A tool for longitudinal assessment of individual differences in judgment calibration during category learning

Warren Woodrich Pettine[1,2,a], Drew E. Winters[3], John D. Murray[2,4], Alan Anticevic[2], and Suma Jacob[5]

[1] University of Utah, Department of Psychiatry & Huntsman Mental Health Institute, Salt Lake City, UT
[2]Yale School of Medicine, Department of Psychiatry, New Haven, CT
[3]University of Colorado School of Medicine, Department of Psychiatry, Aurora, CO
[4]Dartmouth College, Department of Psychological and Brain Sciences, Hanover, NH
[5]University of Minnesota, Department of Psychiatry and Behavioral Sciences, Minneapolis, MN

[a]Direct correspondence to warren.pettine@hsc.utah.edu

Abstract

Accurately judging the quality of one's estimates is a critical psychological process that varies across individuals, fluctuates over time, and is highly domain specific. While there are many methods for assessing judgment calibration during basic sensory tasks and tests of general knowledge, the field lacks a rigorous, controlled tool for assessing this form of metacognitive monitoring during category learning or goal-directed attention. This is especially important given the applicability of category learning and goal-directed attention to both high-level functioning and psychiatric disease. We therefore developed and tested a tool that pairs an established category learning paradigm with a standard metacognitive report, and validated the tool through longitudinal testing. A total of 80 adult subjects completed three sequential sessions online and answered a set of personality and psychiatric trait questionnaires. Test-retest results suggested that our tool provides fair reliability in assessing judgment calibration. We found aloofness and introversion are associated with improved metacognitive resolution. Furthermore, the best calibrated subjects were more likely to report narrowing their attention to a subset of goal-directed stimulus features. Our results suggested specific ways to improve the tools retest stability, while validating its use for longitudinal assessment of individual differences in judgment calibration during category learning.

Introduction

When a doctor informs their patient of potential cancer, the news is accompanied by how confident they are in the diagnosis. The degree to which their estimate corresponds to the true probability of an outcome is known as "judgment calibration." A physician is "well calibrated" if when they estimate a 90% probability of cancer, cancer is diagnosed 90% of the time. Judgment calibration is a type of metacognitive monitoring, which broadly refers to an individual's capacity for introspection on their decision process (Fleming & Frith, 2014). Assessments of judgment calibration are useful for professions such as medicine or national intelligence, education and psychiatry; yet, metacognitive abilities are highly domain-specific. That is, an individual's judgment calibration for basic sensory perception is not necessarily the same as their judgment calibration for questions involving higher-level learning and attention. One form of metacognition might be impaired, while another is intact. Furthermore, assessments of judgment calibration may not be consistent across time. Metacognitive abilities can be influenced by transient states, such as depression (Faissner et al., 2018), and measurement of metacognition is highly sensitive to overall performance on the assessment task (Guggenmos, 2021; Rahnev, 2023). Thus, it is important to have scalable tools with controllable difficulty levels that can reliably assess domain-specific metacognition across multiple points of time.

Given the importance of metacognition for general functioning, it is no surprise that metacognitive capacities vary across individuals, and can be altered in psychiatric conditions. Metacognition has been widely investigated in conditions such as obsessive-compulsive disorder (OCD), schizophrenia, addiction, depression/anxiety, autism spectrum disorder (ASD), and across developmental stages (K. L. Carpenter & Williams, 2023; Hoven et al., 2019; Norman et al., 2019; Seow et al., 2021; Sun et al., 2017). Furthermore, there is a strong focus on metacognition as it relates to the work of professions such as intelligence analysts (Atanasov et al., 2017; Friedman & Zeckhauser, 2018; Grossmann et al., 2023; Mellers et al., 2014; Mellers, Stone, Atanasov, et al., 2015; Mellers, Stone, Murray, et al., 2015; Rieber, 2004). Studies either assess metacognition through self-report questionnaires, or measure it directly through controlled experiments. While self-report questionnaires on metacognition are convenient to deploy and score, the field has largely moved towards controlled experiments where metacognition is directly assessed through confidence reports on specific decisions.

There are several key design choices when implementing these kinds of controlled metacognition experiments. Two of the primary choices pertain to the prompt modality, and the reporting structure (Sandberg et al., 2010). Metacognitive monitoring is modality and setting specific. That is, depending on the type of question being asked, different brain areas and cognitive mechanisms may be recruited (Arkes, 1991; Fleming et al., 2012; Fleming & Dolan, 2012; Fleming & Frith, 2014; Seow et al., 2021). It is thus no surprise that interventions to improve metacognition are largely domain-specific (Callender et al., 2016; J. Carpenter et al., 2019; Lichtenstein & Fischhoff, 1980; Russo & Schoemaker, 1992). One must therefore carefully tailor the structure of metacognitive inquiry to the goal of the study.

First, there is the modality of the prompt. When focused on judgment calibration, studies often use prompts that are close to what human subjects will encounter in the real world, such as factual questions about history (Coane & Umanath, 2021; Lichtenstein et al., 1982; Lichtenstein & Fischhoff, 1980). While ethological, these questions are highly uncontrolled, and do not easily generalize across individuals or groups. On the opposite end, one can use basic stimulus properties as prompts, such as the contrast between two patches. With these, it is possible to carefully calibrate the difficulty of the discrimination and reduce the need for learning. Indeed, due to their controlled nature, these prompt modalities are often used in studies that more generally investigate metacognition and its neurobiological basis. Yet, it is questionable whether prompts using basic stimulus properties generalize to metacognition in other domains. Thus, many studies try to minimize factors external to the task setting, while maximizing relevance for the function of interest. For example, in a study of social metacognition in ASD, subjects were asked to judge the emotional state of a face (McMahon et al., 2016). That allowed investigators to control the presentation of facial features, targeting their assessment of metacognitive monitoring.

Next, reporting of metacognitive judgments can either occur simultaneously with the decision, or occur sequential to it. Simultaneous reports include a wager, or simply reporting the probability of an outcome. For example, in a weighted coin flip, how much does one bet that the coin will land on heads? A large advantage of these report methods is they don't rely on verbal systems, which facilitates cross-species studies (Joo et al., 2021; Kepecs & Mainen, 2012; Lak et al., 2014; Moreira et al., 2018). However, they conflate the decision process with the metacognitive judgment of the decision process. Thus, one can never be sure if one is studying variance in the ability to make a choice (e.g. to pick heads or tails and know the difference), or variance in the ability to judge the quality of a choice (e.g. to choose the one that is more probabilistically likely during this round). Indeed, human metacognitive biases are best captured by models where the decision distribution then informs a metacognitive inference distribution (Fleming & Daw, 2017; Pouget et al., 2016; Sanders et al., 2016). For these reasons, human studies often use a sequential process where the subject first reports a decision, then makes a confidence judgment on the probability they were correct. Those judgments are converted to probabilities, and judgments are assessed using metrics such as the Brier Score, meta-d', gamma and others (Fleming, 2017; Fleming & Lau, 2014; Maniscalco & Lau, 2014; Rahnev, 2023). The relative merit has been extensively discussed, but they largely trade off between flexibility and precision.

Finally, it is important to consider the assessment's longitudinal stability over multiple sessions. Any good system of measurement should provide the same answer when a stable system is measured over multiple points in time. If the answer is not consistent across measurements it can either indicate a change in the measured system or a problem with the measurement tool. One must thus include longitudinal assessments when assessing any measurements, especially those in fields with high response variability, such as psychology and psychiatry. Only then can an investigator be confident that changes in metacognition are due to a state shift (e.g., depression), or an improvement caused by interventions.

We sought to develop a flexible tool for assessing judgment calibration associated with attention and learning. In doing so, we looked to strike a balance between the controllability of task elements, flexibility in architecture and ethological validity of the question domain. Furthermore, we looked to create a tool that can easily be gamified and deployed in scalable settings. In this work, we introduce an online tool that combines category learning and goal-directed attention with metacognitive assessment over multiple sessions. We establish that metacognitive elements of the tool correlate with key trait measures, and we identify specific ways in which the tool can be improved. This work is highly relevant for those looking to investigate the psychological and psychiatric implications of metacognition in high-level cognitive functions, and those developing interventions to improve metacognition.

Methods

Tool implementation

A game for longitudinal assessment of judgment calibration

We designed an online tool for the longitudinal testing of judgment calibration. To do this, we combined a well-established category learning paradigm (Bowman et al., 2020; Bowman & Zeithamova, 2018; Gorlick et al., 2015; Gorlick & Maddox, 2013; Wu et al., 2020; Wu & Fu, 2021; Zhou et al., 2019, 2020) with a standard metacognitive confidence report (Fleming & Lau, 2014) and made specific modifications to facilitate the tracking of goal-directed attention. The category learning framework has several advantages as a basis for metacognitive assessment. One can train subjects on categories during a block before session, thereby reducing the effect of outside knowledge. Next, given the arbitrary nature of the stimuli, one can modulate them to control for difficulty. Furthermore, strategies associated with category learning (prototype, exemplar, rule-based, etc.) have been extensively studied, as have the underlying cognitive systems (Ashby & Maddox, 2005; Pettine et al., 2023; Radulescu et al., 2021).

The category learning paradigm we adapted involves stimuli defined by ten clear visual features, each of which can assume two values. There are two prototypes (A and B) that assume opposite values for each of the ten features. Stimuli of intermediate difficulty are created by taking one of the prototypes and changing a subset of features to that of the other prototype. For example, one could copy prototype A and change a single feature to be that of prototype B, creating a "feature distance" of one. Thus, one can modulate the difficulty level of stimuli. In the classic experiment, there is a training block with feedback followed by a testing block without feedback. Each stimulus is shown only once with, the exception being the prototypes which are shown during the learning block and again during the testing block. During the training block, two training stimuli differed from their prototype on one feature, three differed on two features, three differed on three features and two differed on four features. During the testing block six stimuli differed from their prototype by one feature, seven stimuli by two features, seven by three features, and four by six features. The task paradigm offers a great deal of flexibility. For example, one could increase the number of possible feature values beyond two and then create additional categories (e.g., prototypes C and D). Or, one could easily increase the number of examples and control their difficulty by modulating the feature distance.

For our tool, we made three modifications to the task. First, we changed the stimulus paradigm from cartoons of animals to a "hub and spoke" design with features located at the end of the spokes (Fig. 1A). This design: 1) facilitates eye tracking studies of goal-directed feature attention; 2) reduces influence of semantic feature definitions (e.g., are legs a collection of lines or a body part?); 3) allows for features with continuous morphing between values; and 4) permits stimuli to easily be created programmatically (see the supplemental code package). Second, we added a metacognitive confidence report after categorization during the test block (Fig. 1B). Subjects reported their confidence in four ranges (50-62%, 63-74%, 75-87% and 88-100%). Third, subjects were incentivized by a financial bonus based on the combined quality of their classification performance and judgment calibration (see the "Scoring Rule" section below). Other than these changes, we maintained task structure to ensure consistency with prior studies.

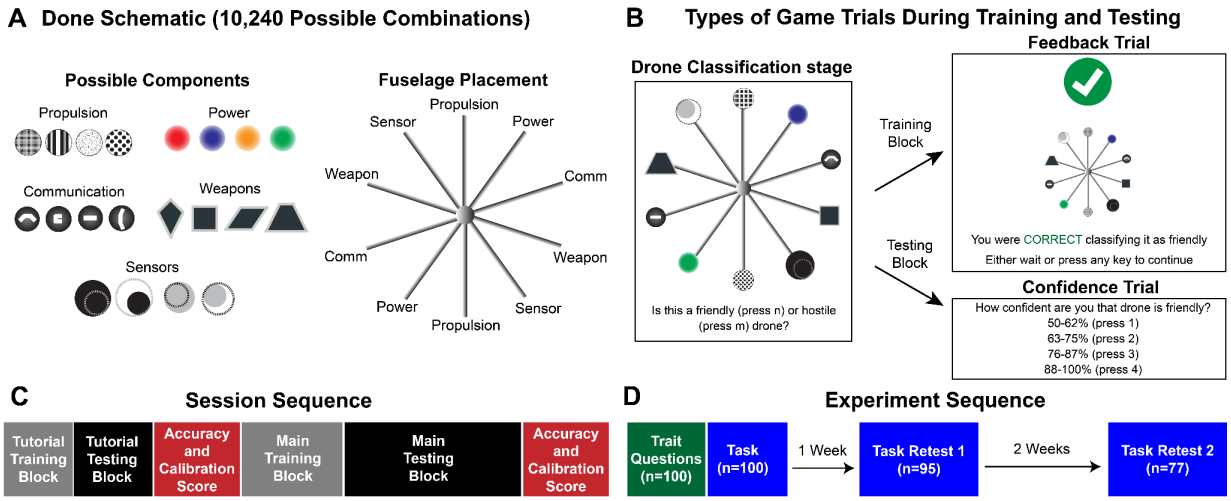

Figure 1. **Development of an online tool for assessing judgment calibration**. (**A**) Schematic of a drone used in categorization. Drone components were used as features. They were placed at the end of spokes in different configurations. (**B**) Subjects were presented with a drone to classify and then either received feedback on their choice or reported their decision confidence. (**C**) Each session was composed of several blocks. (**D**) Subjects participated in three sessions. Prior to the first session, they answered an extensive set of clinical trait questions.

The task was presented as a game where subjects are tasked with classifying "drones" for a fictional intelligence agency. The drones were defined by features at 10 locations, each of which could assume four possible values (Fig. 1A, 10,240 possible drone configurations). For consistency with past literature, only two versions of the features were possible at each location. There were two classes of drones ("friendly"

or “hostile”). One configuration was designated as prototype A (“friendly”), and a version with inverse feature values was designated as prototype B (“hostile”). A working version of the task is available in the supplemental code package.

Measuring judgment calibration

There are many measures that have been developed to assess judgment calibration and metacognition. One of the oldest of these is the Brier score (Brier, 1950). It quantifies the difference between the estimated outcome and the actual observed outcome. That is, if one considers all the instances where one assigned a 70% probability of occurring, did 70% occur? The Brier Score can be decomposed into calibration (how closely the assignment and actual percentages match), outcome variance (did they just guess 50% on each trial?) and resolution (did they provide a wide range of estimates). While more recently developed methods based on receiver operator characteristic curves are less sensitive to overall performance, they are also less flexible and less intuitive. We wanted to maintain the ability to present subjects with more than two options, or to deploy interventions that provide subjects with feedback on calibration. The Brier score and its decomposition met those criteria.

In equations, the Brier score is often denoted by the “probability score” (PS) and calculated as,

$$PS = \frac{1}{N}\sum_{i=1}^{N}(f_i - c_i)^2,$$

where N is the total number of trials, i is the current trial, f is the confidence estimate and c_i is the trial outcome (1 for correct, 0 for incorrect). It is similar to the mean squared error, in that the lower the PS, the better the calibration.

One can then break the Brier score into its components with the equation,

$$PS = O + C - R.$$

In this equation O is the outcome variance, C the calibration and R the resolution.

The outcome variance is calculated using,

$$O = \bar{c}(1 - \bar{c}),$$

where $\bar{c}$ represents the mean outcome, or accuracy. It is maximal at 0.5 (when the outcome is highly variable) and minimum at 0 or 1 (when the outcome is consistent). The lower the outcome variance the better.

Calibration is computed using the equation,

$$C = \frac{1}{N}\sum_{j=1}^{J} N_j(f_j - \bar{c_j})^2,$$

Where j indexes the confidence bin, f_j the subject’s confidence, $\bar{c_j}$ the mean outcome for trials in that bin and N_j the number of trials where that was estimated. In the case of a task with four responses, there are four possible confidence levels, and so four bins. The lower the calibration the better.

The resolution measures the utilization of different confidence estimates and is calculated using the equation,

$$R = \frac{1}{N}\sum_{j=1}^{J} N_j(\bar{c}_j - \bar{c})^2 \quad .$$

If the subject always responds at the same confidence level, they will have a low resolution, whereas if they report a variety of confidence levels they will have a high resolution. A higher resolution is better.

Scoring rule

Incentivizing serious participation by subjects is important for any behavioral study, particularly those conducted online. Given the difficulty of the task, we wanted to encourage subjects to try and to provide honest judgment of confidence. Thus, it was important for subjects to have a clear standard by which their performance is judged. The literature on proper scoring rules is rich with different scoring methods and their effect on subject performance (Gneiting & Raftery, 2007; Johnstone et al., 2011; Merkle & Steyvers, 2013). For our initial tool development, we prioritized feedback legibility. Thus, we scored subjects using a simple average of their performance and their normalized calibration ($\hat{C}$).

$$Score = \frac{\bar{c} + \hat{C}}{2} \quad ,$$

where $\hat{C}$ is normalized on a 0 to 1 (worst to best) scale using,

$$\hat{C} = 1 - (C/C_{ceiling}) \quad ,$$

and $C_{ceiling}$ is the highest calibration score that could be produced by a specific subject's response distribution.

Session structure

Each session involved a tutorial followed by the main task. The tutorial involved simplified versions of the drones with only three features, but otherwise followed the same pattern of a training block before a testing block. Next, they moved onto the main part of the task where they learned the drone classes during training through feedback, then a testing block where they received no feedback, but instead reported their confidence in their selection. At the end of the task they were told both their accuracy and their calibration. If their score was above a threshold of 70, they received a bonus. At the end of each session, subjects reported the strategy they used, and their decision process (Fig. 1C).

Strategy assessment

At the end of the session, subjects reported their strategy through selecting one of four options, "I paid attention to a SUBSET OF COMPONENTS and ignored the others," "I paid attention to ALL COMPONENTS," "I DID NOT HAVE a strategy," or "I WROTE DOWN the components and classifications." They also reported their decision process, with the options: "I CAREFULLY EVALUATED the options," "I trusted MY GUT," or "I chose RANDOMLY."

Trait questionnaires

In psychiatric studies of metacognition, researchers increasingly take a transdiagnostic approach and profile subjects according to clinical traits (Rouault et al., 2018; Seow & Gillan, 2020). To investigate whether the tool can provide insight into psychological and psychiatric traits, we had subjects complete several questionnaires at the beginning of their first session.

Patient Health Questionnaire (PHQ-9)

The PHQ-9 is designed for assessing signs of depression (Kroenke et al., 2001). It is widely used both in clinical and research settings.

Broader Autism Phenotype Questionnaire (BAPQ)

The BAPQ is designed to assess phenotypic traits in those related to those with ASD. There are subscales for behavioral rigidity, pragmatic language deficits and aloofness. It has been commonly deployed in studies to assess for ASD traits (Sucksmith et al., 2011).

The World Health Organization Adult ADHD self-report scale (ASRS)

The ASRS is a screener for attention hyperactivity disorder. The questionnaire has two subscales, one for inattention and the other for hyperactivity/impulsivity. It is a common assessment tool for screening of and research on ADHD (Taylor et al., 2011).

Big Five Inventory- Short Version (BFI-10)

The BFI-10 is a highly abbreviated assessment of the five personality traits: extraversion, agreeableness, conscientiousness, neuroticism, openness (Rammstedt & John, 2007).

Online subject recruitment

We recruited 100 subjects through the online platform Prolific. To be included, subjects were required to be over 18 years of age, located in the US, speak English as a first language and told Prolific they graduated from college. They participated in three sessions. During the first session, they answered several clinical trait questionnaires before participating in the task. They were then invited back to play the task again after one week (n=95), and again two weeks subsequent (n=80, Fig. 1D). Only those who completed all three sessions were included in the behavioral analyses. Of those who completed all sessions, the mean age was 39.54 years, median age was 37 years, 32/80 reported female sex at birth. Furthermore, two of the subjects who reported male at birth did not identify as male. One identified as female and the other as a trans female. During the first sessions subjects were paid $4 for participation and a $2 performance bonus. During the second and third sessions, subjects were paid $3 for participation and a $2 performance bonus.

Questionnaire subscale factor analysis

To identify underlying traits driving responses across questionnaires, we performed factor analysis on the subscale scores. The number of factors was identified through examination of the scree plot and cumulative variance, such that 90% of the variance was captured. When interpreting factor loadings, we set a threshold of 0.6 for significance.

Task behavior analysis

All analyses of task play behavior was done on the block where subjects reported their judgment confidence. The metrics included: overall drone classification accuracy; overall reward; Brier score; as well as the Brier score decomposed into the resolution, calibration and outcome variance (Fleming & Lau, 2014). For each of these metrics, we also measured the intraclass correlation coefficient (ICC) using a random mixed effects model averaged across sessions (McGraw & Wong, 1996).

Results

Internally consistent trait-questionnaire responses

Before using the questionnaires to interpret subject traits, we established response reliability by assessing internal consistency. We found expected correlations between responses to similar questionnaire subscales (Fig. 2A). For example, there was a high positive correlation between the ADHD measures, as well as between the BFI-10 Neuroticism metric and all psychiatric traits. Similarly, there was high internal positive correlation between BFI-10 subscales for Openness, Extraversion, Conscientiousness and Agreeableness, while negative correlation between those measures and traits for autism or depression. This indicates that subjects answered the questions with high internal consistency and supports their use in assessing subject traits.

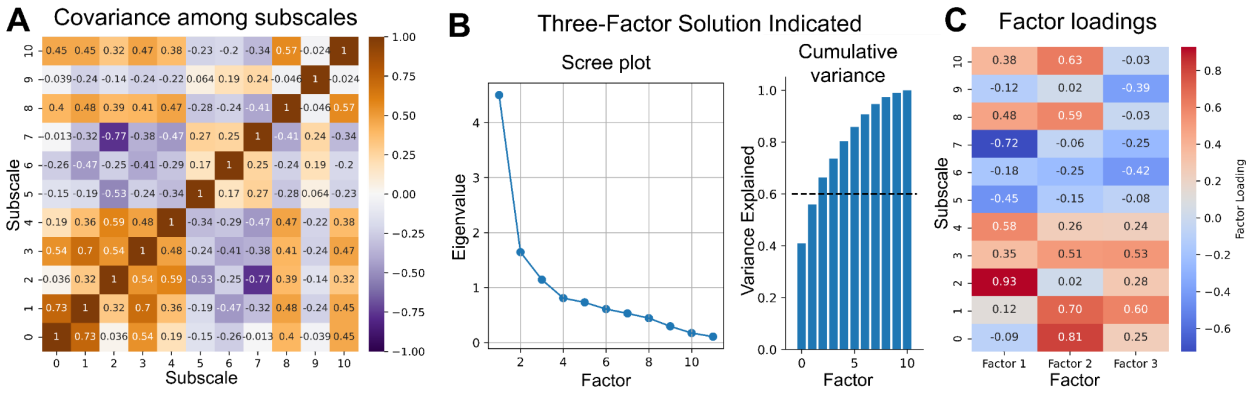

Figure 2. Internal consistency of questionnaire responses and factors identified for social or attention traits. 0 = ASRS: Hyperactivity/Impulsivity, 1 = ASRS: Inattention, 2 = BAPQ: Aloof, 3 = BAPQ: Pragmatic Language, 4 = BAPQ: Rigidity, 5 = BFI-10 Agreeableness, 6 = BFI-10 Conscientiousness, 7 = BFI-10: Extraversion, 8 = BFI-10: Neuroticism, 9 = BFI-10: Openness, 10 = PHQ9. (**A**) High positive correlation among subscales for psychiatric traits, and negative correlation between psychiatric trait subscales and personality inventory subscales. (**B**) Screen plot (left) and cumulative variance plot(right) suggest the solution of three factors. (**C**) Factor 1 captures social traits and Factor 2 captures attention-related traits. BAPQ Aloofness subscale loads positively and the BFI-10 Extraversion subscale negatively on Factor 1, while there is little loading on the ADHD subscales. Factor 2, conversely loads heavily positive for both ADHD subscales and minimally on the social subscales.

Factor analysis identifies a social factor and an attentional factor

Confident in the internal-consistency of responses, we performed factor analysis of the subscale scores. There were three factors with an eigenvalue greater than 1, and three factors were required to capture over 60% of the variance (Fig. 2B). We therefore used three dimensions for our factor analysis. Factor 1 had loadings of the greatest magnitude for BAPQ Aloofness (0.93) and BFI-10 Extroversion (-0.72). Factor 2 had loadings of the greatest magnitude for ASRS Hyperactivity/Impulsivity (0.81), ASRS Inattention

(0.7), and the PHQ-9 (0.63). Of note, the significant loadings in Factor 1 were small in Factor 2, and vice versa. Factor three had a significant loading for ASRS Inattention (0.6). These results show that factor analysis of the questionnaire responses provides lower dimensions for characterizing trait-behavior relationships.

Subjects perform worse as a function of stimulus feature distance from prototypes

We examined whether the feature distance of stimulus examples from their prototype had a significant effect on performance. This is important for establishing that prompt difficulty can be controlled by manipulating the number of features that are shared between an example and a prototype.

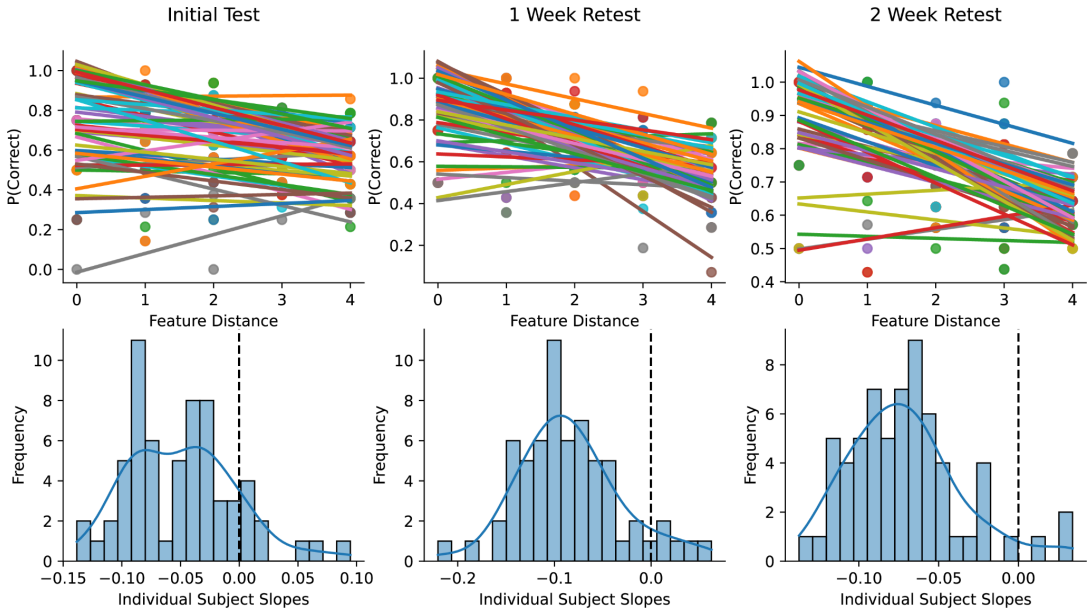

Figure 3. Significant relationship between performance and feature distance. Subjects shown performed above chance across all sessions. Top row: the proportion correct as a function of the example's feature distance for each subject. Bottom row: A histogram of slopes fit to individual subject performance as a function of distance. A slope less than zero indicates that the larger the feature distance, the worse the performance.

To investigate this question, we selected subjects whose performance across sessions was consistently above chance (performance >0.5, N=64/80). For each session, we averaged the performance across all stimuli of the same distance, with the prototypes being distance zero (individual slopes in Fig 3., top row and histogram of slopes in bottom row). We then fit mixed effects models to the proportion correct as a function of distance for each session, with individual subjects as a random intercept. In all sessions, we found a significant effect of distance on performance ($P<0.05$). This was replicated when we included all subjects in the analysis by removing the constraint of performing above chance ($P<0.05$). Thus, we found that feature distance had a significant effect on difficulty.

High correlation between subject accuracy and calibration across all sessions

One well-established issue with Brier scores and the associated calibration measure is the confounding effect of accuracy (Fleming & Lau, 2014; Guggenmos, 2021). Even if a subject's underlying calibration

remains consistent, their calibration metric will show improvement as their accuracy improves. We therefore examined the correlation between subject accuracy and calibration in each session. We found significant correlation between accuracy and calibration during all sessions (Fig. 4, Pearson's correlation $P<0.05$). This suggests that the measure of calibration may be confounded by the variability in accuracy across subjects. The confound could be controlled for by including a well-designed adaptive difficulty mechanism (Rahnev & Fleming, 2019).

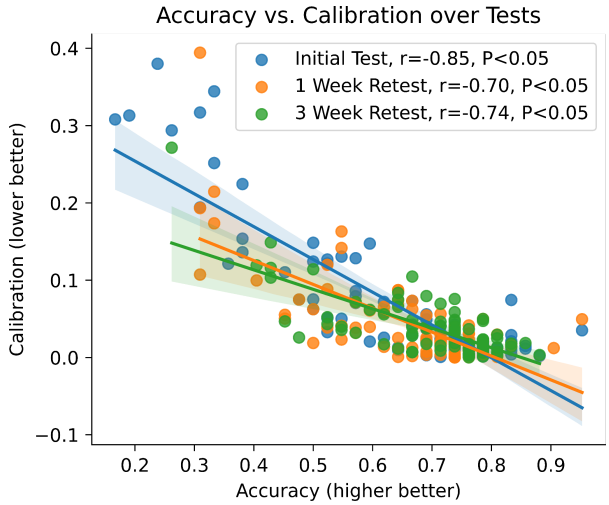

Figure 4. Accuracy and calibration are confounded. Correlation of accuracy and calibration for each session revealed there was a significant relationship.

Test-retest reliability fair for accuracy, Brier score, outcome variance and reward, but poor for calibration and resolution

For a test to be reliable, it must either be consistent across sessions or must change in interpretable ways. We therefore calculated the ICC for each behavioral measure. The ICC produces a number between 0 and 1 where $ICC<0.04$ is poor, $0.40<ICC<0.59$ is fair, $0.60<ICC<0.74$ is good and $ICC>0.75$ is excellent (Cicchetti, 1994). We found that the ICC for accuracy ($ICC=0.56$), the Brier score ($ICC=0.53$), outcome variance ($ICC=0.56$) and rewarded score ($ICC=0.54$) were all in the fair range. The ICC for calibration ($ICC=0.32$) and resolution ($ICC=0.22$) were poor (Fig. 5).

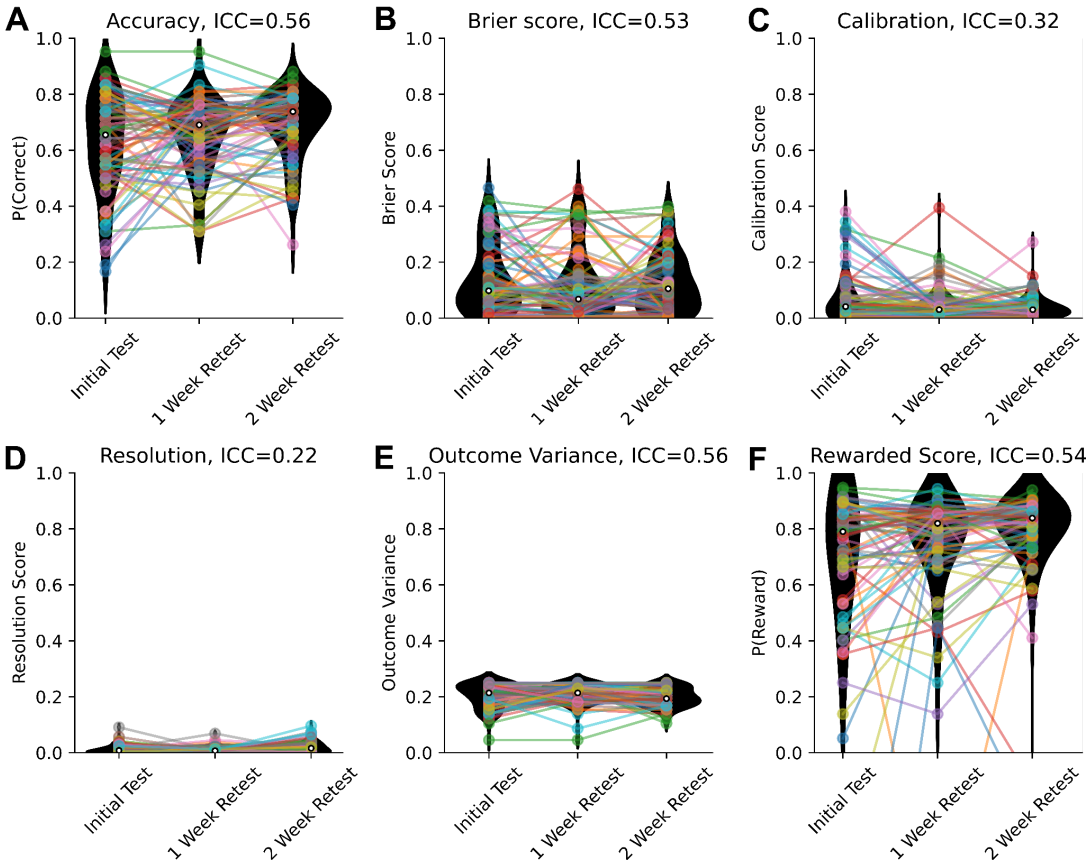

Figure 5. Mixed test-retest reliability when including all subjects. The task metrics across sessions are plotted with the distribution of subjects in black, and the values at each timepoint for individual subjects colored and joined by lines. (**A**) The accuracy across sessions showed significant stability. (**B**) The Brier score was also significantly stable. (**C**) The calibration had a weak correlation across sessions. (**D**) The resolution also had a weak correlation. (**E**) The outcome variance was stable. (**F**) The reward subjects received was also largely stable.

Well-calibrated subjects attend to a subset of features.

We then examined the characteristics of subjects who were reasonably well-calibrated. To do that, we identified subjects whose calibration scores consistently fell below 0.1 across all sessions (N=50/80). We then split the subjects in half based on their calibration score (Fig. 6A). Of the subjects who passed the threshold, we examined their reported strategies, and compared answers between the top half and the bottom half of the calibration rankings. We found that well-calibrated subjects were much more likely to report that they paid attention to a subset of components and ignored the others. They were also much less likely to report having chosen randomly (Fig. 6B). This indicates that attentional deployment is associated with metacognitive performance.

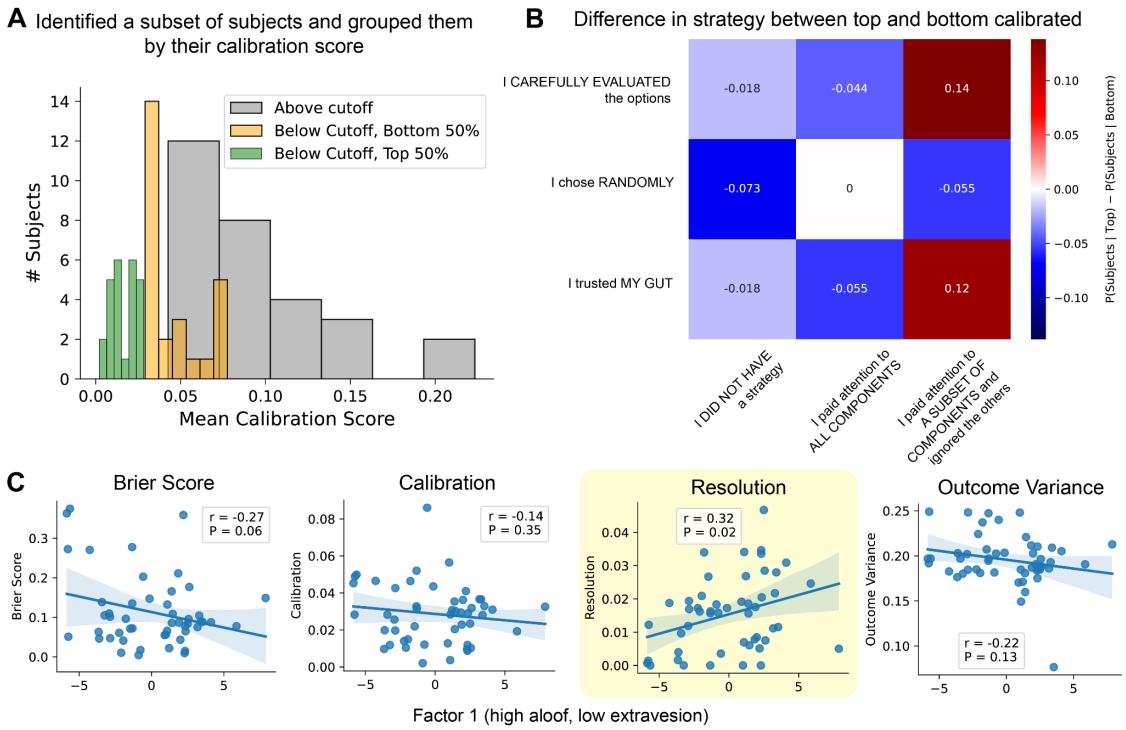

Figure 6. Subjects consistent in calibration scores across sessions. (**A**) We identified a subset of subjects with calibration scores consistently below 0.1 across all sessions, and divided those in half based on their calibration scores. (**B**) Best calibrated subjects attended to a subset of features. We split the top half of calibrated subjects from the bottom and compared their reported strategies. The top half of subjects were more likely to report attending to a subset of drone components and ignoring all others. (**C**) In a regression of the task scores against the subscale factors, we found Factor 1 (the social factor) was significantly correlated with improved resolution.

Aloofness and introversion are associated with improved metacognitive resolution

We then analyzed the correlation between the Factors and the task metrics. We found that Factor 1 (high aloof and low extraversion) was significant for the resolution score (Pearson's correlation, $P<0.05$; Fig. 6C). This indicates that performance on a metacognitive task can provide insight into clinical and personality traits. (Extensive additional trait-behavior analyses are featured in the available code repository.)

Discussion

We developed a tool for longitudinally assessment of individual judgment calibration, then paired it with a strategy report and trait questionnaires. We found poor to fair retest stability in the judgment calibration metrics, and identified several ways in which the retest validity of the tool could be improved. When narrowing to the most well-calibrated subjects, we found a significant correlation between aloofness/introversion and metacognitive resolution, indicating the tool will be useful in studies of cognitive traits. We also found the well-calibrated subjects narrowed their attention to a subset of features, indicating a relationship between attention and metacognition. The findings also suggest the tool could serve as a benchmark for interventions that improve these abilities (Tseng et al., 2020).

Controlling for performance

Empirical studies have shown that even calibration metrics designed to be robust against performance confounds are, in practice, still affected by it (Guggenmos, 2021; Rahnev, 2023). Thus, the most effective way to control for the confound is by ensuring all subjects produce the same level of performance. Indeed, this is one of the reasons we adopted a prompt modality where one can manipulate difficulty through specifying the feature distance. One of the most common ways to ensure a stable level of performance is through staircasing the difficulty of prompts. If subjects are performing better than expected, they then get a series of more difficult questions. If they perform worse, the questions get easier. The level of difficulty thus forms a "staircase" over time as the prompts gradually get more difficult, or easier. However, in studies of metacognition staircasing can introduce an additional form of confounding variance in the form of variable difficulty across stimuli. For this reason, it recommended to use a single difficulty level for computation of metacognition (Rahnev & Fleming, 2019).

Given this is a significant deviation from the category learning paradigm, we did not implement it in the initial assessment of the tool. However, in future versions of the task, the judgment calibration block could be divided into a staircasing sub-block and an assessment sub-block. During the staircasing sub-block, the feature distance would be manipulated until a stable level of performance is achieved. During the assessment sub-block, a single feature distance would be used for all trials.

Trial count

The total number of trials has a significant effect on longitudinal stability of metacognition measurements (Rahnev, 2023). Intuitively, this can be understood that additional trials reduce the effect of extraneous variability. In the current study, we used the structure and trial count from the original category learning task (Gorlick & Maddox, 2013). However, given the combinatorial possibilities offered by the stimulus set, future versions could easily increase the number of trials while still avoiding stimulus repetition.

Proper scoring rule

We used a scoring rule (averaging the performance and normalized calibration score) for ease of explainability rather than robustness. However, it is possible that over the course of sessions subjects focused on task performance at the expense of calibration, or vice versa. If that is the case, an optimized scoring rule could incentivize subjects to actively maximize both performance and judgment calibration in each session (Gneiting & Raftery, 2007; Johnstone et al., 2011; Merkle & Steyvers, 2013).

Trait assessment

We found significant indications that task behavior correlated with psychological traits. However, given the sample size, retest stability issues and lack of a replication sample, we are hesitant to draw strong conclusions from these specific findings. Rather, our results strongly support pairing the tool with extensive trait profiles, such as an extended test of the Big Five, tests for psychotic traits, as well as OCD and rigidity (Barrett et al., 2015; Caprara et al., 1993; Foa et al., 2002; Mossaheb et al., 2012).

Attention and metacognition

Interestingly, we found that the most well-calibrated subjects were more likely to report narrowing their attention to a subset of features. This indicates that subjects may have deployed qualitatively different category learning strategies (Pettine et al., 2023), and that use of these strategies is connected to metacognitive ability. It also supports the use of the hub-and-spoke design so that future studies can use

eye tracking to objectively quantify the integration of feature information (Bahmani et al., 2018; Pettine et al., 2019).

Interventions

With appropriate modifications to increase retest stability, the tool could serve as a benchmark for interventions that improve metacognition in situations involving goal-directed attention, categorization and learning. Possible interventions include explicit instruction on probability estimates (Mellers, Stone, Atanasov, et al., 2015), frequent feedback on within-session judgment calibration (J. Carpenter et al., 2019), or instruction on effective deployment of goal-directed attention (Peng & Miller, 2016) or adaptive programs targeting working memory and information processing speed and accuracy (Tseng et al., 2023). The availability of over 10,000 stimulus configurations means that subjects can engage a near infinite number of sessions. Moreover, the flexible nature of the tool makes it readily adaptable.

Conclusions

We created a tool for assessing attention and metacognition during category learning, and identified specific areas for future development. The tool will be highly useful for investigating these capacities in professions such as national intelligence, studies in psychology and psychiatry, as well as benchmarking interventions (Redish et al., 2021).

Acknowledgements

This work was supported by a SFARI Human Cognitive and Behavioural Science Explorer Award 988485 (J.D.M., S.J., W.W.P.), UMN NeuroPRSMH Seed Award (S.J., W.W.P.) as well as by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at Yale University, administered by Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy and the Office of the Director of National Intelligence (ODNI) (W.W.P.).

Data Availability

Behavioral data are available at https://zenodo.org/records/10155670.

Code Availability

Task code is available at the repository, https://github.com/pettinelab/public_category_learning_metacog_django_webapp. Study management and behavioral analysis code is available at the repository, https://github.com/pettinelab/public_category_learning_metacog_management_analysis. The behavioral analysis code features extensive additional analyses of trait-behavior relationships.

Citations

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*(3), 486–498. https://doi.org/10.1037/0033-2909.110.3.486

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Management Science*, *63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Bahmani, Z., Clark, K., Merrikhi, Y., Mueller, A., Pettine, W., Vanegas, M. I., Moore, T., & Noudoost, B. (2018). *Prefrontal Contributions to Attention and Working Memory*. 1–25. https://doi.org/10.1007/7854_2018_74

Barrett, S. L., Uljarević, M., Baker, E. K., Richdale, A. L., Jones, C. R. G., & Leekam, S. R. (2015). The Adult Repetitive Behaviours Questionnaire-2 (RBQ-2A): A Self-Report Measure of Restricted and Repetitive Behaviours. *Journal of Autism and Developmental Disorders*, *45*(11), 3680–3692. https://doi.org/10.1007/s10803-015-2514-6

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, *9*, e59360. https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract Memory Representations in the Ventromedial Prefrontal Cortex and Hippocampus Support Concept Generalization. *Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, *11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. (1993). The "big five questionnaire": A new questionnaire to assess the five factor model. *Personality and Individual Differences*, *15*(3), 281–288. https://doi.org/10.1016/0191-8869(93)90218-R

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), Article 1. https://doi.org/10/1/Carpenter_MetaTrainingManuscript.pdf

Carpenter, K. L., & Williams, D. M. (2023). A meta-analysis and critical review of metacognitive accuracy in autism. *Autism*, *27*(2), 512–525. https://doi.org/10.1177/13623613221106004

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284.

Coane, J. H., & Umanath, S. (2021). A database of general knowledge question performance in older adults. *Behavior Research Methods*, *53*(1), 415–429. https://doi.org/10.3758/s13428-020-01493-2

Faissner, M., Kriston, L., Moritz, S., & Jelinek, L. (2018). Course and stability of cognitive and metacognitive beliefs in depression. *Depression and Anxiety*, *35*(12), 1239–1246. https://doi.org/10.1002/da.22834

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1). https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Fleming, S. M., & Frith, C. D. (Eds.). (2014). *The cognitive neuroscience of metacognition*. Springer-Verlag Publishing.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496.

Friedman, J. A., & Zeckhauser, R. (2018). Analytic Confidence and Political Decision-Making: Theoretical Principles and Experimental Evidence From National Security Professionals. *Political Psychology*, *39*(5), 1069–1087. https://doi.org/10.1111/pops.12465

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gorlick, M. A., & Maddox, W. T. (2013). Priming for Performance: Valence of Emotional Primes Interact with Dissociable Prototype Learning Systems. *PLOS ONE*, *8*(4), e60748. https://doi.org/10.1371/journal.pone.0060748

Gorlick, M. A., Worthy, D. A., Knopik, V. S., McGeary, J. E., Beevers, C. G., & Maddox, W. T. (2015). DRD4 Long Allele Carriers Show Heightened Attention to High-priority Items Relative to Low-priority Items. *Journal of Cognitive Neuroscience*, *27*(3), 509–521. https://doi.org/10.1162/jocn_a_00724

Grossmann, I., Rotella, A., Hutcherson, C. A., Sharpinskyi, K., Varnum, M. E. W., Achter, S., Dhami, M. K., Guo, X. E., Kara-Yakoubian, M., Mandel, D. R., Raes, L., Tay, L., Vie, A., Wagner, L., Adamkovic, M., Arami, A., Arriaga, P., Bandara, K., Baník, G., … The Forecasting Collaborative. (2023). Insights into the accuracy of social scientists' forecasts of societal change. *Nature Human Behaviour*, *7*(4), Article 4. https://doi.org/10.1038/s41562-022-01517-1

Guggenmos, M. (2021). Measuring metacognitive performance: Type 1 performance dependence and test-retest reliability. *Neuroscience of Consciousness*, *2021*(1), niab040. https://doi.org/10.1093/nc/niab040

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational Psychiatry*, *9*(1), 1–18. https://doi.org/10.1038/s41398-019-0602-7

Johnstone, D. J., Jose, V. R. R., & Winkler, R. L. (2011). Tailored Scoring Rules for Probabilities. *Decision Analysis*, *8*(4), 256–268. https://doi.org/10.1287/deca.1110.0216

Joo, H. R., Liang, H., Chung, J. E., Geaghan-Breiner, C., Fan, J. L., Nachman, B. P., Kepecs, A., & Frank, L. M. (2021). Rats use memory confidence to guide decisions. *Current Biology*. https://doi.org/10.1016/j.cub.2021.08.013

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., & Kepecs, A. (2014). Orbitofrontal Cortex Is Required for Optimal Waiting Based on Decision Confidence. *Neuron*, *84*(1), 190–201. https://doi.org/10.1016/j.neuron.2014.08.039

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*(2), 149–171. https://doi.org/10.1016/0030-5073(80)90052-5

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). *Calibration of probabilities: The state of the art to 1980* (D. Kahneman, P. Slovic, & A. Tversky, Eds.; pp. 306–334). Cambridge University Press. http://www.cambridge.org/emea/

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d′, Response-Specific Meta-d′, and the Unequal Variance SDT Model. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30.

McMahon, C. M., Henderson, H. A., Newell, L., Jaime, M., & Mundy, P. (2016). Metacognitive Awareness of Facial Affect in Higher-Functioning Children and Adolescents with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *46*(3), 882–898. https://doi.org/10.1007/s10803-015-2630-3

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*(1), 1–14. https://doi.org/10.1037/xap0000040

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: A Journal of the Association for Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Merkle, E. C., & Steyvers, M. (2013). Choosing a Strictly Proper Scoring Rule. *Decision Analysis*, *10*(4), 292–304. https://doi.org/10.1287/deca.2013.0280

Moreira, C. M., Rollwage, M., Kaduk, K., Wilke, M., & Kagan, I. (2018). Post-decision wagering after perceptual judgments reveals bi-directional certainty readouts. *Cognition*, *176*, 40–52. https://doi.org/10.1016/j.cognition.2018.02.026

Mossaheb, N., Becker, J., Schaefer, M. R., Klier, C. M., Schloegelhofer, M., Papageorgiou, K., & Amminger, G. P. (2012). The Community Assessment of Psychic Experience (CAPE) questionnaire as a screening-instrument in the detection of individuals at ultra-high risk for psychosis. *Schizophrenia Research*, *141*(2), 210–214. https://doi.org/10.1016/j.schres.2012.08.008

Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in Psychology. *Review of General Psychology*, *23*(4), 403–424. https://doi.org/10.1177/1089268019883821

Peng, P., & Miller, A. C. (2016). Does attention training work? A selective meta-analysis to explore the effects of attention training and moderators. *Learning and Individual Differences*, *45*, 77–87. https://doi.org/10.1016/j.lindif.2015.11.012

Pettine, W. W., Raman, D. V., Redish, A. D., & Murray, J. D. (2023). Human generalization of internal representations through prototype learning with goal-directed attention. *Nature Human Behaviour*, *7*(3), Article 3. https://doi.org/10.1038/s41562-023-01543-7

Pettine, W. W., Steinmetz, N. A., & Moore, T. (2019). Laminar segregation of sensory coding and behavioral readout in macaque V4. *Proceedings of the National Academy of Sciences*, 201819398. https://doi.org/10.1073/pnas.1819398116

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Radulescu, A., Shin, Y. S., & Niv, Y. (2021). Human Representation Learning. *Https://Doi.Org/10.1146/Annurev-Neuro-092920-120559*. https://doi.org/10.1146/annurev-neuro-092920-120559

Rahnev, D. (2023). *Measuring metacognition: A comprehensive assessment of current methods*. PsyArXiv. https://doi.org/10.31234/osf.io/waz9h

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *2019*(1), niz009. https://doi.org/10.1093/nc/niz009

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. https://doi.org/10.1016/j.jrp.2006.02.001

Redish, A. D., Kepecs, A., Anderson, L. M., Calvin, O. L., Grissom, N. M., Haynos, A. F., Heilbronner, S. R., Herman, A. B., Jacob, S., Ma, S., Vilares, I., Vinogradov, S., Walters, C. J., Widge, A. S., Zick, J. L., & Zilverstand, A. (2021). Computational validity: Using computation to translate behaviours across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200525. https://doi.org/10.1098/rstb.2020.0525

Rieber, S. (2004). Intelligence Analysis and Judgmental Calibration. *International Journal of Intelligence and CounterIntelligence*, *17*(1), 97–112. https://doi.org/10.1080/08850600490273431

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, *33*(2), 7–18.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. https://doi.org/10.1016/j.concog.2009.12.013

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, *90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Seow, T. X. F., & Gillan, C. M. (2020). Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. *Scientific Reports*, *10*(1), 2883. https://doi.org/10.1038/s41598-020-59646-4

Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How Local and Global Metacognition Shape Mental Health. *Biological Psychiatry*, *90*(7), 436–446. https://doi.org/10.1016/j.biopsych.2021.05.013

Sun, X., Zhu, C., & So, S. H. W. (2017). Dysfunctional metacognition across psychopathologies: A meta-analytic review. *European Psychiatry*, *45*, 139–153. https://doi.org/10.1016/j.eurpsy.2017.05.029

Tseng, A., Biagianti, B., Francis, S. M., Conelea, C. A., & Jacob, S. (2020). Social Cognitive Interventions for Adolescents with Autism Spectrum Disorders: A Systematic Review. *Journal of Affective Disorders*, *274*, 199–204. https://doi.org/10.1016/j.jad.2020.05.134

Tseng, A., DuBois, M., Biagianti, B., Brumley, C., & Jacob, S. (2023). Auditory Domain Sensitivity and Neuroplasticity-Based Targeted Cognitive Training in Autism Spectrum Disorder. *Journal of Clinical Medicine*, *12*(4), 1635.

Wu, J., & Fu, Q. (2021). The role of working memory and visual processing in prototype category learning. *Consciousness and Cognition*, *94*, 103176. https://doi.org/10.1016/j.concog.2021.103176

Wu, J., Fu, Q., & Rose, M. (2020). Stimulus modality influences the acquisition and use of the rule-based strategy and the similarity-based strategy in category learning. *Neurobiology of Learning and Memory*, *168*, 107152. https://doi.org/10.1016/j.nlm.2019.107152

Zhou, X., Fu, Q., & Rose, M. (2020). The Role of Edge-Based and Surface-Based Information in Incidental Category Learning: Evidence From Behavior and Event-Related Potentials. *Frontiers in Integrative Neuroscience*, *14*. https://www.frontiersin.org/articles/10.3389/fnint.2020.00036

Zhou, X., Fu, Q., Rose, M., & Sun, Y. (2019). Which Matters More in Incidental Category Learning: Edge-Based Versus Surface-Based Features. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00183