The Behavioural Insights Team
Working Paper Series

The UK is overconfident

Working Paper No. 007 - August 2024

Dr. Mark Egan

THE
BEHAVIOURAL
INSIGHTS
TEAM

Abstract

When you test people's knowledge and check how confident they are in their answers, they tend to come out as 'overconfident' - more confident than correct, on average. This is the typical result whether you examine ordinary people or experts, across a range of topics, using a range of methods - most people are simply poorly calibrated about the extent of their own knowledge. However, the vast majority of this research has been conducted in North America, often using samples of university students. This paper contributes to this literature by reporting comprehensive calibration results for a representative sample of 2,024 UK adults, one of the largest and most detailed datasets collected of its type. Respondents got a list of 30 general knowledge questions, and were asked to say whether each was true or false and how certain they were about their answer. On average the sample had a confidence level of 83% (where 100% means total certainty that one's answer is correct) but only got 68% correct. This 15pp gap represents significant overconfidence, and is in line with past studies. Older and less educated respondents were more likely to be overconfident, but there were no notable differences by gender.

Author information

mark.egan@bi.team is a Principal Research Advisor at The Behavioural Insights Team

Notes

- The views expressed herein are those of the author and do not necessarily reflect the views of the Behavioural Insights Team.
- BIT working papers are circulated for discussion and comment purposes. They have been reviewed internally following BIT's quality assurance guidelines. This review is conducted at various stages of the project by senior researchers and advisors external to the project team who hold expertise in the topic/ research area.
-
- To cite this paper: Egan, M. (2024). The UK public is overconfident (BIT Working Paper Number 006).

1. Introduction

> “Real knowledge is to know the extent of one's ignorance.”
> - Confucious
>
> “Ignorance more frequently begets confidence than does knowledge.”
> - Charles Darwin

If you measure
(i) whether a person knows the correct answer to something, and
(ii) how confident they are that their answer is correct,
then you can also examine the difference between those two things.

If a person’s confidence exceeds their correctness, that is overconfidence. If a person is more correct than confident, that’s underconfidence. And if a person is about as correct as they are confident in their knowledge, such that these two things tend to track each other fairly reliably across a range of topics, then they are ‘well-calibrated’. When well-calibrated people say there is an 80% chance of something happening or something being true, they tend to be correct 80% of the time. An overconfident person might be 100% sure about something and still turn out to be wrong. Poor calibration has been found to contribute to worse project management (Buehler, Griffin, & Ross, 1994) and investment performance (Barber & Odeon, 2001), and inaccurate medical diagnoses Christensen-Szalanski & Bushyhead, 1981) and political forecasts (Moore et al, 2017).

Being well-calibrated turns out to be surprisingly rare - as we will see from decades of past research, and the results of a novel dataset from a representative sample of 2,024 UK adults, among the largest and most detailed datasets ever reported of this nature. This short paper provides new evidence on the phenomenon, and discusses insights on potential solutions to the problem.

2. What we know about calibration

A review of several hundred studies about cognitive calibration or ‘metaknowledge’, published in 2000 in the Journal of Consumer Research (Alba & Hutchinson, 2000), found widespread overconfidence - people think they know more than they actually do, with the gap often being ~15 percentage points (e.g. 85% confident vs 70% correct). This tendency towards overconfidence has been documented across many domains, including general knowledge, memory for events, predictions about the future, and assessments of one's own abilities. The finding was also robust to measuring calibration by, for example, (i) comparing average correctness and average confidence across a range of questions, or (ii) having participants provide a range within which they believed the true value of something lies (e.g. a 90% confidence interval estimate of how much revenue Microsoft made in 2023).

Most studies in that review described the results of ‘lab experiments’ - carefully controlled, artificial environments set up by the academic authors and typically using university student participants. Another 2002 review covered over 100 studies which focused more on real-world decision-making by experts in five domains: medicine, meteorology, law, business,

and sports (Koehler, Brenner, & Griffin, 2002). Even these experts often displayed systematic miscalibration in routine assessments they did as part of their everyday activities. There were exceptions. Economists and weather forecasters, who tended to have the most training and technical assistance in statistical modelling approaches, tended to be better calibrated than physicians, stockbrokers, and sports commentators. In other words, when weather forecasters were highly confident it would rain, it usually did. In contrast, lawyers and doctors often overestimated their chances of winning a case or accurately diagnosing a disease. Expert bridge players have also been documented as having nearly perfect calibration (Keren, 1987).

Subsequent research has confirmed the ubiquity of overconfidence, even among many experts. Tetlock (2005) found that political forecasters' predictions were only slightly better than chance and often overconfident. Lin & Bier (2008) examined over 5,000 real-world estimates by experts from 27 technical panels covering topics from assessing the toxicity of different chemicals to the risks of space debris, cranes, and volcanos and found that only 58% of true values fell within the experts' 90% probability intervals, indicating substantial overconfidence. A 2021 study asked a group of 164 experts (e.g. epidemiologists, statisticians, clinicians) in April 2020 to predict the number of coronavirus infections and deaths that would occur by December 2020, then examined how accurate these predictions turned out to be. Although the experts were more accurate than the public, they were still overconfident (Recchia, Freeman, & Spiegelhalter, 2021).

A smaller subset of studies has identified groups with excellent or near-perfect calibration. Mellers et al. (2016) found that 'superforecasters' had essentially perfect calibration over two years, such that they were capable of accurately distinguishing between fine-grained probabilities (e.g. a 62% or 67% chance of an event occurring). These individuals tended to have high crystallised intelligence, a high need for cognition, open-mindedness, and 'scope-sensitivity' (i.e., good at fine-tuning assessments based on new information). These characteristics suggest avenues for training others to become better-calibrated, and short training modules have indeed shown promising results (Chang, Chen, Mellers, & Tetlock, 2016).

Lastly, although most studies in this literature use the broad term ‘overconfidence’, others have identified three specific variations of it (Moore & Healy, 2008). ‘Over-estimation’ is the tendency to over-predict one’s level of performance or ability, e.g. a person who thinks they got 90% of items correct on a quiz but in reality got 50%. ‘Over-placement’ is incorrectly ranking your performance as better than others, e.g. thinking you’re in the top 10% in a group but are really below-average. ‘Over-precision’ is excessively narrow certainty, e.g. thinking a project will take 12-14 days to complete when it ends up taking 25.

One limitation of these past decades of research is that the vast majority of it has been conducted in North America, often using samples of university students. This geographical and demographic bias raises questions about the universality of these findings across different cultures and populations, particularly given that, as shown in Stankov & Lee (2014), overconfidence can vary notably by country. This study therefore seeks to expand the

UK-specific evidence, using the largest and most detailed survey reported to date for that country.

3. Data and methodology

This paper's contribution to this literature is to present new evidence on the calibration performance of the UK general population. The study involved a representative sample of 2,024 UK adults, collected in June 2024 using Predictiv, the in-house online survey & experiments platform of The Behavioural Insights Team.

The data were collected via a simple online survey. Participants were recruited online through our established panel networks. They were randomly assigned to see one of two lists (List A or B), each of which contained 30 statements of comparable difficulty. The randomisation element was used to test a greater number of questions in a limited amount of time, not to test for differences between List A and List B. For each statement, participants were asked:
(i) whether they thought it was true or false, and
(ii) how sure they were in your answer, on a scale of 50% ("*I have no idea whether my answer is right or wrong*") to 100% ("*I'm totally sure that my answer is correct"*).

Overall calibration was measured by calculating the (i) proportion of questions which participants answered correctly, and (ii) their average level of confidence across all their answers, and then examining the difference between these two numbers. For example, a person might answer the questions with 90% confidence on average, but only get 70% correct. They would be 20pp more confident than correct, representing significant overconfidence. A person who answered with only 65% average confidence, but who got 80% correct, would be underconfident - more correct than confident. A person who was about as correct as they were confident, e.g. 73% correct and 75% confident, would be well-calibrated.

Example statements tested included:
(i) "*Will Smith won the 2022 Oscar for Best Actor.*"
- 54% of respondents correctly said this was true, with an average confidence of 74% (20pp more confident than correct, representing significant overconfidence).

(ii) "*Bangalore is a city in Australia.*"
- 93% of respondents correctly said this was false with 89% average confidence (confidence & correctness within 5pp of each other, representing good calibration).

(iii) "*Benedict Arnold was the name of a United States President*".
- 88% of respondents correctly said this was false with 78% confidence (confidence 10pp lower than correctness, representing underconfidence).

After answering the 30 calibration items, participants answered questions about their demographic background. Recruitment quotas were used to ensure the sample was approximately representative of the UK adult population; the key proportions for the final sample are shown in Table 1. An attention check was used to ensure participants were engaging appropriately with the survey - participants who failed this were excluded from the

final sample. The survey had no time limit. Participants could in theory have searched for the answers, but there was no meaningful incentive for them to do so. They were paid a modest fee for completing the survey, and the payment did not vary depending on whether they got more answers correct.

Table 1. Demographic characteristics of the sample (n=2,024)

Gender		Age		Ethnicity	
Female	50.1%	18-24	9.5%	Asian	6.6%
Male	49.7%	25-54	46.1%	Black	4.7%
Other	<1%	55+	44.4%	Other	2.7
				White	86%

4. Results

The median time to complete the whole survey was 8 minutes. Table 2 shows the key results. On average, across the whole sample and in line with the broader literature, respondents were significantly overconfident - about 15 percentage points more confident in their answers than they were actually correct. This is about the same average effect size found in past studies. Average correctness was 5pp lower in List A (65.5%) vs List B (70.2%), but average confidence was about the same (82.7% vs 83.3%).

Table 2. Key calibration results

	Total (n = 2,024)	**List A** (n = 1,062)	**List B** (n = 962)
% average correctness	67.7%	65.5%	70.2%
% average confidence	83.0%	82.7%	83.3%
Calibration score	**15.2pp**	**17.1pp**	**13.1pp**

Participants were randomly assigned to see either List A or List B, both of which contained 30 items. Calibration score = confidence - correctness. Positive scores imply overconfidence, negative scores imply underconfidence.

Figure 1 shows the distribution of scores across the entire sample. The best performers are those with a score of 0, or close to it. These are people who had a tight correspondence between average correctness and average confidence across the 30 items they answered, e.g. someone who got 70% of the answers correct with an average confidence of 70%.

Figure 1. Distribution of calibration scores (n=2,024)

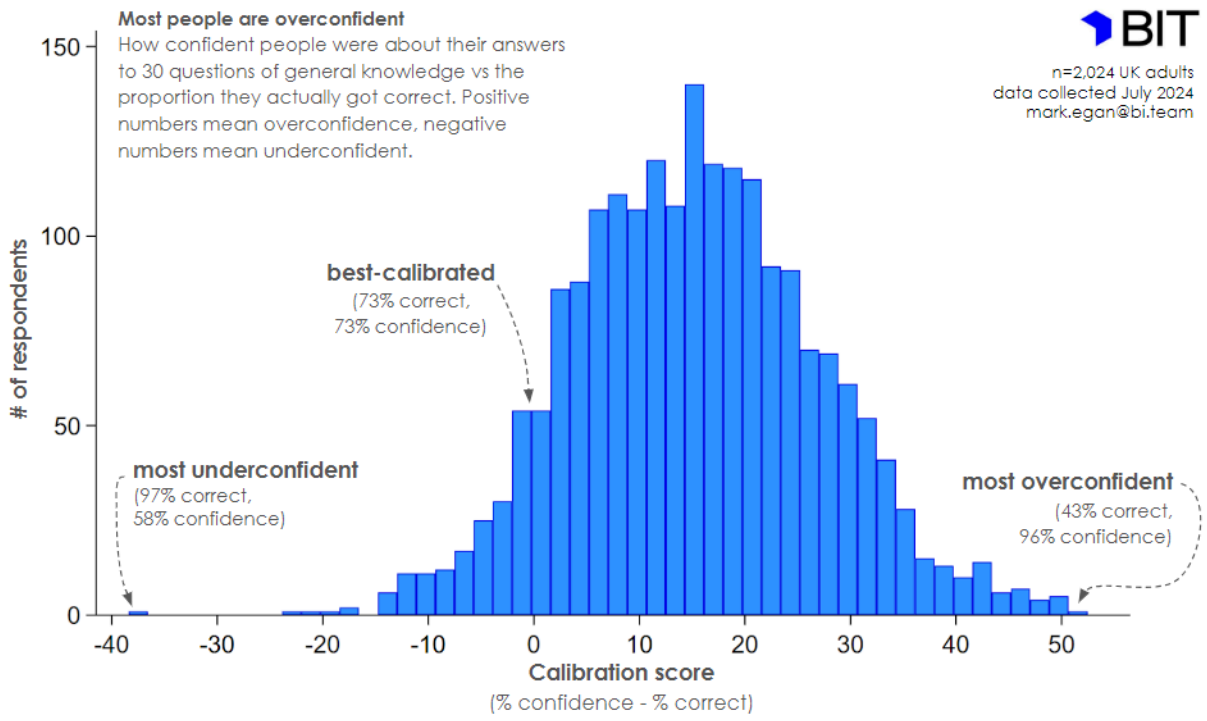

Table 3 shows the proportion of respondents by 'calibration category', with key demographic breakdowns. Across the whole sample,
- 81% were overconfident (average confidence at least 5pp higher than average correctness)
- 16% were well-calibrated (confidence and correctness within 5pp of each other)
- 3% were underconfident (confidence at least 5pp lower than correctness)

Even when using a stricter definition of overconfidence (average confidence at least 15pp higher than average correctness), a majority of respondents (51%) still came out as overconfident. There was striking variation by age - those aged 55+ were 11pp more likely to fall into the overconfident category than those aged 18-24 (difference between those two groups significant at $p<0.01$). This was over twice as large as ~5pp difference between the most- and least-educated respondents (difference between those two groups significant at $p<0.05$). Perhaps surprisingly, there was no notable difference by gender.

Table 3. Proportion of respondents in each calibration category.

	Overconfident	Well-calibrated	Underconfident
Total	80.8%	15.8%	3.5%
Age			
18-24	73.1%	22.3%	4.7%
25-54	79.3%	16.7%	4.0%
55+	84.0%	13.4%	2.7%
Gender			
Male	80.2%	16.7%	3.1%
Female	81.5%	14.7%	3.9%
Education			
O-levels or below	83.6%	12.6%	3.8%
A-levels or further qualifications	80.6%	16.1%	3.3%
Degree	78.9%	17.8%	3.3%

Overconfident: average confidence scores at least 5pp higher than average correctness,
Well-calibrated: average confidence scores within 5pp of average correctness,
Underconfident: average confidence scores at least 5pp lower than average correctness.

5. Conclusion

Our study adds to the substantial literature on overconfidence by documenting the phenomenon in a large, representative sample of UK adults. We find a striking correlation between overconfidence and age, with older groups being more likely to be overconfident. This relationship has been observed in some previous studies (Prims & Moore, 2017), but its causes remain unclear. It could be due to age-, period- or cohort-effects, or a combination.

Depending on how overconfidence is defined, between 5 and 10 (strict definition) and 8 in 10 (less strict) of respondents fell into the category. But what do those results really mean? They are based on people's responses to a low-stakes online quiz, completed by most in just a few minutes. If we could peer into the lives of these respondents, would we observe them routinely making catastrophically overconfident errors of judgement? Probably not. In real-life situations, especially for those with significant consequences like choosing a career, or saving for retirement, or deciding whether to have children, people would likely engage in more deliberate thinking and seek advice from others. That said, overconfidence has been found to persist even in professional settings with organisational safeguards (Meikle, Tenney,

& Moore, 2016), and in life-and-death situations like aviation (Shappell & Wiegmann, 2003) and pandemic management (Halpern, 2024).

Good calibration is rare - only 16% of our sample achieved it - but there is evidence that even brief training can improve it (Chang et al, 2016). Past research also suggests that environments providing systematic, regular, and unambiguous performance feedback can help foster better calibrated judgements. Structured approaches to group discussion can also help (Silver, Mellers, & Tetlock, 2021). These insights could be applied to inform the design of processes and incentives within organisations, which should in turn nudge people towards better calibration. Perhaps more broadly, cultures should reconsider the common assessment of confidence as something to universally strive for and encourage - confidence is valuable, but ideally should be tethered to an underlying, well-calibrated assessment of reality.

References

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

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1), 261-292.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, 67(3), 366–381.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting. *Judgment and Decision Making, 11*(5), 509–526.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4), 928–935.

Halpern, D. (2024). Witness statement of David Halpern. UK COVID-19 Inquiry, paragraph 73. Retrieved from https://covid19.public-inquiry.uk/wp-content/uploads/2023/11/01172222/INQ000188738.pdf

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, 39(1), 98–114.

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715). Cambridge University Press

Lin, S.-W., & Bier, V. M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety*, 93(5), 711-721.

Meikle, N. L., Tenney, E. R., & Moore, D. A. (2016). Overconfidence at work: Does overconfidence survive the checks and balances of organizational life? *Research in Organisational Behavior*, 36, 121-134.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3), 267–281.

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

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, 63(11), 3552-3565.

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

Recchia, G., Freeman, A. L. J., & Spiegelhalter, D. (2021). How well did experts and laypeople forecast the size of the COVID-19 pandemic? *PLOS ONE*, 16(5), e0250935.

Shappell, S. A., & Wiegmann, D. A. (2003). A human error analysis of general aviation controlled flight into the terrain accidents occurring between 1990-1998. Washington, DC: Federal Aviation Administration. Retrieved from https://commons.erau.edu/ publication/1217

Silver, I., Mellers, B. A., & Tetlock, P. E. (2021). Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology*, 96, Article 104157.

Stankov, L., & Lee, J. (2014). Overconfidence across world regions. *Journal of Cross-Cultural Psychology*, 45(5), 821–837.

Tetlock, P. E. (2005). Expert political judgment: How good is it? How can we know? Princeton University Press.