and filler items are excluded. An example question looking at optimism is ‘In uncertain times, I usually expect the best’. Higher LOT-R scores indicate an optimistic disposition. The LOT-R displayed excellent reliability ($\alpha = .90$).

3.2.3.6. New General Self-Efficacy (NGSE; Chen et al., 2001)

Self-efficacy relates to one’s belief that one can meet the demands of a given situation (Wood and Bandura, 1989). The NGSE contains 8 items presented on a 5-point scale ranging from ‘strongly disagree’ to ‘strongly agree’, and an example statement is ‘I am confident that I can perform effectively on many different tasks’. The NGSE has been shown to contain greater construct validity than other measures of self-efficacy and exhibits high reliability (αs = .85 − .88; Chen et al., 2001). In the current study, reliability was .93.

3.2.3.7. Wishful thinking scale (WTS; Sigall et al., 2000)

WTS is described as the extent one’s cognitions impact motivations, often via the desire for a specific outcome. The WTS contains 25 items asking participants to provide likelihood ratings for an event to occur for themselves and ‘an average person’. An example item is ‘Personal achievements are described in a newspaper’. Items are rated on an 11-point scale ranging from ‘extremely unlikely’ to ‘extremely likely’. Wishful thinking is determined by taking the difference between self and other ratings (i.e., self – other = wishful thinking) with higher scores indicating more wishful thinking.

*3.2.3.8. Illusion of Control (IOC;McKenna, **1993**)*

The IOC is similar to optimism, in that one expects positive outcomes but they attribute them to aspects in one’s control. The IOC contains 12 items presented on an 11-point scale ranging from ‘much less likely’ to ‘much more likely’, with the midpoint being ‘average’. An example item is ‘Compared to the average driver, how likely do you feel you are to be involved in an accident in which the vehicle you are in skids on black ice?’. Higher scores indicate more IOC.

*3.2.3.9. The Brief Social Desirability Scale (BSDS; Haghighat, **2007**)*

The BSDS was included as a control for socially desirable answering, and it contains 4 binary response questions (yes/no). An example question is ‘Would you ever lie to people?’. Higher scores signify increased socially desirable responses.

3.2.3.10. Demographic questions

Participants were also asked a set of demographic questions, including age, gender, education, income, ethnicity, and a set of political ideology/partisanship questions.

3.2.3.10.1. Procedure. Participants first completed the overconfidence measures (GOT, CRT, and numeracy tasks). The GOT was always presented first with the order of the CRT and numeracy task being counterbalanced in Study 1a (the CRT and numeracy task were not used in Study 1b). Next, participants saw either outcome measures (Study 1a) or were presented with a subset of exploratory measures (Study 1b). Demographic questions were completed at the end of the session with attention checks throughout the survey. The first attention check occurred prior to our primary data collection, and incorrect responders were removed from the study. All data and materials (including preregistrations) are available on OSF (https://osf.io/tkmua/).

4. Results

4.1. Outcome measures

First, correlations were run across overconfidence measures (see Table 2). All overconfidence measures were significantly correlated (*rs* .14 − .32, *ps* < .05); however, the correlation between the GOT

Table 2. *Means, standard deviations, and correlations across overconfidence measures.*

Measure	*M*	*SD*	1	2	3	4
1. GOT estimate	1.79	1.66	1	.85***	.34***	.22***
2. GOT overconfidence	0.86	2.45	.79***	.86	.27***	.19**
3. Numeracy overconfidence	−0.01	1.57	.25***	.18**	.52	.53***
4. CRT overconfidence	1.62	1.84	.18**	.14*	.32***	.70

Abbreviations: CRT, Cognitive Reflection T; GOT, Generalized Overconfidence Task.
Note: M and *SD* are used to represent mean and standard deviation, respectively. Correlations above the diagonal are disattenuated, while values on the diagonal denote Guttman's lambda 6 values (reliability). A conservative value of 1 was used for the GOT estimate as it was a single-item response.
*indicates $p < .05$.
**indicates $p < .01$.
***indicates $p < .001$.

and the CRT overconfidence (r = .18, p = .001) scores was weaker than that found with numeracy overconfidence (r = .25, p < .001). This indicates some consistency across overconfidence measures. Nonetheless, these correlations were far from what would be expected if the three overconfidence outcomes were measuring the same underlying trait. Instead, it appears that the GOT is measuring something distinct from (but that is related to) the other overconfidence measures (which are distinct from each other as well). This is expected given that the task-related concerns highlight in the introduction. We, therefore, turn to our analysis of the GOT's ability to predict behavioral outcomes associated with overconfidence.

Table 3 outlines the main findings for overconfidence measures and its associated outcomes. Notably, in most instances, the GOT was equally, if not more strongly, associated with outcome predictors relative to overconfidence on the numeracy and CRT tasks. Consistent with our expectations, estimated performance on the GOT demonstrated strong positive correlations with belief in conspiracies (r = .34, p < .001), BSR (r = .34, p < .001), and overclaiming (r = .30, p < .001), along with a strong negative correlation with the ability to distinguish true and false headlines (r = −.28, p < .001). Interestingly, the GOT—despite being the result of a single performance estimate—was as predictive (if not more predictive in some cases) as actual accuracy on the CRT and numeracy tests (both measures that are highly established in the literature). Naturally, there were no such associations with performance on the GOT. In fact, the direction of many of the associations was flipped (albeit non-significantly in most cases). This leads to an additional ancillary finding: For the GOT, estimated performance is a less noisy measure than overconfidence (i.e., the difference between estimated and actual performance). Since performance on the task is entirely unrelated to estimated performance (and largely the product of guessing), subtracting actual performance from estimated performance likely only adds noise to the measure. Alternatively stated, assuming chance performance on the GOT while using estimated scores is a more reliable measure of overconfidence.

Similar findings, albeit not as strong, were found when looking at numeracy overconfidence. More specifically, numeracy overconfidence was significantly associated with belief in conspiracies (r = .13, p = .024), BSR (r = .11, p = .047), overclaiming (r = .23, p < .001), and headline discernment (r = −.23, p < .001). However, one difference was that IH significantly correlated with numeracy overconfidence (r = −.12, p = .040) but not the GOT estimate (r = −.08, p = .166). However, unlike the GOT, these correlations appear to be mostly driven by actual performance, which was a strong predictor of most outcomes (except for IH).

The pattern of results was weaker when looking at CRT overconfidence. Only headline accuracy (r = −.23, p < .001) and belief in conspiracies were significantly (r = .18, p = .002) associated with CRT overconfidence, although overclaiming was marginally significant (r = .11, p = .055). An explanation may stem from the tendency for CRT responses to be held with high confidence (Mata, 2023). In turn, the CRT may be particularly poorly suited to capture generalized overconfidence because it was created specifically to inflate confidence. Put differently, some people who would not otherwise appear

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Table 3. *Correlations (Pearson's r) between overconfidence measures and outcome predictors.*

Task	Measure	BCTI	BSR	OCQ	OCQ false alarms	Headline accuracy	IH
GOT	Estimated accuracy	.34***	.34***	.30***	.18**	−.28***	−.08
	Actual accuracy	−.08	−.08	−.12*	−.15**	.05	.08
	Overconfidence	.28***	.28***	.28***	.22***	−.22***	−.10
CRT	Estimated accuracy	−.16**	−.11	−.10	−.05	−.02	−.06
	Actual accuracy	−.32***	−.19***	−.20***	−.02	.22***	−.02
	Overconfidence	.18**	.10	.11	−.02	−.23***	−.03
Numeracy	Estimated accuracy	−.11	−.08	−.02	.08	.00	−.04
	Actual accuracy	−.24***	−.19***	−.24***	−.10	.22***	.07
	Overconfidence	.13*	.11*	.23***	.19***	−.23***	−.12*

Abbreviations: BCTI, Belief in Conspiracy Theories Inventory; BSR, bullshit receptivity; CRT, cognitive reflection test; GOT, generalized overconfidence task; IH, intellectual humility; OCQ, overclaiming questionnaire.

*** $p \leq .001$.

** $p \leq .01$.

* $p \leq .05$.

overconfident are led to overestimate their performance due to the intuitiveness of the incorrect answers on the CRT, thus undermining it as a measure of overconfidence as a trait.

To determine the predictive capabilities of the overconfidence measures, we then ran separate univariate regressions with overconfidence outcomes acting as the dependent variable and the overconfidence scores as independent variables (see Table 4). The goal of this analysis was to investigate which of the measures (overconfidence for numeracy and CRT, estimated performance for the GOT) uniquely predicted the outcomes in question. Across 4 of the 6 outcomes, the GOT estimate displayed stronger predictive capabilities than either of the other overconfidence measures: conspiracy beliefs ($\beta = .34, p < .001$), BSR ($\beta = .34, p < .001$), overclaiming ($\beta = .30, p < .001$), and headline accuracy ($\beta = -.28, p < .001$). The only instances the GOT estimate was not a stronger predictor than the other overconfidence measures were with IH and false alarms; however, only numeracy overconfidence significantly predicted IH.

In contrast, numeracy overconfidence was predictive of all the outcomes, while CRT overconfidence was predictive of less than half of the associated outcomes (more precisely, two). Although numeracy was predictive of all outcomes, it was regularly outperformed by the GOT estimate except with false alarms ($\beta = .19, p = .001$) and intellectual humility ($\beta = -.12, p = .040$). CRT was predictive of conspiracy beliefs ($\beta = .18, p = .002$) and headline veracity ($\beta = -.23, p < .001$). This inconsistency is noteworthy: Across all the outcome variables, only conspiracies and headline discernment were significantly predicted by each of the overconfidence measures. One explanation relates back to the generalizability concerns when using specific tasks. The variation between numeracy and CRT overconfidence suggests that findings from task-specific measures, while likely tapping into general overconfidence (to some extent), are impacted by individual differences in performance. Comparatively, the GOT estimate provides a broader picture of the negative consequences that can arise from dispositional overconfidence as it is not confounded by task performance. Altogether, these findings suggest that the GOT provides an alternative way to measure overconfidence.

4.2. Additional exploratory measures

Next, we examined whether the GOT could be explained by similar non-confidence-based measures (see Table 5). Importantly, there was no evidence suggesting that estimated performance on the GOT shares substantial similarity to any of the measures investigated. Across all 8 measures, the strongest correlations found were related to a lack of openness toward alternative opinions. More precisely,

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Table 4. *Univariate regression results using various outcome predictors as the criterion.*

Conspiracy beliefs

Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	0.34	[0.23, 0.44]	6.17	< .001
Numeracy overconfidence	0.13	[0.02, 0.24]	2.27	.024
CRT overconfidence	0.18	[0.06, 0.29]	3.11	.002

BSR

Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	0.34	[0.24, 0.45]	6.30	< .001
Numeracy overconfidence	0.11	[0.00, 0.23]	2.00	.047
CRT overconfidence	0.10	[–0.02, 0.21]	1.69	.092

Overclaiming

Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	0.30	[0.19, 0.41]	5.50	< .001
Numeracy overconfidence	0.23	[0.12, 0.34]	4.15.	< .001
CRT overconfidence	0.11	[0.00, 0.22]	1.93	.055

Overclaiming false alarms

Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	0.18	[0.07, 0.29]	3.11	.002
Numeracy overconfidence	0.19	[0.08, 0.31]	3.43	.001
CRT overconfidence	–0.02	[–0.14, 0.09]	–0.37	.709

Headline accuracy

Predictor	B	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	–0.28	[–0.39, –0.17]	–5.03	< .001
Numeracy overconfidence	–0.23	[–0.34, –0.12]	–4.14	< .001
CRT overconfidence	–0.23	[–0.34, –0.12]	–4.12	< .001

Intellectual humility

Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	–0.08	[–0.19, 0.03]	–1.39	.166
Numeracy overconfidence	–0.12	[–0.23, 0.01]	–2.07	.040
CRT overconfidence	–0.03	[–0.15, 0.08]	–0.55	.584

Abbreviations: CRT, cognitive reflection test; GOT estimate, generalized overconfidence.
Note: *β* indicates the standardized regression weights. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

closed-minded thinking ($r = .26, p < .001$), DOG ($r = .26, p < .001$), and actively open-minded thinking ($r = -.24, p < .001$) were the strongest correlations found with estimated performance on the GOT. Even when examining these cases, measures displayed small-to-medium (Cohen, 1988) effect sizes, leading us to conclude that the GOT was not redundant, but, rather, may simply be factors that lead

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Table 5. *Correlations (Pearson's r) with GOT and cognitive and divergent overconfidence measures.*

Measure	*GOT r's*	*t*	*P*
4-CTSQ: Actively open-minded thinking	−.24	–3.84	< .001
4-CTSQ: Close-minded thinking	.26	4.18	< .001
4-CTSQ: Preference for effortful though	−.02	–0.36	.721
4-CTSQ: Preference for intuitive though	.13	2.01	.045
BFI: Agreeableness	.12	1.90	.059
BFI: Conscientiousness	.07	1.11	.268
BFI: Extraversion	.16	2.64	.009
BFI: Neuroticism	−.16	–2.61	.010
BFI: Openness	−.01	–0.23	.820
Narcissistic Personality Inventory	.16	2.53	.012
DOG	.26	4.21	< .001
Life orientation test (optimism)	.08	1.25	.212
Need for general self-efficacy	.16	2.54	.012
WTS	.10	1.56	.119
IOC	.03	0.50	.621

Abbreviations: BFI, Big Five Inventory; DOG, Dogmatism; GOT, generalized overconfidence task estimated performance; IOC, illusion of control; 4-CTSQ, 4-Component Thinking Styles Questionnaire; WTS, wishful thinking scale.

to (and/or are impacted by) these other individual differences. Similarly, narcissism ($r = .16$, $p = .012$) and self-efficacy ($r = .16$, $p = .012$) were found to positively correlate with the GOT estimates. Lastly, extraversion ($r = .16$, $p = .009$) and neuroticism ($r = -.16$, $p = .010$), but not openness ($r = -.01$, $p = .820$) were associated with estimates on the GOT[3].

5. Study 2

Having established the predictive capability of the GOT, we turn next to reliability. For this, we assessed test–retest reliability over 15 days with the GOT, CRT, numeracy task, BSR and overclaiming. The CRT and numeracy tasks were included to allow for a direct comparison with the GOT—that is, how overconfidence measures vary in terms of reliability. The inclusion of the BSR and overclaiming scales allowed us to extend this comparison to see how well overconfidence measures performed relative to other well-established scales of correlated concepts.

Additionally, in Study 2b, we provide a replication and extension of our findings by increasing the precision of our overconfidence measurements. We did this in two ways. First, we collected a continuous confidence rating for each item (for the GOT, CRT, and numeracy tests), which allowed us to investigate whether reliability could be improved compared to using a single-item measure of overestimation. Second, we asked participants to estimate the performance of other Prolific users for each of the overconfidence measures (GOT, CRT, and numeracy task), which allowed us to calculate another form of overconfidence: overplacement.

5.1. Participants

Sample characteristics for Studies 2a and 2b can be found below (see Table 6). Individuals with missing values for any of the overconfidence measures were removed from the dataset. In Study 2a, 315 participants were recruited from Prolific for the first part of the survey. One individual was removed for indicating they were not willing to complete the second portion of the study and another for not proving

[3] Previous work by Schaefer et al. (2004) found the only personality dimension significantly related to overconfidence was extraversion ($r = .19$, $p < .05$).

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Table 6. *Sample characteristics for Studies 2a and 2b.*

Study 2a	Source	Initial N	DNF	Removed	Final N	$\bar{x}_{\text{age}}$	N Female	N Male
Test	Prolific	315	14	2	299	33	148	146
Retest	Prolific	255	3	0	252	33	123	124
Study 2b	**Source**	**Initial N**	**DNF**	**Removed**	**Final N**	$\bar{x}_{\text{age}}$	**N Female**	**N Male**
Test	Prolific	307	0	7	300	41	151	142
Retest	Prolific	241	0	0	241	42	111	123

Abbreviation: DNF, did not finish (including participants who opened the survey and quit immediately).
Note: Participants were removed for skipping the estimate questions. Those who failed an initial attention check at the beginning of the study were removed immediately. N for female; male does not equal the final *N* because some individuals indicated something other than male/female or did not answer the question.

a CRT estimate. In total, 299 participants were retained to complete the second portion of the study at a later date. After 15 days, participants were invited to complete the retest portion. This resulted in 255 individuals opening the survey. Of these, only 3 did not finish the survey and an additional individual was removed for only having completed the second portion. This resulted in 251 individuals completing both portions of the survey, yielding a high completion rate of 84%.

In Study 2b, 307 participants were recruited from Prolific for the first part of the survey. Six individuals were removed for indicating they were not willing to complete the second portion of the study and another for incorrectly answering the initial screener. In total, 300 participants were retained to complete the second portion of the study at a later date[4]. Two hundred and forty-one people opened and completed the second portion of the survey (see Table 6), resulting in an 80% completion rate (comparable to Study 2a). Participants in both studies were paid commensurate to the recommended £9 per hour (~15 pence per minute).

5.2. Materials

The GOT, CRT, numeracy task, BSR, and overclaiming measures from Study 1 were used. However, each task was randomly separated into two portions with additional items being included for the CRT, numeracy, BSR, and overclaiming to ensure balanced sets across sessions. In Study 2b, confidence scores (sliders going from 0 to 100) were collected after participants responded to each item on the GOT/CRT/numeracy tests (i.e., ‘How confident are you that you are correct?’). After completing each overconfidence task, participants provided an estimated performance score for both themselves (same as Study 1) and for other Prolific users (e.g., ‘How many of the images do you believe other Prolific users identified correctly beyond the chance level of about 2?’).

For exploratory purposes, we also included the overconfidence test (OCT) and a self-report question of overconfidence (‘How overconfident are you in how you make decisions?’) from Lawson et al. (2023) in the retest portion of Study 2b. The OCT contains three items (e.g., ‘One-hundred people are guessing the number of jellybeans in a jar. The closest 10 guesses win $100. How likely are you to be one of the winners?’) with responses ranging from 0 to 100 in 5-point increments. The OCT displayed adequate reliability (α = 0.59) in the study. Self-reported overconfidence was collected on a 0–100 slider.

5.3. Procedure

Participants first completed the overconfidence measures (GOT, CRT, and numeracy) in a randomized order. The BSR and overclaiming questionnaire came next. Demographic questions were collected at

[4] One participant provided their estimated numeracy score after completion (via direct message), which was manually entered.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

the end of the first survey only. After 15 days, participants were then invited to complete the second session which had the same structure; however, contained the remaining items not seen in session one and the OCT and self-report overconfidence question for the retest portion of Study 2b. Materials, preregistration, and data are available on OSF (https://osf.io/tkmua/).

6. Results

Correlations for the test–retest can be found in Table 7. Notably, estimated performance on the GOT significantly correlated over 15 days in Study 2a ($r = .54, p < .001$) and Study 2b ($r = .53, p < .001$), with similar findings for CRT ($rs = .58$ and $.40, p < .001$) and numeracy ($rs = .65$ and $.63, p < .001$) estimates. This highlights that there is stability in estimated performance. Following expectations, accuracy on the

Table 7. *Correlations (Pearson’s r) between sessions 1 and 2, separated by 15 days.*

Task	Measure	Study	*r*	*t*	*P*
GOT	Estimated accuracy	2a	.54	10.15	< .001
		2b	.53	9.63	< .001
	Actual accuracy	2a	.05	0.83	.408
		2b	.01	0.22	.829
	Overconfidence	2a	.35	5.83	< .001
		2b	.32	5.27	< .001
	Estimated other scores	2b	.51	9.16	< .001
	Overplacement	2b	.11	1.77	.078
	Confidence sliders	2b	.77	18.30	< .001
CRT	Estimated accuracy	2a	.58	11.18	< .001
		2b	.40	6.70	< .001
	Actual accuracy	2a	.60	11.78	< .001
		2b	.55	10.28	< .001
	Overconfidence	2a	.42	7.33	< .001
		2b	.36	5.97	< .001
	Estimated other scores	2b	.32	5.29	< .001
	Overplacement	2b	.44	7.50	< .001
	Confidence sliders	2b	.57	10.74	< .001
Numeracy	Estimated accuracy	2a	.65	13.45	< .001
		2b	.63	12.47	< .001
	Actual accuracy	2a	.42	7.21	< .001
		2b	.35	5.80	< .001
	Overconfidence	2a	.37	6.29	< .001
		2b	.18	2.84	.005
	Estimated other scores	2b	.39	6.62	< .001
	Overplacement	2b	.24	3.77	< .001
	Confidence sliders	2b	.63	12.44	< .001
BSR		2a	.73	16.81	< .001
		2b	.77	18.41	< .001
Overclaiming accuracy		2a	.72	16.17	< .001
		2b	.69	14.85	< .001
Overclaiming false alarms		2a	.51	9.25	< .001
		2b	.67	13.76	< .001

Abbreviations: BSR, bullshit receptivity; CRT, cognitive reflection test; GOT, generalized overconfidence task.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

GOT was not correlated between sessions in either study (*rs* = .01 − .05, *ps* ≥ .408), replicating findings from Study 1. However, this was not the case when looking at accuracy for the CRT (*rs* = .55 − .60, *ps* < .001) or numeracy (*rs* = .35 − .42, *ps* < .001) tasks, which were roughly as stable as performance estimates. Across each of the 4 sessions, estimated and actual scores on the GOT were non-significant (*rs* = .03 − .06, *ps* ≥ .396), suggesting that participants were unaware of their performance on the task.

Next, we examined the test–retest reliability of overconfidence scores (for estimated performance) across sessions. That is, we wanted to know whether the other overconfidence scores display similar levels of reliability as the GOT estimate. Both CRT (*rs* = .36 − .42, *ps* < .001) and numeracy (*rs* = .18 − .37, *ps* < .001) overconfidence produced weaker correlations than the GOT estimate (*rs* = .40 − .54, *ps* < .001). This was not the case when comparing the CRT and numeracy overconfidence scores to GOT overconfidence (*rs* = .32 − .35, *ps* < .001), however. Given the lack of calibration on the GOT, computing a GOT overconfidence score appears counterproductive as it adds noise and reduces the reliability of the measure (as noted in Study 1). Using a post hoc Steiger's (1980) test, the GOT estimate manifested a stronger correlation than numeracy overconfidence in Study 2a (*z* = 2.4, *p* = .020) and Study 2b (*z* = 4.45, *p* < .001), but produced mixed (albeit marginally significant) differences when contrasted with CRT overconfidence (Study 2a: *z* = 1.74, *p* = .08; Study 2b: *z* = 2.33, *p* = .02). Altogether, these findings suggest that the GOT estimate is a more reliable measure of overconfidence than the other measures examined.

We then focused on contrasting the test–retest reliability of the GOT with the 2 established measures (i.e., BSR and overclaiming). In line with our expectations, the BSR (*rs* = .73 − .77, *ps* < .001) and overclaiming (*rs* = .69 − .72, *ps* < .001) measures produced larger test–retest correlations in both studies than any of the overconfidence measures, including the GOT estimate. Another post hoc Steiger's (1980) test suggested that the difference between the GOT estimate and overclaiming was statistically different in Study 2a (*z* = 3.4, *p* < .001) and Study 2b (*z* = 2.61, *p* < .001), as well when looking at the BSR (Study 2a: *z* = 3.6 *p* < .001; Study 2b: 4.43, *p* < .001). Even when participants were asked to report their level of confidence along with an estimate of other participants' performance (i.e., Study 2b), these findings remained consistent. While the GOT estimate outperformed the other traditional measures of overconfidence examined, there is still a clear discrepancy between reliability when compared to measures that are not (ostensibly) performance-based. It is noteworthy, however, that estimated performance performs quite well despite being measured with a single item at each time point.

Post hoc tests were conducted to see whether combining overconfidence scores across sessions in Study 2a would bolster their associations with BSR and overclaiming. In all instances, the strength of the correlations increased when using a combined score as opposed to any single session (see Table 8). For instance, the combined GOT estimate was more strongly correlated with BSR (*r* = .34) than the single-session scores (*r* = .25, .30). The pattern was similar for overclaiming (combined GOT estimate: *r* = .38, individual sessions: *r* = .34 and .31). These findings suggest that the inclusion of multiple data points for overconfidence measures may enhance reliability.

Next, to test whether the reliability of the GOT would be increased by multiple confidence measurements, confidence ratings (0-100 sliders) after each trial were collected in Study 2b. These confidence ratings were collected for each overconfidence measure (i.e., GOT, CRT, and numeracy) and were aggregated across subjects for each task to create a confidence score. Confidence was found to be quite reliable for the GOT (*r* = .77, *p* < .001), CRT (*r* = .57, *p* > .001), and numeracy (*r* = .63, *p* < .001) tasks. Notably, test–retest reliability for GOT confidence met or exceeded benchmarks set by the BSR (*rs* = .73 − .77, *ps* < .001), overclaiming (*rs* = .69 − .72, *ps* < .001), and false alarms (*rs* = .51 − .66, *ps* < .001). Moreover, post hoc analyses suggest that GOT confidence was comparable to GOT estimates when looking at correlations with BSR (*rs* = .48 and .48, *ps* < .001), overclaiming (*rs* = .40 and .45, *ps* < .001), and false alarms (*rs* = .48 and .42, *ps* < .001). To determine whether there was any additional benefit of combining GOT confidence and the GOT estimate, we applied a z-score transformation to GOT confidence and GOT estimate, which was then averaged. Across all instances, except numeracy overconfidence, the combined GOT measure produced stronger correlations than

Table 8. *Correlations with overconfidence and outcomes of overconfidence.*

Variable	1	2	3	4	5	6	7	8	9	10
1. Combined GOT	.97	.91***	.89***	.53***	.45***	.47***	.29***	.31***	.42***	.30***
2. GOT item confidence	.87***	.94	.54***	.50***	.41***	.46***	.36***	.29***	.38***	.24***
3. GOT estimate	.87***	.52***	1	.43***	.38***	.37***	.16*	.25***	.36***	.27***
4. BSR	.51***	.47***	.42***	.94	.60***	.45***	.14	.18**	.43***	.24***
5. Overclaiming	.44***	.39***	.38***	.58***	.99	.75***	.13	.12	.48***	.23***
6. False alarms	.46***	.44***	.36***	.42***	.73***	.96	.27***	.18**	.42***	.21***
7. BNT overconfidence	.23***	.27***	.12	.11	.10	.21**	.63	1***	.71***	.38***
8. BNT overplacement	.30***	.28***	.25***	.17**	.12	.17**	.79***	1	.51***	.44***
9. CRT overconfidence	.26***	.23***	.23***	.27***	.30***	.26***	.36***	.32***	.40	1***
10. CRT overplacement	.29***	.24***	.27***	.23***	.23***	.21**	.30***	.44***	.84***	1

Note: M and *SD* are used to represent mean and standard deviation, respectively. Correlations above the diagonal are disattenuated, while values on the diagonal denote Guttman's lambda 6 values (reliability). A conservative value of 1 was used for the GOT estimate, BNT overplacement, and CRT overplacement as it includes single-item responses. The combined GOT reliability was an average between the GOT item confidence and GOT estimate reliabilities. Any corrected correlation that was greater than 1 was reported as 1.

*indicates $p < .05$.

**indicates $p < .01$.

***indicates $p < .001$.

Table 9. *Univariate regression results using various outcome predictors as the criterion.*

		BSR		
Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	0.37	[0.25, 0.49]	6.11	< .001
Overconfidence test	0.33	[0.21, 0.45]	5.35	< .001
		Overclaiming		
Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	0.37	[0.25, 0.48]	6.09	< .001
OCT	0.34	[0.22, 0.46]	5.63	< .001
		False alarms		
Predictor	*β*	*β* 95% CI [LL, UL]	*t*	*p*
Generalized overconfidence	0.34	[0.22, 0.46]	5.60	< .001
OCT	0.46	[0.34, 0.57]	7.94	< .001

Abbreviations: GOT estimate; generalized overconfidence; OCT, overconfidence test.

Note: Only scores from Study 2b retest for all measures were used as the OCT was not included in any other study. *β* indicates the standardized regression weights. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

either the GOT estimate or GOT confidence alone (see Table 8). Given the improved reliability and predictive capabilities, it is recommended that the GOT is administered with both confidence ratings and estimated performance scores (see Appendix B).

Test–retest reliability for overplacement was comparable to their overestimation counterparts for numeracy ($r = .24$, $p < .001$) and CRT ($r = .44$, $p < .001$); however, this was not the case for the GOT ($r = .11$, $p = .087$). This is not surprising though as overplacement tends to be a noisier measure of overconfidence—and this may have been magnified by the low familiarity of the GOT task. For instance, Griffin and Tversky (1992) showed that accuracy was poorer when predicting the performance

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Table 10. *Mean and standard deviations of estimated performance on the GOT for party and gender broken down by comparison group.*

Variable	Comparison groups	*M*	*SD*
GOT estimate	Sample	2.02	1.71
Party	Republican	2.38	1.71
	Democrats	2.00	1.88
Gender	Male	2.19	1.79
	Female	1.88	1.63

of others compared to themselves (68% versus 81%). However, the GOT estimate was significantly related to numeracy ($r = .30, p < .001$) and CRT ($r = .32, p < .001$) overplacement.

Similar findings were found when looking at the OCT and overplacement scores on the CRT ($r = .32$, $p < .001$) and numeracy test ($r = .30, p < .001$). In fact, the GOT estimate and OCT strongly correlated with each other ($r = .42, p < .001$) and displayed similar predictive capabilities when looking at the BSR, overclaiming, and false alarms (see Table 9 for univariate regressions).

Self-report overconfidence was also significant with overconfidence measures (*rs* range from .19 to .32, including overplacement scores (*rs* = .15 − .28). This may suggest that people are partially aware of their general tendency toward overconfidence, albeit to a diminished extent. Further, correlations between self-reported overconfidence with outcomes, such as BSR ($r = .25, p < .001$), overclaiming ($r = .18, p = .005$), and false alarms ($r = .24, p < .001$), although highly significant, were substantially weaker than the GOT estimate or OCT.

7. Associations between overconfidence and demographic factors

Lastly, we conducted a post hoc meta-analysis of our samples to explore the connection between general overconfidence and various demographic variables. Combining the data across all four studies resulted in 1,291 observations, which allowed us to estimate effects with high statistical power for the following demographic variables: age, gender, education, income, party affiliation, economic conservatism, social conservatism, and attention checks. Attention checks were included in the meta-analyses since it seems plausible that overconfident individuals could be less attentive, undermining the efficacy of interventions aimed at reducing overconfidence (Table 10).

Using a random intercept model, 5 of the 8 demographic variables examined were significantly related to general overconfidence (via GOT estimates). More precisely, overconfident individuals tend to be less educated ($\beta = -.11, p = .003$), Republican ($\beta = .32, p = .033$), and more conservative both in terms of social ($\beta = .21, p < .001$) and economic ($\beta = .10, p = .014$) issues. Females ($\beta = -.28$, $p = .003$) were less overconfident. Figure 2 provides an overview of the standardized beta values across all demographics for the random intercept model.

8. General discussion

Understanding the tendency to be overconfident holds the potential to inform us about errors in human judgment and decision-making. However, the existence of a general overconfidence has been critiqued. We found that the generalized overconfidence task (GOT), an alternative way to measure overconfidence that is unconfounded by task performance, was successful in predicting a broad range of behavioral outcomes, including conspiracy beliefs, BSR, overclaiming, and the ability to discern news headlines. Indeed, effect sizes for these correlations were generally as strong (and in some

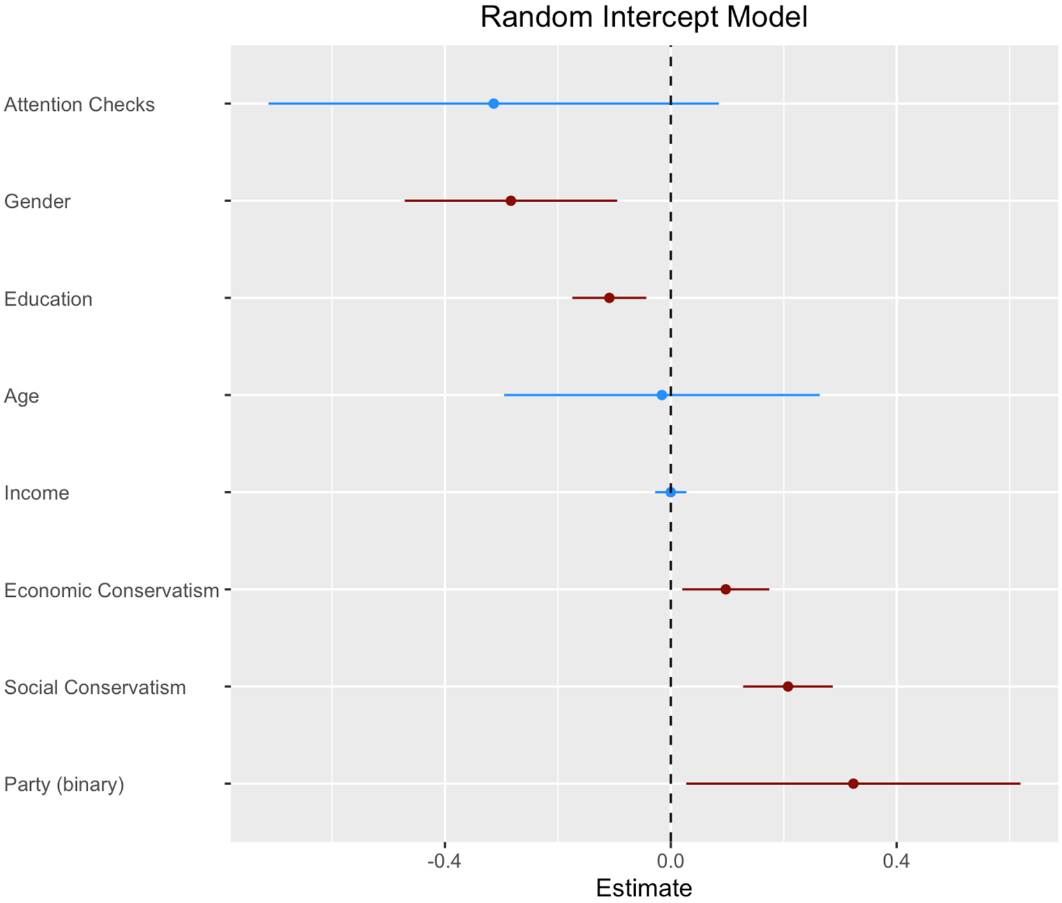


Figure 2. *Meta-analytic standardized beta values. Red estimates and 95% CIs indicate statistically significant results (p < .05).*

cases stronger) than for established performance-based measures, such as the CRT and numeracy (Pennycook, 2023; Pennycook et al., 2015; Peters, 2012; Reyna and Brainerd, 2023). The GOT was also more predictive of these outcomes than other overconfidence tasks (i.e., CRT and numeracy). In fact, CRT and numeracy overconfidence were predictive of less than half of the outcomes, while the GOT was significant in nearly all instances (with the exception of intellectual humility and false alarms; however, see also Costello et al., 2023). A plausible explanation for why the GOT accounts for a variety of domains stems back to concerns around performance (e.g., task difficulty, *a priori* beliefs, and familiarity) that are inherent with other overconfidence measures. Even within the present work, discrepancies between the predictive nature of CRT and numeracy overconfidence were found showing that the GOT did a better job than these other performance-based tasks. Altogether, the GOT provides an alternative means to measure overconfidence and suggests that a general overconfidence can, in fact, be measured.

A key strength of the GOT is that it is consistent with a particular conceptual definition of overconfidence: Participants who indicate being good at the task are not, in fact, any better at it than those who indicate a low competency on the task. This tendency is not driven by differences in performance monitoring or *a priori* beliefs (e.g., based on prior experiences) about the task; rather, those who indicate being good at the task likely do so because they just generally think they are good at cognitive tasks (i.e., they are overconfident in a more general sense). Nonetheless, a valid concern

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

is whether the GOT fully captures general overconfidence. Overconfidence has been defined in various ways, with a particular focus on 3 specific measurement approaches: overestimation, overplacement, and overprecision (Moore and Schatz, 2017). The current work primarily focuses on overestimation; however, estimated performance on the GOT was significantly correlated with CRT and numeracy overplacement and displayed substantial overlap with the OCT (Lawson et al., 2023) which involves 3 items aimed at capturing a core overconfidence. Furthermore, continuous confidence ratings across trials on the GOT were just as predictive as performance estimates (and, in fact, had stronger test–retest reliability; likely owing to the fact that the same question is asked several times). Collectively, the data indicate meaningful consistency in measures conceptually related to a general overconfidence.

Another related issue is that the GOT may be tapping into an overconfidence that is specific to *perception* (given the nature of the task) that coincidentally captures a broad pattern of behaviors associated with overconfidence or other latent abilities. Even though participants are unaware of their performance on the task, this does not preclude them from applying *a priori* beliefs they perceive to be relevant when completing the GOT. Unescapably, it may not be feasible to fully address the role of previous experiences and beliefs even on a task (and stimulus) as novel as the adversarial images that we used. Nonetheless, we contend that the GOT provides a substantial improvement. Moreover, while the present work demonstrates that the GOT is predictive across several domains, it may not encapsulate *all* aspects of overconfidence. For instance, our focus was on 'epistemically suspect beliefs', but other areas, such as political extremism (Ortoleva and Snowberg, 2015) or anti-science beliefs (Light et al., 2022; Motta et al., 2018), would enhance our understanding of the breadth of overconfidence captured by the GOT. Exploring other sorts of tasks where familiarity and calibration are low would help expand our understanding as further work is required to validate the GOT as a measure of general overconfidence and how to interpret its findings.

Further, the test–retest reliability of a single-point estimate over 15 days is suggestive of an enduring nature of overconfidence, as one would suspect with a general overconfidence. Even with a single-item measure (i.e., estimated performance on the GOT), test–retest reliability was adequate and similar to that found with emotional and self-regulation (Beauchamp et al., 2017; Enkavi et al., 2019). Furthermore, aggregated confidence ratings met or exceeded other validated measures, showing that multiple responses greatly improved reliability. Yet, further work is needed to further develop and optimize the GOT and other measures that follow from this approach.

Normally, the presence of individual differences in task performance is accounted for by subtracting their scores from estimated performance. Although the GOT removes individual effects associated with ability on a given task, individual variation for the slope of the reverse calibration curve (i.e., subjective probability estimates as a function of objective probabilities; see Erev et al., 1994) remains unaccounted for. That is, it is unclear the proportion that the GOT captures a general overconfidence versus individual differences on the reverse calibration curve. For instance, if people differ in regression away from extreme values, it could produce differences in overconfidence on the GOT. Future work may focus on fixing performance at a higher level, as well at chance levels, to examine individual differences on the reverse calibration curve.

9. Conclusion

Do individual differences in overconfidence exist? We found that some individuals consistently overestimate their abilities, even when presented with a nonsensical task. Specifically, those who were overconfident were more likely to endorse epistemically suspect beliefs and displayed stable levels of confidence over 15 days. This has implications when considering the development of interventions aimed at stemming negative behavioral outcomes of overconfidence, including epistemically suspect beliefs. It is plausible that the efficacy of these interventions will be directly affected by the malleability (or rigidity) of one's general tendency to be overconfident. Regardless, even in the absence of skill, it appears that being unaware is a hallmark of overconfidence.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Acknowledgments. This research was supported by grants from the Social Sciences and Humanities Research Council of Canada and the John Templeton Foundation.

Competing interest. The authors declare none.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/J.TICS.2017.05.004.

Altemeyer, B. (2002). Dogmatic behavior among students: Testing a new measure of dogmatism. *The Journal of Social Psychology*, *142*(6), 713–721.

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality*, *40*(4), 440–450.

Beauchamp, J. P., Cesarini, D., & Johannesson, M. (2017). The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and uncertainty*, *54*, 203–237.

Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2019). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment*, *26*(3), 351–363.

Boltz, M. G., Kupperman, C., & Dunne, J. (1998). The role of learning in remembered duration. *Memory & Cognition*, *26*, 903–921.

Bowes, S. M., Ringwood, A., & Tasimi, A. (2023). Is intellectual humility related to more accuracy and less overconfidence? *The Journal of Positive Psychology*, 538–553.

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

Burt, C. D., & Kemp, S. (1994). Construction of activity duration and time management potential. *Applied Cognitive Psychology*, *8*(2), 155–168.

Cavojova, V., Šrol, J., & Brezina, I. (2022). Why people overestimate their bullshit detection abilities: Interplay of cognitive factors, self-esteem, and dark traits.

Chen, G., Crossland, C., & Luo, S. (2015). Making the same mistake all over again: CEO overconfidence and corporate resistance to corrective feedback. *Strategic Management Journal*, *36*(10), 1513–1535. https://doi.org/10.1002/smj.2291.

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, *4*(1), 62–83.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Lawrence Erlbaum.

Cokely, E., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*, 25–47.

Costello, T. H., Newton, C., Lin, H., & Pennycook, G. (2023). A metacognitive blindspot in intellectual humility measures. https://osf.io/preprints/psyarxiv/gux95.

Deaves, R., Lüders, E., & Schröder, M. (2010). The dynamics of overconfidence: Evidence from stock market forecasters. *Journal of Economic Behavior & Organization*, *75*(3), 402–412.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology*, *44*, 249–286. https://doi.org/10.1016/B978-0-12-385522-0.00005-6.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477.

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

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases.". *European Review of Social Psychology*. 83–115. doi:10.1080/14792779143000033.

Glaser, M., Langer, T., & Weber, M. (2005). Overconfidence of professionals and lay men: Individual differences within and between tasks?. https://madoc.bib.uni-mannheim.de/2646/1/dp05_25.pdf.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435.

Haghighat, R. (2007). The development of the brief social desirability scale (BSDS). *Europe's Journal of Psychology*, *3*(4), 10–5964.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*, 92–107.

Kirchler, E., & Maciejovsky, B. (2002). Simultaneous over-and underconfidence: Evidence from experimental asset markets. *Journal of Risk and Uncertainty*, *25*, 65–85.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

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

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

Lawson, A., Larrick, R. P., & Soll, J. B. (2023). Forms of overconfidence: Reconciling divergent levels with consistent individual differences. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4558486.

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, *43*(6), 793–813. https://doi.org/10.1177/0146167217697695.

Leman, J., Kurinec, C., & Rowatt, W. (2023). Overconfident and unaware: Intellectual humility and the calibration of metacognition. *The Journal of Positive Psychology*, *18*(1), 178–196.

Lerner, J. S., Gonzalez, R. M., Small, D. A., & Fischhoff, B. (2003). Effects of fear and anger on perceived risks of terrorism: A national field experiment. *14*(2), 144–150. https://doi.org/10.1111/1467-9280.01433

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183.

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), eabo0038.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *21*(1), 37–44. https://doi.org/10.1177/0272989X0102100105.

Littrell, S., & Fugelsang, J. A. (2024). Bullshit blind spots: The roles of miscalibration and information processing in bullshit detection. *Thinking & Reasoning*, *30*(1), 49–78.

Liu, B., & Tan, M. (2021). Overconfidence and forecast accuracy: An experimental investigation on the hard–easy effect. *Studies in Economics and Finance*, *38*(3), 601–618.

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.

Mata, A. (2023). Overconfidence in the cognitive reflection test: Comparing confidence resolution for reasoning vs. general knowledge. *Journal of Intelligence*, *11*(5), 81.

McKenna, F. P. (1993). It won't happen to me: Unrealistic optimism or illusion of control? *British Journal of Psychology*, *84*(1), 39–50.

Moore, D. A., & Dev, A. S. (2017). Individual differences in overconfidence. *Encyclopedia of Personality and Individual Differences*. https://learnmoore.org/mooredata/EPID.pdf

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*(8), 1–12. https://doi.org/10.1111/spc3.12331.

Motta, M., Callaghan, T., & Sylvester, S. (2018). Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine*, *211*, 274–281.

Newton, C., Feeney, J., & Pennycook, G. (2021). On the disposition to think analytically: Four distinct intuitive-analytic thinking styles, 906–923. http://doi.org/10.31234/Osf.Io/R5wez.

Niu, X., & Harvey, N. (2022). Outcome feedback reduces over-forecasting of inflation and overconfidence in forecasts. *Judgment and Decision Making*, *17*(1), 124–163.

Ortoleva, P., & Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, *105*(2), 504–535.

Paulhus, D. L. (2012). Overclaiming on personality questionnaires. https://psycnet.apa.org/record/2014-20300-010.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904. https://doi.org/10.1037/0022-3514.84.4.890.

Pennycook, G. (2022). A framework for understanding reasoning errors: From fake news to climate change and beyond.

Pennycook, G. (2023). A framework for understanding reasoning errors: From fake news to climate change and beyond. In *Advances in experimental social psychology* (Vol. 67, pp. 131–208). Academic Press. https://doi.org/10.1016/bs.aesp.2022.11.003.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*(1), 25293.

Pennycook, G., Binnendyk, J., & Rand, D. (2022). Overconfidently conspiratorial: Conspiracy believers are dispositionally overconfident and massively overestimate how much others agree with them. https://osf.io/preprints/psyarxiv/d5fz2.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*(6), 549–563.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, *48*, 341–348. https://doi.org/10.3758/s13428-015-0576-1.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, *24*(6), 425–432. https://doi.org/10.1177/0963721415604610.

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, *24*, 1774–1784. https://doi.org/10.3758/s13423-017-1242-7.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31–35.

Phillips, L. D. (1987). On the adequacy of judgmental forecasts. In G. Wright & P. Ayton (Eds.), Judgmental forecasting (pp. 11–30). John Wiley & Sons.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212.

Reyna, V. F., & Brainerd, C. J. (2023). Numeracy, gist, literal thinking and the value of nothing in decision making. *Nature Reviews Psychology*, *2*(7), 421–439.

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, *33*(2), 7–17.

Schaefer, P. S., Williams, C. C., Goodie, A. S., & Campbell, W. K. (2004). Overconfidence and the big five. *Journal of Research in Personality*, *38*(5), 473–480.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, *67*(6), 1063.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology. General*, *141*(3), 423–428. https://doi.org/10.1037/a0025391.

Sigall, H., Kruglanski, A., & Fyock, J. (2000). Wishful thinking and procrastination. *Journal of Social Behavior and Personality*, *15*(5), 283–296.

Skala, D. (2008). Overconfidence in psychology and finance-an interdisciplinary literature review. *Bank I Kredyt*, *4*, 33–50.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1984). Behavioral decision theory perspectives on risk and safety. *Acta Psychologica*, *56*(1–3), 183–203.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342–357. https://doi.org/10.1037/0022-0663.89.2.342.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245.

Swami, V., Barron, D., Weis, L., Voracek, M., Stieger, S., & Furnham, A. (2017). An examination of the factorial and convergent validity of four measures of conspiracist ideation, with recommendations for researchers. *PLOS ONE*, *12*(2), e0172617. https://doi.org/10.1371/JOURNAL.PONE.0172617.

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113.

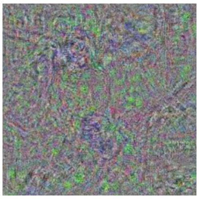
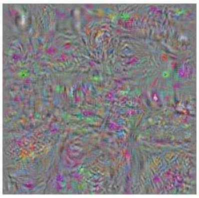
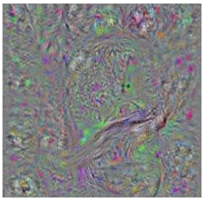
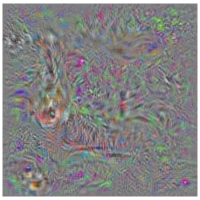
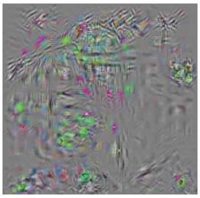
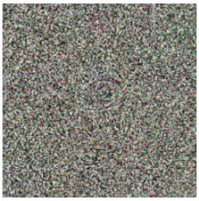
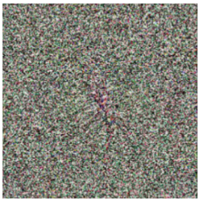
Vitriol, J. A., & Marsh, J. K. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology*, *48*(7), 955–969. https://doi.org/10.1002/ejsp.2504.

Wood, R., & Bandura, A. (1989). Impact of conceptions of ability on self-regulatory mechanisms and complex decision making. *Journal of Personality and Social Psychology*, *56*(3), 407.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1–9.

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Appendix A. GOT Images and Estimate

1)  Was that a chimpanzee or a baseball player?	2)  Was that a fire station or a confectionary?
3)  Was that an eel or a tiger shark?	4)  Was that a horse or a golden retriever?
5) Was that a parking meter or a stop sign?	6)  Was that a barn or a greenhouse?
7) Was that an armadillo or a basketball?	8) Was that an apple or a robin?
9)  Was that a mug or a bubble?	10)  Was that a centipede or a crayon?

https://doi.org/10.1017/jdm.2024.22 Published online by Cambridge University Press

Estimated performance:

‘Because there were only two options on each of the perceptual task questions, people who guessed randomly would have (on average) correctly answered the questions 5 times out of 10.’

How many of the images do you believe you identified correctly beyond the chance level of 5, if any?

[Please enter a number from 0 to 5 to indicate how many of the images you think you got correct above chance.]

Appendix B. GOT Recommended Usage

The recommended version of the GOT can be accessed via the OSF link (https://osf.io/tkmua/) and by downloading the qsf file labeled ‘GOT’. This version uses both confidence ratings and an estimated performance score. When conducting analyses, it is suggested to first z-score both the aggregated confidence ratings across trials and performance estimate and then use the mean of these values (i.e., mean of zConfidence rating and zEstimate).

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