How large is "large enough"? Large-scale experimental investigation of the reliability of confidence measures

Clémentine Bouleau, Nicolas Jacquemet, Maël Lebreton

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

WORKING PAPER N° 2025-06

How large is "large enough"? Large-scale experimental investigation of the reliability of confidence measures

Clémentine Bouleau
Nicolas Jacquemet
Maël Lebreton

JEL Codes:
Keywords: Confidence; Accuracy; Reliability; Design of experiments; Multiple trials

anr agence nationale de la recherche
AU SERVICE DE LA SCIENCE

How large is "large enough"? Large-scale experimental investigation of the reliability of confidence measures*

Clémentine Bouleau† Nicolas Jacquemet† Maël Lebreton‡

January 2025

Abstract

Whether individuals feel confident about their own actions, choices, or statements being correct, and how these confidence levels differ between individuals are two key primitives for countless behavioral theories and phenomena. In cognitive tasks, individual confidence is typically measured as the average of reports about choice accuracy, but how reliable is the resulting characterization of within- and between-individual confidence remains surprisingly undocumented. Here, we perform a large-scale resampling exercise in the Confidence Database to investigate the reliability of individual confidence estimates, and of comparisons across individuals' confidence levels. Our results show that confidence estimates are more stable than their choice-accuracy counterpart, reaching a reliability plateau after roughly 50 trials, regardless of a number of task design characteristics. While constituting a reliability upper-bound for task-based confidence measures, and thereby leaving open the question of the reliability of the construct itself, these results characterize the robustness of past and future task designs.

Keywords: Confidence; Accuracy; Reliability; Design of experiments; Multiple trials.

*Authors are listed in alphabetical order. We thank participants to the "*Psychological belief formation*" workshop (2024, Paris), the "*Models of learning and decision-making: an interdisciplinary approach*" workshop (2024, Les Treilles), the "*Market, Cooperation and Voting*" workshop (2024, Madrid), the 2024 *ASFEE* annual conference (Grenoble) and the workshop "*Behavioural Science Meets Metascience*" (2024, Oxford) for insightful feedback. Financial support from the European Research Council (Starting Grant 948671) and from the French National Research Agency (program *Investissements d'Avenir*, ANR-10–LABX-93-0 and ANR-17–EURE-0001) are gratefully acknowledged.

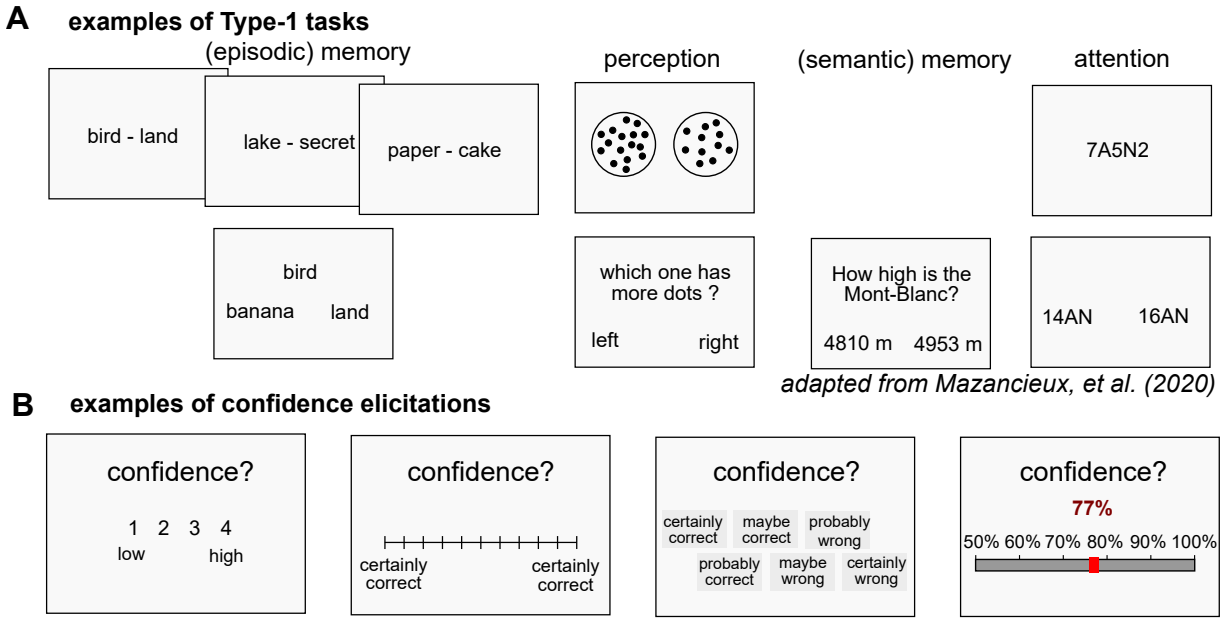
†Paris School of Economics and U. Paris 1 Panthéon-Sorbonne. Centre d'Economie de la Sorbonne (CES), Maison des Sciences Economiques, 106-112 boulevard de l'Hôpital 75013 Paris. clementine.bouleau@psemail.eu; nicolas.jacquemet@univ-paris1.fr.

‡Paris School of Economics, 48 Bd Jourdan, 75014 Paris; and Swiss Center for Affective Sciences, Université de Genève, 24 rue du Général-Dufour, 1211 Genève 4. mael.lebreton@psemail.eu.

The concept of individual self-confidence is key to theoretical and empirical studies in behavioral sciences. Confidence is an important determinant of individual beliefs (Bénabou and Tirole, 2002; Möbius et al., 2022; Zimmermann, 2020), corporate investment (Malmendier and Tate, 2005), financial decision-making (Scheinkman and Xiong, 2003), self-employment (Koellinger et al., 2007), voting and political behavior (Ortoleva and Snowberg, 2015; Rollwage et al., 2018), management strategy (Russo et al., 1992) and medical errors (Berner and Graber, 2008). Differences in confidence across individuals of specific socio-demographic groups (e.g., gender, or cultures) might moreover contribute to biases that are both socially undesirable and market-inefficient (Bhandari and Deaves, 2006; Niederle and Vesterlund, 2007; Lundeberg et al., 1994). On the clinical side, anomalies in individual levels of confidence have been shown to underpin neuro-psychiatric conditions such as anxiety, depression or compulsivity (Hoven et al., 2019, 2023; Rouault et al., 2018b). Empirically testing those theories and developing potential applications requires individual confidence measures that reliably capture individual confidence level and between individual differences in confidence.

In recent years, with the rise of research sub-disciplines like computational political psychology (Rollwage et al., 2019) and computational psychiatry (Huys et al., 2016), empirical studies have increasingly leveraged (meta-)cognitive tasks to measure individual confidence and used this behavioral measure as an explanatory factor for important socio-economic or clinical outcomes of interests (Hoven et al., 2023). Individual confidence measures take the form of an averaging of beliefs — typically reported on a rating scale — about the accuracy of binary choices, over multiple trials of, e.g., a perceptual, memory or reasoning decision task (see, e.g., Mazancieux et al., 2020; Rouault et al., 2023; Lehmann et al., 2022; Rouault et al., 2018a; Mazancieux et al., 2023; West et al., 2023, see Figure 1). This approach is highly convenient and increasingly popular, because it generally allows for a tight control of both decision difficulty and choice accuracy, and lends itself to sophisticated computational modeling able to extract latent components of decision or meta-cognitive processes (Fleming, 2024; Guggenmos, 2022; Salem-Garcia et al., 2023; Boundy-Singer et al., 2023; Navajas et al., 2017). An important implicit assumption of this approach is that the reliability and precision of confidence and metacognitive measures is achieved thanks to the implementation of a large number of trials, that can deliver high within-individual statistical power. However, it remains unclear whether this class of procedures produces measures of individual confidence that can reliably capture both individual levels of confidence (e.g., if measured twice, or across two conditions) and differences in confidence levels between individuals (Rouault et al., 2018a; Guggenmos, 2021). Importantly, while shorter cognitive tasks might be desirable in the context of longitudinal studies or when testing cognitively fragile populations (e.g., patients, elderly people, see Hauser et al., 2022), little is known about how many trials of a cognitive task are sufficient to produce a reliable confidence average — but see Fox et al. (2024).

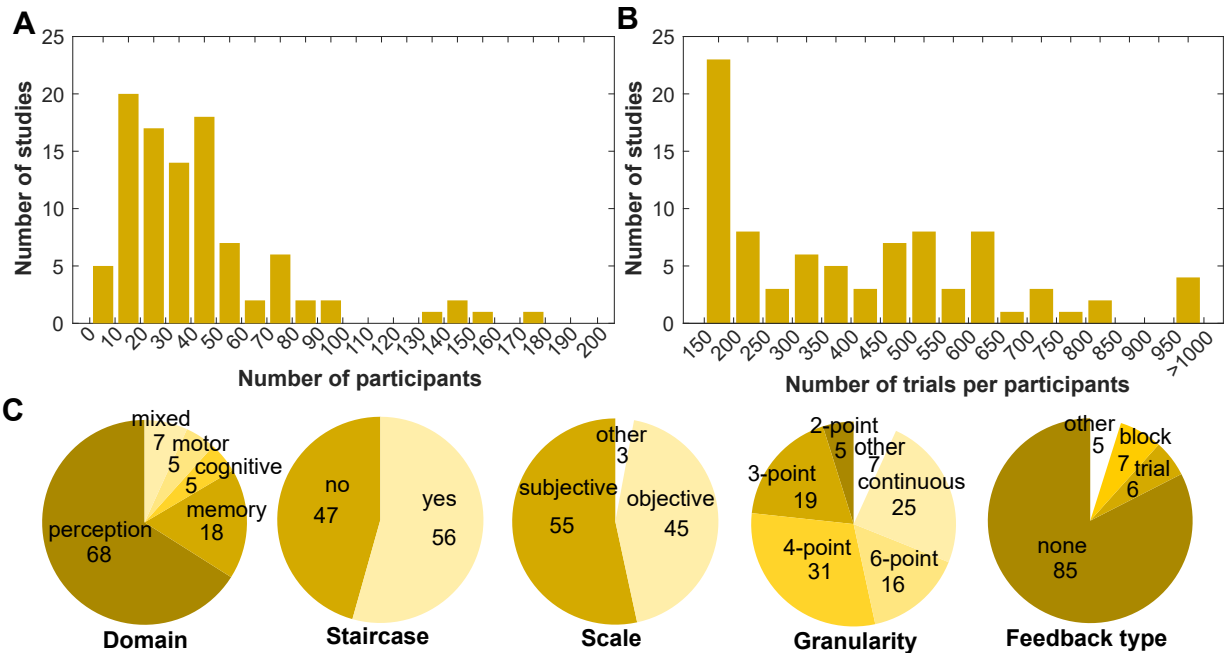
Figure 1: Confidence elicitation in the Confidence Database



Note. Examples of experimental tasks used to elicit confidence. **Panel A**: Examples of type-1 tasks related to different domains. **Panel B**: Examples of the variety of confidence scales used to elicit confidence regarding performance in task-1 elicitation, featuring different levels of granularity and referring either to objective or subjective measures.

Herein we address these questions by performing a large-scale, comprehensive exploration of the reliability of within- and between-individual confidence measures elicited in meta-cognitive tasks, as a function of the number of trials and of the characteristics of the task. To that end, we take advantage of the Confidence Database (CD, Rahnev et al., 2020), a large open source dataset of confidence studies spanning a broad range of paradigms, participants and populations. We selected a subset of 103 studies (over 6,000 participants and 2,000,000 trials, see Figure 2, Panels A and B), which satisfied a list of key minimal constraints (see Methods, Section 3), and spanned a variety of domains (visual, memory, cognitive) and confidence elicitation methods (various scales or binary choices), various choice difficulty level distributions, and different feedback availability rules (see Figure 2.C). The CD tasks generally feature manipulations of one or several experimental factors and can be subject to uncontrolled variations of practice effects, mood or attention, creating large trial-specific heterogeneity (Desender et al., 2022; Weilnhammer et al., 2023; Mei et al., 2023). The common practice is to smooth-out this heterogeneity by averaging confidence reports across multiple trials, to obtain a measure of individual-specific confidence. To assess the sensitivity of this procedure to the number of trials, we thus focus on the reliability of the elicited measure, i.e., the degree to which it yields similar results when repeated under equivalent conditions (Cook and Beckman, 2006; Matheson, 2019; Karvelis et al., 2023).

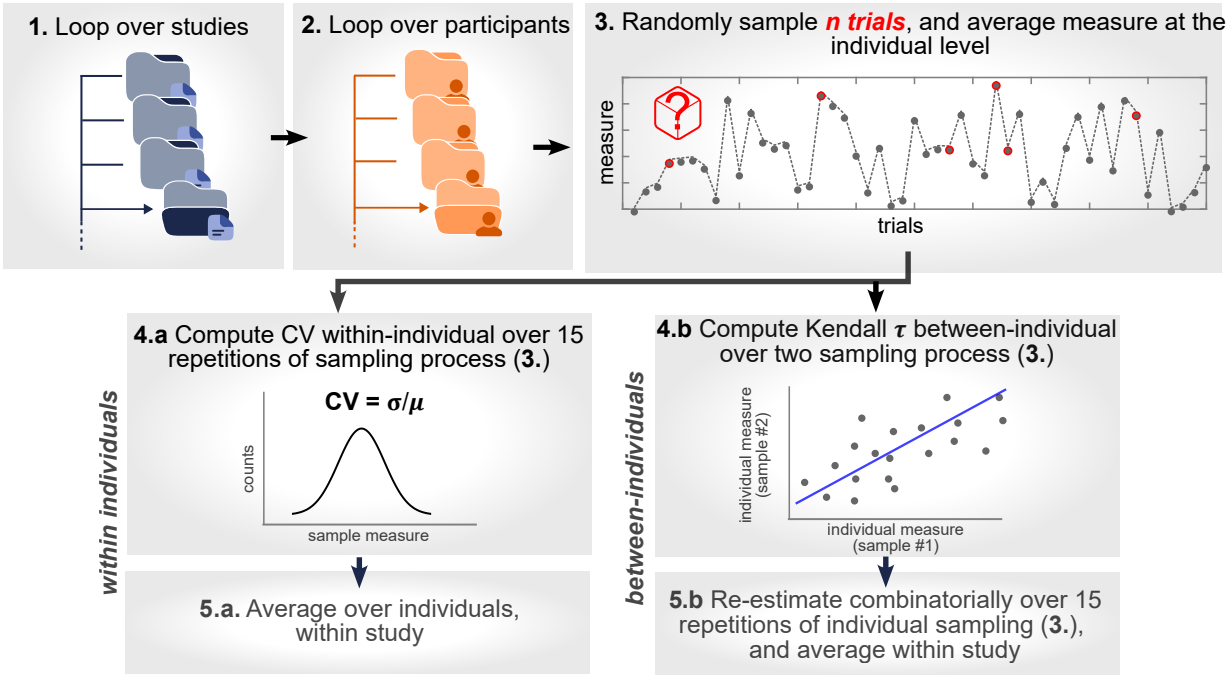
Figure 2: Design heterogeneity in the final sample of studies



Note. Panel A: Distribution of the number of participants per study. **Panel B**: Distribution of the total number of trials per participant in the working sample of $J = 103$ studies (see Methods, Section 3, for more details). **Panel C**: Distribution of studies included in the working sample over the domain of the type-1 task, whether a staircase procedure is applied to performance in this task, whether the confidence scale refers to objective or subjective measures and its level of granularity (see Figure 1.A for examples), and whether some feedback about performance at the task is provided to participants. For details about studies classified as "other", refer to Methods, Section 2.

To measure reliability at both the within- and the between-individual levels, we designed two re-sampling exercises performed on the pooled individual data from all included studies (summarized in Figure 3, see Methods, Section 4, for more details on the measures). First, we repeatedly sampled a fixed number of random trials (n) for each individual in each study, and measured within-individual confidence reliability thanks to the Coefficient of Variation ($\text{CV}_n^{\text{Conf}}$) of the average confidence measures in these samples: this measure lies between 0 and 1, and is smaller when reliability is higher. Second, we sampled a fixed number of random trials for each individual of a same study to compute individual confidence measures, and then estimated the Kendall coefficient of correlations (τ) across individuals, over couples of sampling instances, to measure between-individuals confidence reliability. We then convert this correlation into a measure of Ranking Stability ($\text{RS}_n^{\text{Conf}}$), i.e., the probability that two individuals are ranked similarly by their individual confidence estimate, when it is estimated from a random sample of trials. The higher this probability, the more reliable is the between-individual ranking provided by the confidence measure.

Figure 3: Mains steps of the resampling exercise used to measure reliability



Note. Main steps of the resampling exercise for a given choice of the number of trials, n. For each study (**step 1**) and each participant (**step 2**), we randomly draw (with replacement) n trials and compute the average confidence (/accuracy) for that individual (**step 3**). Within-individual reliability is based on the coefficient variation of the confidence (/accuracy) measure over $R = 15$ replications of this sampling exercise (**step 4a**), averaged at the study level (**step 5a**). Between-individuals reliability is based on the average value of the Kendal's τ of the confidence (/accuracy) ranking between all pairs of individuals within a study (**step 4b**), averaged at the study level (**step 5b**).

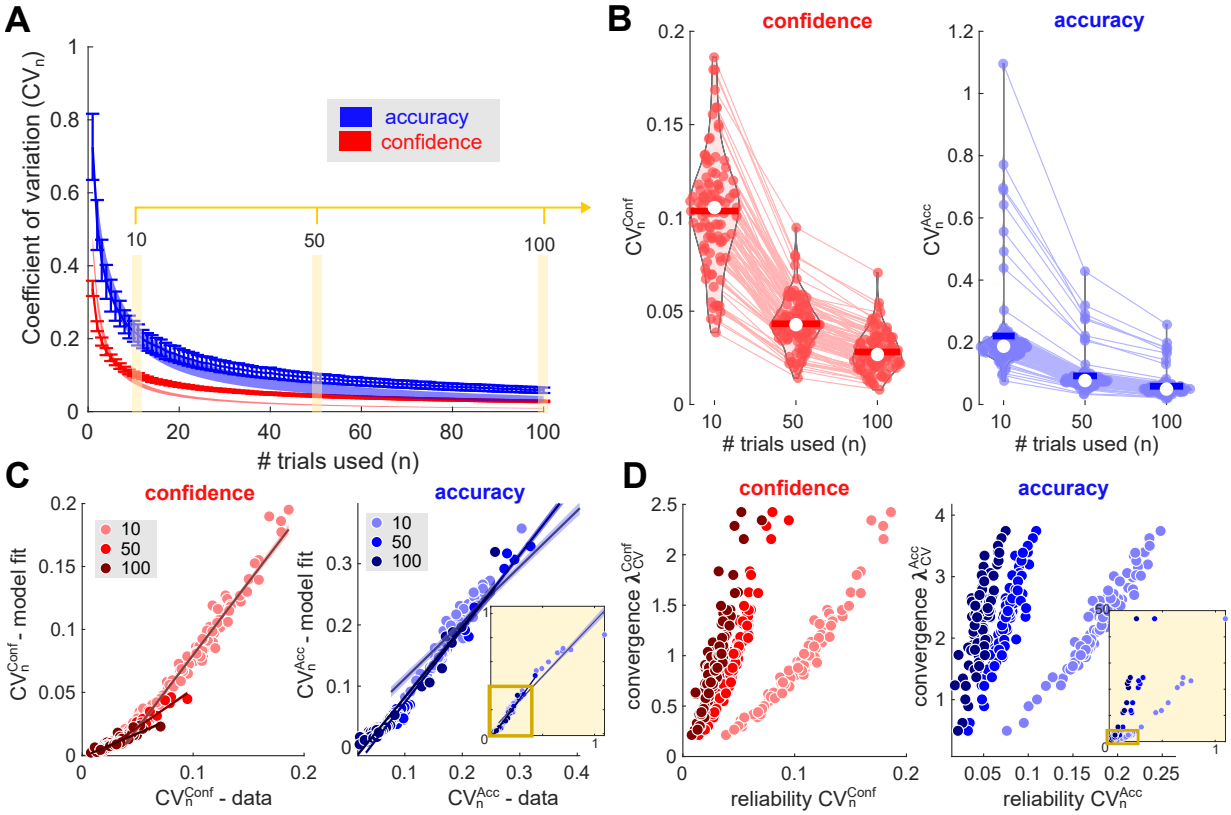
As a reference, we replicated the same analyses on choice accuracy (CV_n^{Acc} and RS_n^{Acc}), which allowed us to assess how individual confidence is stable per se, and not just a consequence of a stable individual choice accuracy. Overall, our results show that individual confidence estimates are more stable than their choice-accuracy counterpart, reaching a reliability plateau after 50 trials, regardless of a number of task design characteristics. To go beyond this aggregate evidence and better document both the dynamic of reliability and its heterogeneity across studies, we structurally estimate study-specific convergence parameters that fit well both the observed dynamic of (within- and between-subjects) reliability and its heterogeneity across studies.

Results

Most of the feasible within-individual reliability is achieved in less than 50 trials

We first focus on the within-individual reliability of confidence measures, i.e., how stable are individual averages of confidence reports over random sampling of trials. Despite all experimental

Figure 4: Within-individual reliability of confidence and accuracy measures



Note. Summary statistics on the within-individual reliability of confidence and accuracy measures (measured by the Coefficient of Variation over replication samples, see step 4a in Figure 3) as a function of the number of trials, n. **Panel A**: Reliability as a function of the number of trials. Lines and error bars indicate the empirical mean +/- 95% CI of $\text{CV}_n^{\text{Conf}}$ (red) and CV_n^{Acc} (blue). Shaded areas indicate the mean +/- 95% CI of fitted CV, obtained from the convergence model in (1), for confidence (red) and accuracy (blue). **Panel B**: Empirical distribution of reliability measures at $n = 10, 50$, and 100, for accuracy (left, blue) and confidence (right, red). Violin plots represent the sample distribution of CV. Connected, colored dots represent the estimates from each individual study. Horizontal bars indicate the sample means and white dots the sample medians. **Panel C**: Fitted reliability (λ_{CV}) of accuracy (left, blue) and confidence (right, red) by the convergence model in (1), as a function of empirically measured CV, at $n = 10, 50$, and 100 (resp. light, medium and dark coloured). Each dot represents an individual study, and the coloured lines correspond to the best fit of linear regressions, at $n = 10, 50$, and 100. **Panel D**: Correlation at the study level between the estimated convergence parameter from this same model and the empirical measures of reliability at $n = 10, 50$, and 100.

and non-experimental factors (choice difficulty, attention, mood, etc.) which generally induce variations in confidence reports, the coefficients of variation of confidence measures quickly and smoothly drop (indicating an increase in reliability) as the number of trials increases (Figure 4.A; from $\text{CV}_5^{\text{Conf}} = .15 \pm .01$ (95% CI) to $\text{CV}_{25}^{\text{Conf}} = .06 \pm .00$ and $\text{CV}_{50}^{\text{Conf}} = .04 \pm .00$). Remarkably, the same analysis performed on choice accuracy reveals a similar profile, although the reliability of accuracy is lower than confidence reliability ($\text{CV}_5^{\text{Acc}} = .32 \pm .04$ to $\text{CV}_{25}^{\text{Acc}} = .13 \pm .02$ and $\text{CV}_{50}^{\text{Acc}} = .09 \pm .01$). To better characterize this dynamic, we extracted the study level summary statistics for three levels of the number of trials ($n = 10, 50$ and 100; Figure 4.B). This analysis confirmed

that the increase in reliability is significant in magnitude in the first half ot the trials distribution (increase from $n = 10$ to $n = 50$: $\Delta\text{CV}^{\text{Conf}} = -.06 \pm .00$; $\Delta\text{CV}^{\text{Acc}} = -.13 \pm .012$) but significantly decreases in the second half of the distribution (increase from $n = 50$ to $n = 100$: $\Delta\text{CV}^{\text{Conf}} = -.02 \pm .00$; $\Delta\text{CV}^{\text{Acc}} = -.03 \pm .00$; differences in increase between first and second half, $-.05 \pm .00; t_{102} = -31.4$; $p < .001$ for confidence; $-.09 \pm .01; t_{102} = -14.4$; $p < 0.001$ for accuracy; see the SI, Section 6 about the statistical tests used in the text). These results suggest that a reliable within-individual measure of confidence can be reached with 50 trials, and that additional trials only marginally (yet significantly from a statistical point of view) improve the reliability of the measure (also see the SI, Figure A).

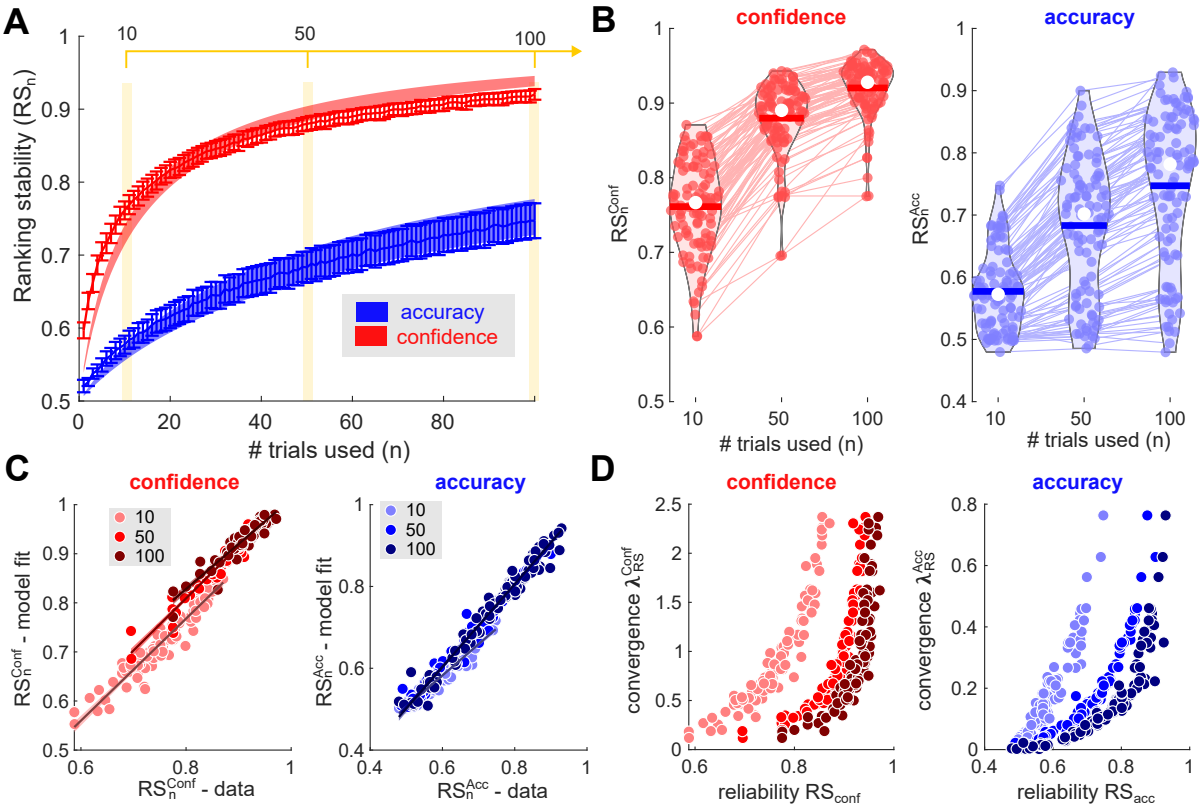
Although, according to these summary statistics, 50 trials appear to be sufficient on average, it is nonetheless possible that the individual reliability of different tasks exhibit different dynamics, such that some tasks are already relatively more reliable with fewer trials, or, on the contrary, only become relatively more reliable when more trials become available. To address this possibility and provide a full, parametric, characterization of the dynamics of reliability, we adapted a descriptive model inspired by the structural definition of the convergence curve proposed by Kadlec et al. (2024). To that end, the CV of confidence and accuracy are defined as a non-linear function of the number of trials and a convergence parameter, λ (see Methods, Section 5).

We fit this equation using non-linear least squares at the study, j, level. The resulting vector of study-specific convergence parameters, $[\boldsymbol{\lambda}_{\text{CV}}^{Conf}; \boldsymbol{\lambda}_{\text{CV}}^{Acc}]$, summarizes the convergence in both confidence and accuracy reliability as a function of the number of trials for each study using a single parameter (see Methods, Section 5). This descriptive model fits well both the between studies variations in confidence reliability, and their change according to the number of trials (Figure 4.C, left). It performs similarly well on fitting the reliability of accuracy (Figure 4.C, right). Most importantly, the estimated study-level parameter is tightly monotonically related to the reliability of studies as proxied by the CV computed with various number of trials (10, 50, 100), for both confidence and accuracy (Figure 4.D). This suggests that the relative reliability of tasks for within-individual confidence estimates evaluated by the CV is somewhat independent from the number of trials used to compute the CV, and that the convergence parameter provides a very accurate summary of the relative level of reliability achieved by the various studies (see also the SI, Figure B). Reciprocally, this also implies that the within-individual confidence measure reliability achieved by a task, as well as the convergence of its reliability curve, can be robustly approximated with CVs estimated with $n = 10, 50$, or 100 trials.

Most of the feasible between-individuals reliability is (also) achieved in less than 50 trials

We next turned to between-individual measures of reliability, i.e., the ability of averaged confidence reports to robustly differentiate individual confidence levels, which we assessed with the Ranking

Figure 5: Between-individuals reliability of confidence and accuracy measures



Note. Summary statistics on the between-individuals reliability of confidence and accuracy measures (measured by the Ranking Stability, i.e., the probability that any two individuals are ranked the same way in replications samples, see step 4b in Figure 3) as a function of the number of trials, n. **Panel A**: Reliability as a function of the number of trials. Lines and error bars indicate the empirical mean +/- 95% CI of the RS for accuracy (blue) and confidence (red). Shaded areas indicate the mean +/- 95% CI of fitted RS, obtained from the convergence model in (1), for accuracy (blue) and confidence (red). **Panel B**: Empirical distribution of reliability measures at $n = 10, 50$, and 100, for accuracy (left, blue) and confidence (right, red). Violin plots represent the sample distribution of RS. Connected, colored dots represent the estimates from each individual study. Horizontal bars indicate the sample means and white dots the sample medians. **Panel C**: Fitted reliability (λ_{RS}) of accuracy (left, blue) and confidence (right, red) by the convergence model in (1), as a function of empirically measured RS, at $n = 10, 50$, and 100 (resp. light, medium and dark coloured). Each dot represents an individual study, and the coloured lines correspond to the best fit of linear regressions, at $n = 10, 50$, and 100. **Panel D**: Correlation at the study level between the estimated convergence parameter from this same model and the empirical measures of reliability at $n = 10, 50$, and 100.

Stability. Note that the tasks in the confidence database are, for the most part, not primarily designed to assess between-individual differences in confidence. Besides, since between-individuals reliability relies on comparisons in the target measure (of either confidence or accuracy), this outcome cumulates the measurement noise over individuals that are compared. Nonetheless, our results show that the ranking stability again quickly increases (denoting an improvement in reliability) until reaching a plateau (from $RS_5^{Conf} = .71 \pm .01$ (95% CI) to $RS_{25}^{Conf} = .83 \pm .01$ and $RS_{50}^{Conf} = .88 \pm .01$; Figure 5.A). Again, choice accuracy exhibited a similar profile, but with a lower reliability on average (from $RS_5^{Acc} = .55 \pm .01$ to $RS_{25}^{Acc} = .63 \pm .02$ and $RS_{50}^{Acc} = .68 \pm .02$; Figure 5.A).

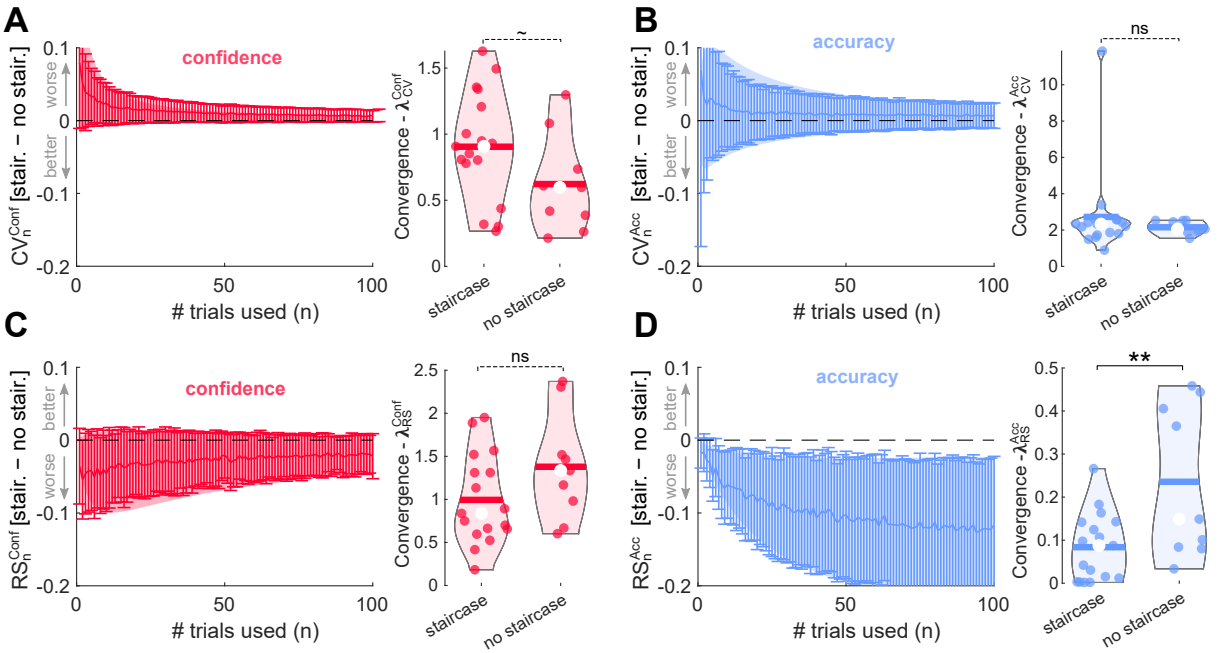
A detailed look at specific points on the curve ($n = 10, 50$ and 100, Figure 5.B) again confirmed that the increase in reliability is significant in magnitude in the bottom half of the trials distribution (from $n = 10$ to $n = 50$: $\Delta \text{RS}^{\text{Conf}} = .12 \pm .01$; $\Delta \text{RS}^{\text{Acc}} = .11 \pm .01$) but decreases in the top half of the distribution (from $n = 50$ to $n = 100$: $\Delta \text{RS}^{\text{Conf}} = .04 \pm .00$; $\Delta \text{RS}^{\text{Acc}} = .06 \pm .00$; differences in increases from first- to second-half $.04 \pm .01, t_{102} = 23.5$, $p < .001$ for confidence; $.04 \pm .01, t_{102} = 8.9$, $p < .001$ for accuracy). These results suggest that reliable measures of between-individual differences in confidence can be reached with 50 trials, and that additional trials only marginally improve the reliability of the measure (see also the SI, Figure A).

Again, we considered the possibility that the between-individual reliability of different tasks exhibit different dynamics with respect to the number of trials used to compute confidence and accuracy measures. To address this concern, we summarize this dynamics in reliability using a similar parametric model as in the previous section, by estimating the vectors of convergence parameters $[\boldsymbol{\lambda}_{\text{RS}}^{Conf}; \boldsymbol{\lambda}_{\text{RS}}^{Acc}]$ fitting the RS data based on a descriptive convergence model (see Methods, Section 5). For all values of $n = 10, 50$ and 100, the reliability predicted from these estimates almost perfectly coincides with the observed reliability of both confidence and accuracy (Figure 5.C). Again, study-level convergence parameters appeared tightly monotonically related to the relative level of reliability observed for different numbers of trials across studies, for both confidence and accuracy (see Figure 5.D). This suggests that the relative reliability of tasks for inter-individual confidence differences evaluated by the RS is somewhat independent from the number of trials used to compute the RS, and that the convergence parameter provides a very accurate summary of the relative level of reliability achieved by the various studies (see also the SI, Figure B). Reciprocally, this also implies that the between-individual confidence measure reliability achieved by a task, as well as the convergence of its reliability curve, can be robustly approximated with RSs estimated with $n = 10, 50$, or 100 trials.

Evaluating the effect of staircase and subjective confidence scales on individual accuracy and confidence measure reliability

Regarding both within- and between-individuals measures, 50 trials turns-out as a satisfactory rule of thumb to achieve reliable measures of both confidence and accuracy. Our results nonetheless show important disparities between studies, regarding the absolute level of reliability and the dynamic of reliability convergence as a function of the number of trials used to compute confidence measures (see the SI, Figure A). Because our resampling exercise in the confidence database naturally pooled data from heterogeneous studies, this raises the obvious question of the impact of specific task features on within- and between-individual reliability of confidence and accuracy measures. Some design choices could indeed mechanically decrease the ability of averaged trial reports to constitute a reliable within- or between-individual measure. We first focus on two main

Figure 6: Variations in reliability induced by staircase procedures

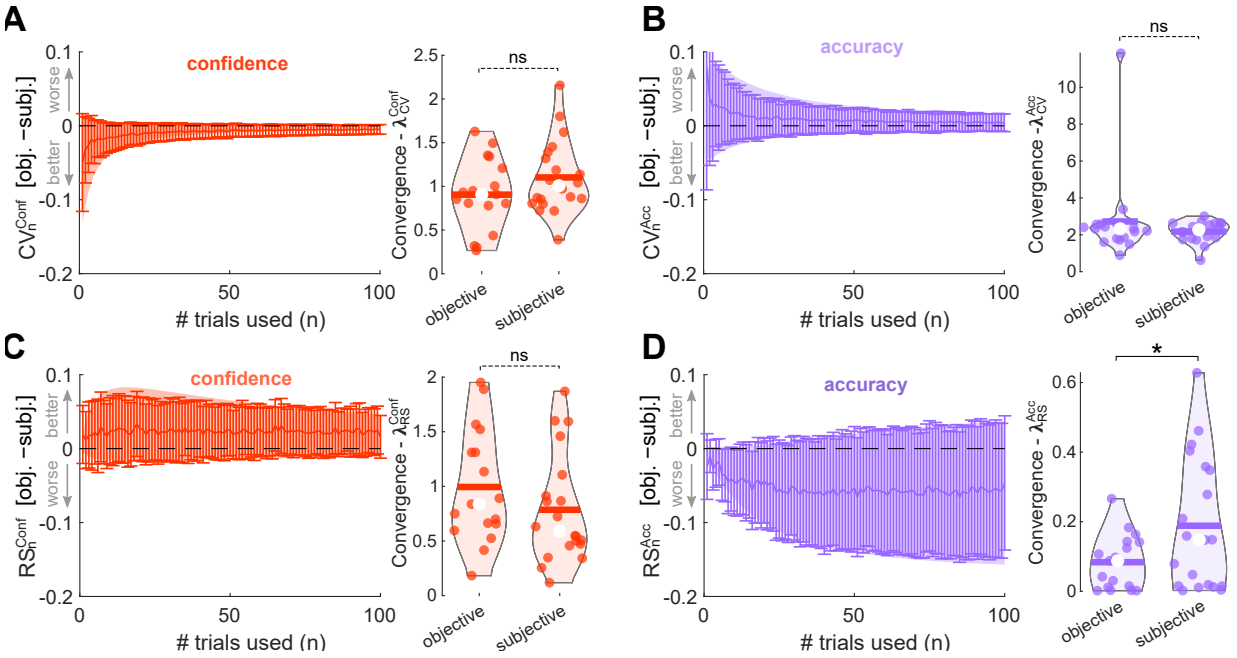


Note. Distribution (average value and 95% CI) of the contrast between studies implementing a staircase procedure to the cognitive task ($N = 9$ studies) and those which don't ($N = 17$), regarding the reliability of within-individual confidence (**Panel A**) and accuracy (**Panel B**), as well as between-individuals confidence (**Panel C**) and accuracy (**Panel D**). In each panel, the additional plots on the right-hand side provide the empirical distribution of the convergence parameter for the corresponding measure in each group. The sample is restricted to studies relying on a perceptual task with no feedback and an objective scale. See the SI, Figure C, for a replication on the entire sample.

task features, which we hypothesized could have such an effect: the presence of a staircase procedure for accuracy, and the nature of the confidence report (subjective or objective scale). Our rationale is as follows. On the one hand, staircase procedures aim at canceling all variations in Type-1 accuracy, between trials and between individuals, by varying the level of difficulty endogenously and individually so as to reach a predefined performance target. Therefore, the presence of this task feature should increase within-individuals and decrease between-individuals reliability of accuracy measures — though we remain agnostic about its effect on confidence measure. On the other hand, subjective confidence scales add a layer of subjective interpretation on how to report confidence: some individuals might be more reluctant than others to report "*highly confident*", because they interpret this level as "*almost certain*". This subjective layer should be absent for objective confidence scale, where e.g. "*I believe my choice to have an 80% chance of being correct*" should have the same meaning for all individuals. We therefore hypothesized that subjective confidence scales induce an additional individual bias, thereby increasing the between-individual reliability of confidence measures.

Because the non-experimental nature of the variations in the design of the confidence-elicitation task in the sample comes with strong limitations in our ability to isolate the consequences of each

Figure 7: Variations in reliability induced by the nature of the scale



Note. Distribution (average value and 95% CI) of the contrast between studies relying on an objective scale ($N = 17$ studies) and those relying on a subjective scale ($N = 20$), regarding the reliability of within-individual confidence (**Panel A**) and accuracy (**Panel B**), as well as between-individuals confidence (**Panel C**) and accuracy (**Panel D**). In each panel, the additional plots on the right-hand side provide the empirical distribution of the convergence parameter for the corresponding measure in each group. The sample is restricted to studies relying on a perceptual task with no feedback and implementing a staircase procedure. See the SI, Figure D, for a replication on the entire sample.

characteristic of the implementation, we evaluate the effects of staircase and confidence scale on reliability, while controlling for a limited number of other key implementation characteristics (see the SI, Section B, for a multivariate heterogeneity analysis of the reliability measures). The within-individual reliability of either accuracy or confidence does not seem to be much affected by the presence or absence of a staircase procedure though it appears to marginally reduce confidence reliability consistently throughout the range of trials used to compute the measure ($\boldsymbol{\lambda}^{Acc}_{\text{CV}}|_{\text{staircase}} = 2.72 \pm 1.24$ and $\boldsymbol{\lambda}^{Acc}_{\text{CV}}|_{\text{no staircase}} = 2.15 \pm .28$, difference: $t_{24} = .70$, $p = .493$; $\boldsymbol{\lambda}^{Conf}_{\text{CV}}|_{\text{staircase}} = 0.91p \pm .21$ and $\boldsymbol{\lambda}^{Conf}_{\text{CV}}|_{\text{no staircase}} = .62 \pm .28$, difference: $t_{24} = 1.72$, $p = .099$; Figure 6, Panels A and B). In contrast and as hypothesized, at the between-individual level, the reliability of accuracy is significantly lower when a staircase procedure applies to the target task ($\boldsymbol{\lambda}^{Acc}_{\text{RS}}|_{\text{staircase}} = .08 \pm .04$ and $\boldsymbol{\lambda}^{Acc}_{\text{RS}}|_{\text{no staircase}} = .24 \pm .14$, difference: $t_{24} = -3.04$, $p = .006$; Figure 6, Panels C and D), while the reliability of confidence remains somewhat unaffected or even marginally improved ($\boldsymbol{\lambda}^{Conf}_{\text{RS}}|_{\text{staircase}} = .99 \pm .27$ and $\boldsymbol{\lambda}^{Conf}_{\text{RS}}|_{\text{no staircase}} = 1.38 \pm .49$, difference: $t_{24} = -1.67$; $p = .108$). Note that these analyses replicate when using the full dataset rather than carefully controlling for the other heterogeneity dimensions (see the SI, Figure C).

The nature of the scale has little effect on the within-individual and between-individual reliability of accuracy (difference in $\boldsymbol{\lambda}_{\text{CV}}^{Acc}$: $t_{35} = .99$, $p = .331$; difference in $\boldsymbol{\lambda}_{\text{RS}}^{Acc}$, $t_{35} = -2.12$, $p = 0.041$; Figure 7, Panels B and D). The within-individual and between-individual reliability of confidence tends to be slightly higher with an objective scale ($\boldsymbol{\lambda}_{\text{CV}}^{Conf}|_{\text{subjective}} = 1.10 \pm 0.39$ and $\boldsymbol{\lambda}_{\text{CV}}^{Conf}|_{\text{objective}} = .91 \pm .43$ difference: $t_{35} = -1.43$, $p = .161$; $\boldsymbol{\lambda}_{\text{RS}}^{Conf}|_{\text{subjective}} = .79 \pm .23$ and $\boldsymbol{\lambda}_{\text{RS}}^{Conf}|_{\text{objective}} = .99 \pm .27$, difference: $t_{35} = 1.25$, $p = .221$; Figure 7, Panels A and C). This is consistent with the fact that subjective scales are likely interpreted heterogeneously by participants, hence generating more noisy data. Again, these analyses replicate when using the full dataset rather than carefully controlling for the other heterogeneity dimensions (see the SI, Figure D).

Discussion

The question of the reliability of behavioral measures has recently re-surfaced, notably applied to risk elicitation (Frey et al., 2017; Pedroni et al., 2017), cognitive control (Enkavi et al., 2019), and reinforcement-learning (Mkrtchian et al., 2023; Pike et al., 2022; Schurr et al., 2024; Vrizzi et al., 2023). Most have highlighted the fragility of behavioral measures especially when evaluated in test-retest setups, as well as their inability to capture relevant inter-individual variance (Hedge et al., 2018). Although confidence has been shown to be a primitive of a wide range of outcomes in behavioral sciences, from economics to management, psychology and neurosciences, little is still known about the reliability of confidence measures. Here, we filled this gap and evaluated the reliability of confidence measures obtained by averaging judgments about choice accuracy over multiple trials of typical cognitive tasks. Thanks to resampling exercises leveraging more than a hundred individual studies included in the Confidence Database, we show that reliable individual measures can be robustly obtained when averaging confidence judgments over 50 trials or more.

We extend our resampling exercise to the reliability of comparisons in confidence levels between individuals, hence asking whether a stable confidence ranking can be inferred from individual confidence measures. This is key to the large strand of research that aims at explaining differences in behavior between individuals by differences in individual characteristics, from cognitive neurosciences to clinical or political psychology (Lebreton et al., 2019). Although such rankings derive from comparisons between individual confidence measures that are themselves noisy, we show that a similar rule of thumb still applies to the reliability of between-individuals confidence measures. Once again, the increase in reliability achieved when increasing the number of trials beyond 50 is at most marginal.

We also replicated our analysis on a second dimension of participants' behavior, the performance — or choice accuracy — at the primary task, which gives us an opportunity to contrast the reliability of confidence with the reliability of the behavioral output that confidence is meant to evaluate. At both the within- and between-individual levels, we show that the convergence

of the reliability of both accuracy and confidence are parallel, hence providing support to the literature investigating the empirical correlation between the two (Jin et al., 2022), and for theoretical models building metacognitive evaluation on variables that are shared with the type-1 decision-processes (Fleming and Daw, 2017). Our results nonetheless show that the reliability of accuracy is systematically significantly lower than that of confidence, regardless of the dimension (within or between-individual) considered, hence contributing to the recent literature challenging the reliability of (Type-1) cognitive tasks (Kadlec et al., 2024; Vrizzi et al., 2023; Enkavi et al., 2019; Pedroni et al., 2017). In addition, the differences in reliability between choice accuracy and confidence measures raise a cautionary note on the interpretation of observed dissociations between these two dimensions, e.g., regarding the effect of heterogeneity factors like gender.

At both the within- and between-individual levels, we also show that the marginal benefit in terms of reliability of increasing the number of trials above 50 is typically very small. While rarely addressed explicitly, the optimization of trial number is inherent in almost any behavioral task design: one typically looks for the minimal number of trials that can deliver a reliable assessment of the behavioral phenomenon of interest. The cost of adding trials can dramatically vary depending on the setup and potential application of the behavioral elicitation of interest. Importantly, such behavioral assessments are increasingly included in the context of longitudinal study, -notably web or smartphone-based - where the attention span of participants is shorter (Crump et al., 2013; Hauser et al., 2022), in the context of costly or precisely-timed intervention (neuroimaging, pharmacological intervention with specific pharmacodynamics), or with vulnerable population who cannot engage in longer tasks (kids, elderly, patients suffering from various motor or cognitive pathologies). In all these cases, a proper evaluation of the marginal benefit on behavioral assessment reliability of increasing the number of trials could inform about optimal parameters of task designs. Besides, optimizing tasks for a lower number of trials contribute to minimizing the uncontrolled variations of practice effects, mood or attention (Desender et al., 2022; Weilnhammer et al., 2023; Mei et al., 2023). We suggest that quantification exercises similar to the one we proposed in the present report could therefore positively contribute to behavioral sciences.

Our conclusions build from a heterogeneous set of studies, covering a large variety of populations and a wide range of cognitive and elicitation tasks. Although these variations happen in a non-experimental way in our sample, our multivariate analysis shows these results are highly robust to important design and implementation characteristics. The pool of studies contained in the confidence database also allows us to analyze more precisely two important dimensions of confidence elicitation: the implementation of a staircase procedure and the nature of the scale. We find little variation in the convergence of both kinds of confidence reliability across these dimensions of design heterogeneity; although, as expected, a staircase procedure tends to slow down the convergence of between-individual measures of accuracy. Yet, currently, the confidence

database still only contains limited information regarding other important task factors like, e.g., the provision and characteristics of feedback (Haddara and Rahnev, 2022) or the incentivization of confidence measures (Lebreton et al., 2018; Smith and Walker, 1993; Schlag et al., 2015), which prevented the extension of our analyses to those dimensions.

Overall, our results confirm that individual-specific psychological factors can be extracted from confidence elicitation tasks when a large enough number of trials is considered. They constitute a useful benchmark to guide the design of confidence measures in future works. Given the minor effect of key task design factors on our main results, our conclusions thus likely generalize to a broader range of experimental paradigms the results obtained by Fox et al. (2024), whose innovative gamified application has been shown to produce stable individual confidence estimates that correlate with inter-individual dimensions of anxiety-depression or compulsivity and intrusive thoughts within 40 trials, and of Binnendyk and Pennycook (2023) who show that participants' overconfidence estimated in 10 trials of a difficult perception test reliably predicts of a host of behavioural outcomes, including conspiracy beliefs, bullshit receptivity, overclaiming, and the ability to discern news headlines.

In contrast with these studies, we could not extend our reliability analysis to investigations of the correlation between confidence measures and individual behavior or psychometric profile, due to the limited information available in the confidence database. This illustrates the inherent complementarity between researchers' ability to assess the generalizability of behavioral findings to different tasks, and their ability to robustly assess the internal and external validity of the proposed behavioral measures: lab-like experiments similar to Fox et al. (2024) and Binnendyk and Pennycook (2023) allow for a rich evaluation of individuals' characteristics, but are generally limited in the variety of behavioral tasks that they can propose to each participant; a database approach like ours allow to test for the generalizability of task-measure robustness with an unprecedented breadth, but are inherently limited in their access to standardized measures of individual characteristics that would allow to evaluate the external validity of the measures. Relatedly, our approach is restricted to assessing the reliability of the various measures used in the literature to capture confidence, but remains agnostic about the internal validity of the underlying construct itself. Addressing these questions, in terms of both the external validity of confidence measures, and the internal validity of the construct(s) targeted by these measures, require to combine the two approaches by collecting large-scale data using a comprehensive set of both confidence measures and individual characteristics. This is next on our agenda.

Material and Methods

1 Data

We used data included in the "Confidence Database" up to July 2023, and available at https://osf.io/s46pr/. This consisted in individual data of 171 studies, along with a set of commonly formatted variables among which: the number of trials per subject, the stimulus-response pairings for every trial, and a measure of participant confidence recorded for each trial (albeit using a variety of confidence scales). Depending on the study, supplementary variables such as reaction times, task difficulty, feedback mechanisms, and participant demographics were also available. Such variables, that were not uniformly available across all datasets, were excluded from our analysis.

2 Coding of variables

The confidence database documents several design heterogeneity factors characterizing the different studies: a classification of the task category (cognition, perception, memory, motor, or mixed domains), of the confidence scale (e.g., continuous or n-point), and of the stimulus type (*e.g.*, Gabords/ellipse, letter and colors, ball throw), of the granularity of the confidence scale (i.e. number of possible confidence levels). It also includes information on whether confidence judgments were elicited simultaneously with decision-making and on any additional manipulation specific to the study (e.g., variations in the task difficulty). We computed a set of additional variables based on the information available either from the study-specific CD readme files or from the published manuscript: a classification of the type of feedback, whether the task difficulty is adjusted using a staircase procedure, and whether this scale is subjective or objective. To create this variable, we define "*objective*" scales as those associated with a probabilistic interpretation, i.e., for which confidence is expressed as a percentage chance of correctness (ranging typically from 0 to 100 or 50 to 100, or conveyed through labels such as "*chance/guess*" to "*certain*"). In contrast, scales that lack a probabilistic basis are recorded as "*subjective*"; this includes measures featuring non-probabilistic extremes like "*high vs. low*" or "*unsure vs. sure*", or no specific interpretation at all (represented, e.g., by a slider or a numerical value).

Note that in some exceptional cases, studies could not be unambiguously coded on specific dimensions, or featured characteristics that were not shared by enough studies to constitute a meaningful category (classified as "other" in Figure 2); Scale: 3 studies elicited confidence as wager rather than ratings, hence were not classified as objective or subjective. Granularity: 2 studies feature a 5-point scale, 1 study features a 9-point scale, 3 studies feature a 11-point scale, and 1 study alternates between a 4-point scale, and a continuous scale. Feedback: 5 studies alternate between feedback and no-feedback.

3 Exclusion criteria and final dataset

The re-sampling exercise that generates the four main outcomes analyzed in this study (within- and between-individuals reliability of both confidence and accuracy) requires to observe a total number of trials that exceeds the sample size of the re-sampling — at the risk of artificially inflating the correlation between draws otherwise. Since we consider reliability for up to 100 trials, we restrict our sample to studies containing at least 150 trials per individual. In the resulting set of 137 studies, we also excluded (13) studies whose measure of accuracy is non-binary, since the inclusion of these studies would have implied to arbitrarily choose an accuracy threshold. An additional 21 studies were excluded due to data availability issues. Specifically, we excluded 7 studies in which visibility or subjective difficulty were elicited instead of confidence, 8 studies because their design was too far from other studies (e.g., the confidence scale changes across trials) and 6 studies because of data quality concerns. Overall, there are $6,024$ participants and $25,438$ trials in total in the excluded studies, which represents around 22% of the CD data.

4 Measures of reliability

The study aims at documenting the link between the number of trials, n, and the reliability of the measures of both confidence and accuracy, denoted $y_{i,t}$, $y \equiv \{Conf; Acc\}$, for each individual i in each trial t. Given the distribution of the total number of trials in the dataset, we restrict our analysis to a maximum number of 100 trials (also see Section 3 below).

At the within-individual level, our measure of reliability is the coefficient of variation, defined as the ratio between the standard deviation and the sample mean of a sequence of random draws. Assuming that trials are random draws in the distribution of confidence for each individual, we build our measures based on a resampling exercise at the subject-trial level. For each possible value of $n = 1, \ldots, 100$, we generate for each subject $i = 1, \ldots, N$, a total of $R = 15$ samples of n randomly drawn (with replacement) trials. We then compute the average of the coefficient of variation at the individual level over all replications,

$$\mathrm{CV}^{y}_{i,n} = \frac{1}{R} \cdot \sum_{r=1}^{R} \mathrm{CV}^{y}_{i,n,r} = \frac{1}{R} \cdot \sum_{r=1}^{R} \frac{\sum_{t=1}^{n} (y_{i,t} - 1/n \sum_{t=1}^{n} y_{i,t})^2}{\sum_{t} y_{i,t}}$$

This procedure is applied to both outcomes, and thus generates a panel dataset of the coefficient of variation in confidence and accuracy at the individual level for each possible value of n.

At the between-individuals level, our notion of reliability focuses on the ordinal ranking of subjects, *i.e.*, whether a given number of trials is enough to elicit meaningful comparisons within the sample regarding the level of confidence and accuracy. To that end, we apply a similar re-sampling exercise to the computation of Kendall correlation coefficients for all possible values of the total number of trials, n. Specifically, for each possible value of $n = 1, \ldots, 100$, we generate

for each subject $i = 1, \ldots, N_j$ in study j, a total of $R = 15$ samples of n randomly drawn (with replacement) trials. We then compute the mean of the outcome over trials in each replication, and the full set of correlation coefficients between all pairs of individuals participating to the same study over all combinations of replications.

We measure these correlations using the Kendal's τ rather than Spearman's correlation. Although the two measures are similar, the Kendall measure is generally considered as more robust and better suited to non-continuous data (see, e.g., Kruskal, 1958). The correlation is computed based on the number of "concordant" pairs. Given the two samples defined by all pairs of replications for two individuals i and l, $\{(\boldsymbol{y}_{i_n}; \boldsymbol{y}_{l_n})\}$, two pairs are concordant if the ranking between the two variables is the same in the two pairs, and discordant otherwise. The Kendall's τ^y of the corresponding outcome y is defined as the difference in the number of occurrences of these two cases divided by the total number of different pairs in the sample. It thus lies between 0 and 1 and equals 0 for independent variables.

The measure of between-individual reliability is computed as the average of this quantity at the study, $j = 1, ..., J$, level:

$$\begin{aligned} \text{RS}^y_{j,n} &= \frac{1}{N_j(N_j-1)/2} \times \sum_{\{i,l\} \in j; i \neq l} \tau^y_{i_n l_n} \\ &= \frac{1}{N_j(N_j-1)/2} \times \sum_{\{i,l\} \in j; i \neq l} \times \frac{\sum_{r=1}^{R} \sum_{m=1}^{R} \mathbb{1}_{[y_{i,r} < y_{i,m}]} + \mathbb{1}_{[y_{i,r} < y_{i,m}]}}{R(R-1)/2} \end{aligned}$$

This measure is closer to 1 the more stable across replications is the relative ranking in the outcome between any two individuals who participated to the same study.

5 Structural models of reliability convergence

We adapted a descriptive model inspired by the structural definition of the convergence curve proposed by Kadlec et al. (2024). This study investigates the reliability of individual behavioral measures obtained from various cognitive tasks based on the mean Pearson's correlation (P) across participants on different subsets of data from a given behavioral measure. Assuming there is no learning (i.e., samples are independent, and two consecutive trials are independent), and assuming each participant i has a true proficiency (i.e., a true ability level for a given task), they show that the dynamic of this measure of reliability is related to the number of trials, n, and the single trial error variance over true-score variance in classical test theory, V_i, as (see derivations in Kadlec et al., 2024):

$$P_{i,n} = \frac{n}{n + V_i} \tag{1}$$

As a result, participants are expected to each converge to a stable mean reflecting this proficiency. In this case, as we average across more trials, reliability at the individual level will increase, which will drive an increase in reliability across individuals. Because our setup and reliability measures slightly differ from these definitions, we adjusted the convergence as follows

For within-individual reliability convergence, because the coefficient of variation corresponds to an inverse measure of reliability (the higher the measure, the more variance between the same measure estimated from various samples), we defined the following convergence descriptive model:

$$\mathrm{CV}^{y}_{i,n} = \frac{\lambda^{y}_{\mathrm{CV},i}}{n + \lambda^{y}_{\mathrm{CV},i}},\ y \in \{Conf, Acc\}, \forall i$$

from which we estimate the study-specific within-individual convergence parameters $\lambda^{y}_{\mathrm{CV},i}$.

For between-individual reliability convergence, because our rank stability is bounded between 0.5 and 1, we adapted the original convergence model as follows:

$$\mathrm{RS}^{y}_{i,n} = 0.5 + 0.5 \times \frac{n}{n + \delta^{y}_{\mathrm{RS},i}},\ y \in \{Conf, Acc\}, \forall i$$

For plotting and statistics, we define the corresponding study-specific between-individual convergence parameters using the rescaled transformation $\lambda^{y}_{\mathrm{RS},i} = 10/\delta^{y}_{\mathrm{RS},i}$, as it is more directly linked to reliability estimated with various trial numbers (Figure 5, Panel D), was more normally distributed in our sample, and had a similar interpretation as the parameter estimated in the within-individual dimension. Note that in both cases, the descriptive convergence models lose their link with classical test theory (which they have in Kadlec et al., 2024), and are only used as satisfactory, one-parameter descriptive models of reliability convergence. The models were fit to the data using non-linear least-square methods (`fitnlm` in Matlab®).

6 Statistics

Unless otherwise specified, the statistical assessment of the difference of variables as a function of the number of trials is based on within-individuals paired t-test; the assessment of the effects of task characteristics (contrast approach) is based on between-individuals two-sample t-test.

References

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. Quarterly Journal of Economics, 117(3):871–915.

Berner, E. S. and Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. American Journal of Medicine, 121(5):S2–S23.

Bhandari, G. and Deaves, R. (2006). The demographics of overconfidence. Journal of Behavioral Finance, 7(1):5–11.

Binnendyk, J. and Pennycook, G. (2023). Individual differences in overconfidence: A new measurement approach. Available at SSRN 4563382.

Boundy-Singer, Z. M., Ziemba, C. M., and Goris, R. L. T. (2023). Confidence reflects a noisy decision reliability estimate. Nature Human Behaviour, 7(1):142–154.

Cook, D. A. and Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. American Journal of Medicine, 119(2):166–e7.

Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS One, 8(3):e57410.

Desender, K., Vermeylen, L., and Verguts, T. (2022). Dynamic influences on static measures of metacognition. Nature communications, 13(1):4208.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., and 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.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. Annual Review of Psychology, 75(1):241–268.

Fleming, S. M. and Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological Review, 124(1):91.

Fox, C. A., McDonogh, A., Donegan, K. R., Teckentrup, V., Crossen, R. J., Hanlon, A. K., Gallagher, E., Rouault, M., and Gillan, C. M. (2024). Reliable, rapid, and remote measurement of metacognitive bias. Scientific Reports, 14(1):14941.

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., and Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. Science Advances, 3(10):e1701381.

Guggenmos, M. (2021). Measuring metacognitive performance: Type 1 performance dependence and test-retest reliability. Neuroscience of Consciousness, 2021(1):niab040.

Guggenmos, M. (2022). Reverse engineering of metacognition. ELife, 11:e75420.

Haddara, N. and Rahnev, D. (2022). The impact of feedback on perceptual decision-making and metacognition: Reduction in bias but no change in sensitivity. Psychological Science, 33(2):259–275.

Hauser, T. U., Skvortsova, V., De Choudhury, M., and Koutsouleris, N. (2022). The promise of a model-based psychiatry: Building computational models of mental ill health. The Lancet Digital Health, 4(11):e816–e828.

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior Research Methods, 50(3):1166–1186.

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., and van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. Translational Psychiatry, 9(1):268.

Hoven, M., Rouault, M., van Holst, R., and Luigjes, J. (2023). Differences in metacognitive functioning between obsessive–compulsive disorder patients and highly compulsive individuals from the general population. Psychological Medicine, 53(16):7933–7942.

Huys, Q. J. M., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience, 19(3):404–413.

Jin, S., Verhaeghen, P., and Rahnev, D. (2022). Across-subject correlation between confidence and accuracy: A meta-analysis of the Confidence Database. Psychonomic Bulletin & Review, 29(4):1405–1413.

Kadlec, J., Walsh, C. R., Sadé, U., Amir, A., Rissman, J., and Ramot, M. (2024). A measure of reliability convergence to select and optimize cognitive tasks for individual differences research. Communications Psychology, 2(1):1–18.

Karvelis, P., Paulus, M. P., and Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. Neuroscience & Biobehavioral Reviews, page 105137.

Koellinger, P., Minniti, M., and Schade, C. (2007). "I think I can, I think I can": Overconfidence and entrepreneurial behavior. Journal of Economic Psychology, 28(4):502–527.

Kruskal, W. H. (1958). Ordinal Measures of Association. Journal of the American Statistical Association, 53(284):814–861.

Lebreton, M., Bavard, S., Daunizeau, J., and Palminteri, S. (2019). Assessing inter-individual differences with task-related functional neuroimaging. Nature Human Behaviour, 3(9):897–905.

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., van Holst, R. J., and Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. Science Advances, 4(5):eaaq0668.

Lehmann, M., Hagen, J., and Ettinger, U. (2022). Unity and diversity of metacognition. Journal of Experimental Psychology: General, 151(10):2396.

Lundeberg, M. A., Fox, P. W., and Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. Journal of Educational Psychology, 86(1):114.

Malmendier, U. and Tate, G. (2005). CEO Overconfidence and Corporate Investment. Journal of Finance, 60(6):2661–2700.

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. PeerJ, 7:e6918.

Mazancieux, A., Fleming, S. M., Souchay, C., and Moulin, C. J. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. Journal of Experimental Psychology: General, 149(9):1788.

Mazancieux, A., Pereira, M., Faivre, N., Mamassian, P., Moulin, C. J. A., and Souchay, C. (2023). Towards a common conceptual space for metacognition in perception and memory. Nature Reviews Psychology, 2(12):751–766.

Mei, N., Rahnev, D., and Soto, D. (2023). Using serial dependence to predict confidence across observers and cognitive domains. Psychonomic Bulletin & Review, 30(4):1596–1608.

Mkrtchian, A., Valton, V., and Roiser, J. P. (2023). Reliability of decision-making and reinforcement learning computational parameters. Computational Psychiatry, 7(1):30.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing Self-Confidence: Theory and Experimental Evidence. Management Science, 68(11):7793–7817.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., and Bahrami, B. (2017). The idiosyncratic nature of confidence. Nature Human Behaviour, 1(11):810–818.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? Quarterly Journal of Economics, 122(3):1067–1101.

Ortoleva, P. and Snowberg, E. (2015). Overconfidence in Political Behavior. American Economic Review, 105(2):504–535.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., and Rieskamp, J. (2017). The risk elicitation puzzle. Nature Human Behaviour, 1(11):803–809.

Pike, A. C., Tan, K., Ansari, H. J., Wing, M., and Robinson, O. J. (2022). Test-retest reliability of affective bias tasks. PsyArXiv Preprints.

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., Double, K. S., Duyan, Y. A., Faivre, N., Fallow, K., Filevich, E., Gajdos, T., Gallagher, R. M., de Gardelle, V., Gherman, S., Haddara, N., Hainguerlot, M., Hsu, T.-Y., Hu, X., Iturrate, I., Jaquiery, M., Kantner, J., Koculak, M., Konishi, M., Koß, C., Kvam, P. D., Kwok, S. C., Lebreton, M., Lempert, K. M., Ming Lo, C., Luo, L., Maniscalco, B., Martin, A., Massoni, S., Matthews, J., Mazancieux, A., Merfeld, D. M., O'Hora, D., Palser, E. R., Paulewicz, B., Pereira, M., Peters, C., Philiastides, M. G., Pfuhl, G., Prieto, F., Rausch, M., Recht, S., Reyes, G., Rouault, M., Sackur, J., Sadeghi, S., Samaha, J., Seow, T. X. F., Shekhar, M., Sherman, M. T., Siedlecka, M., Skóra, Z., Song, C., Soto, D., Sun, S., van Boxtel, J. J. A., Wang, S., Weidemann, C. T., Weindel, G., Wierzchoń, M., Xu, X., Ye, Q., Yeon, J., Zou, F., and Zylberberg, A. (2020). The Confidence Database. Nature Human Behaviour, 4(3):317–325.

Rollwage, M., Dolan, R. J., and Fleming, S. M. (2018). Metacognitive Failure as a Feature of Those Holding Radical Beliefs. Current Biology, 28(24):4014–4021.e8.

Rollwage, M., Zmigrod, L., de-Wit, L., Dolan, R. J., and Fleming, S. M. (2019). What Underlies Political Polarization? A Manifesto for Computational Political Psychology. Trends in Cognitive Sciences, 23(10):820–822.

Rouault, M., Lebreton, M., and Pessiglione, M. (2023). A shared brain system forming confidence judgment across cognitive domains. Cerebral Cortex, 33(4):1426–1439.

Rouault, M., McWilliams, A., Allen, M. G., and Fleming, S. M. (2018a). Human metacognition across domains: Insights from individual differences and neuroimaging. Personality Neuroscience, 1:e17.

Rouault, M., Seow, T., Gillan, C. M., and Fleming, S. M. (2018b). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. Biological Psychiatry, 84(6):443–451.

Russo, J. E., Schoemaker, P. J., et al. (1992). Managing overconfidence. Sloan Management Review, 33(2):7–17.

Salem-Garcia, N., Palminteri, S., and Lebreton, M. (2023). Linking confidence biases to reinforcement-learning processes. Psychological Review, 130(4):1017.

Scheinkman, J. A. and Xiong, W. (2003). Overconfidence and Speculative Bubbles. Journal of Political Economy, 111(6):1183–1220.

Schlag, K. H., Tremewan, J., and Van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. Experimental Economics, 18:457–490.

Schurr, R., Reznik, D., Hillman, H., Bhui, R., and Gershman, S. J. (2024). Dynamic computational phenotyping of human cognition. Nature Human Behaviour, 8(5):917–931.

Smith, V. L. and Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. Economic Inquiry, 31(2):245–261.

Vrizzi, S., Najar, A., Lemogne, C., Palminteri, S., and Lebreton, M. (2023). Comparing the test-retest reliability of behavioral, computational and self-reported individual measures of reward and punishment sensitivity in relation to mental health symptoms. PsyArXiv Preprints.

Weilnhammer, V., Stuke, H., Standvoss, K., and Sterzer, P. (2023). Sensory processing in humans and mice fluctuates between external and internal modes. PLoS Biology, 21(12):e3002410.

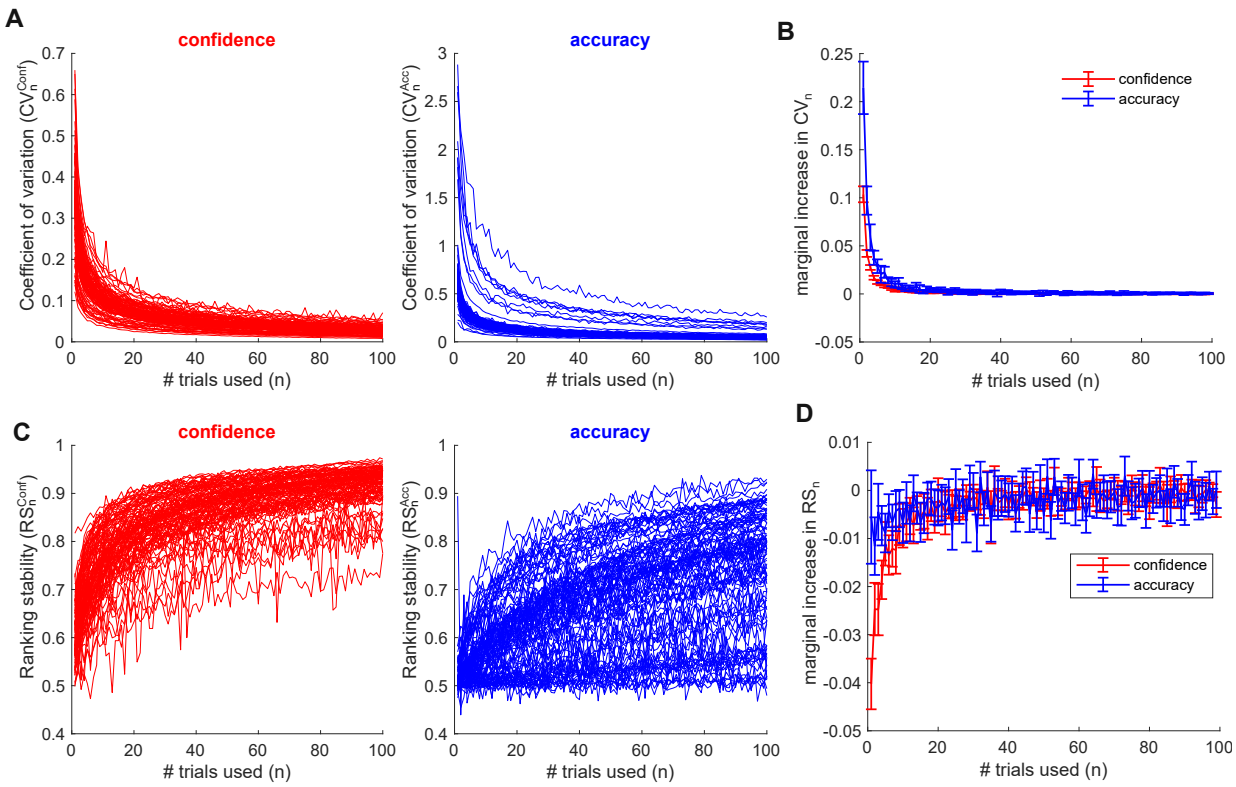
West, R. K., Harrison, W. J., Matthews, N., Mattingley, J. B., and Sewell, D. K. (14 juil. 2023). Modality independent or modality specific? Common computations underlie confidence judgements in visual and auditory decisions. PLOS Computational Biology, 19(7):e1011245.

Zimmermann, F. (2020). The Dynamics of Motivated Beliefs. American Economic Review, 110(2):337–361.

Supplementary Information

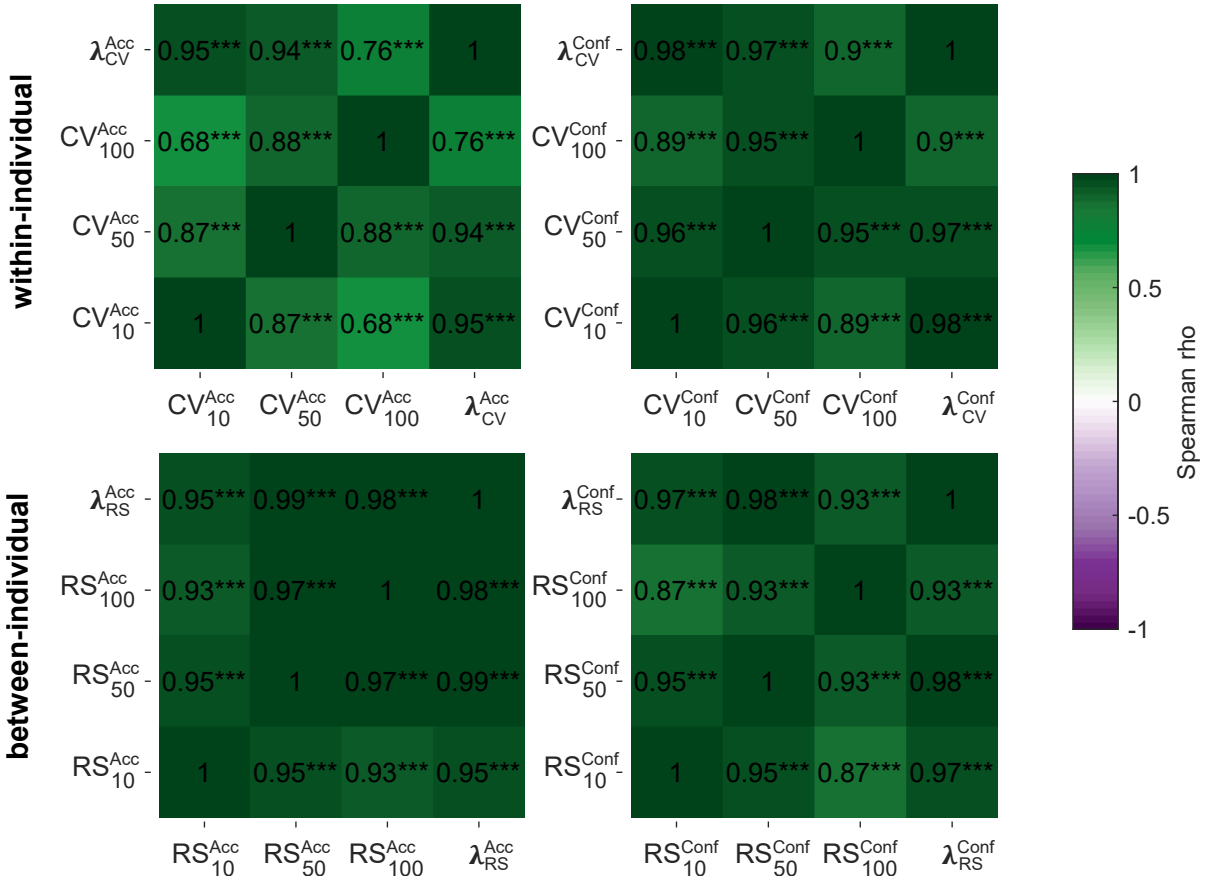
A Additional figures

Figure A: Reliability at the study level, and marginal increase from an additional trial



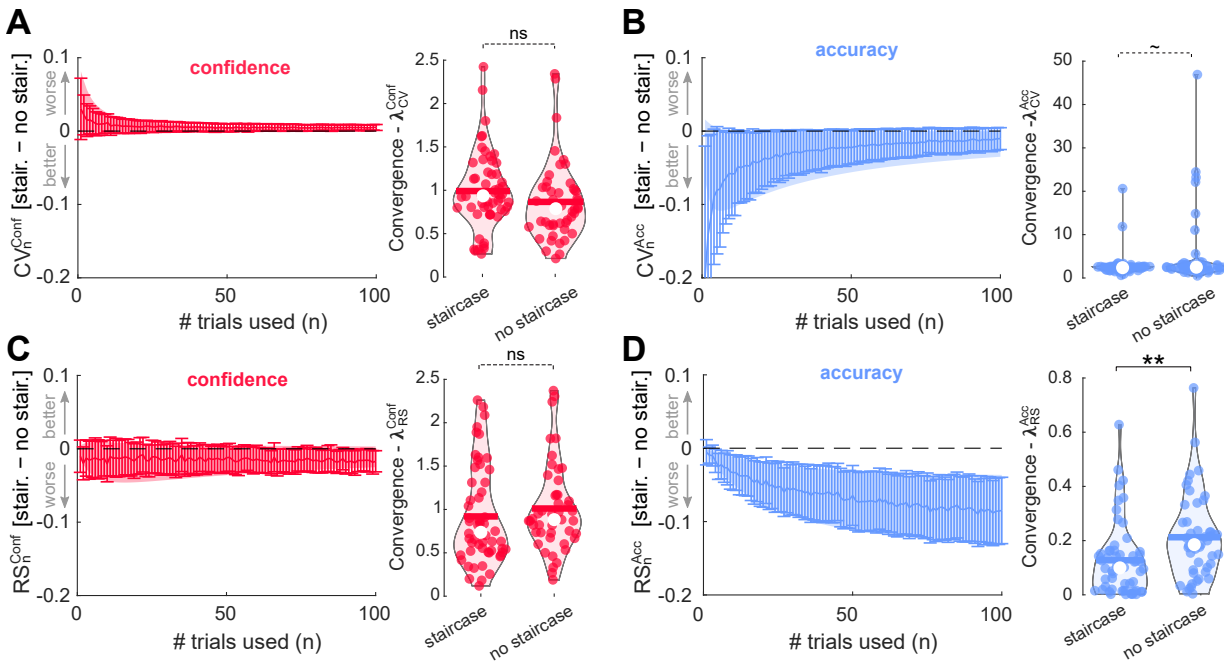
Note. Panel A: Empirical distribution of the measure of within-subjects reliability (CV) at the study level for both confidence (left-panel) and accuracy (right-panel). **Panel B**: Marginal increase (and 95% CI) in the CV (computed as $\Delta \text{CV}_n = \text{CV}_n - \text{CV}_{n-1}$) of both confidence and accuracy as a function of the number of trials (n). **Panel C**: Empirical distribution of the measure of between-subjects reliability (RS) at the study level for both confidence (left-panel) and accuracy (right-panel). **Panel B**: Marginal increase (and 95% CI) in the RS (computed as $\Delta \text{RS}_n = \text{RS}_n - \text{RS}_{n-1}$) of both confidence and accuracy as a function of the number of trials (n).

Figure B: Correlations between reliability measures



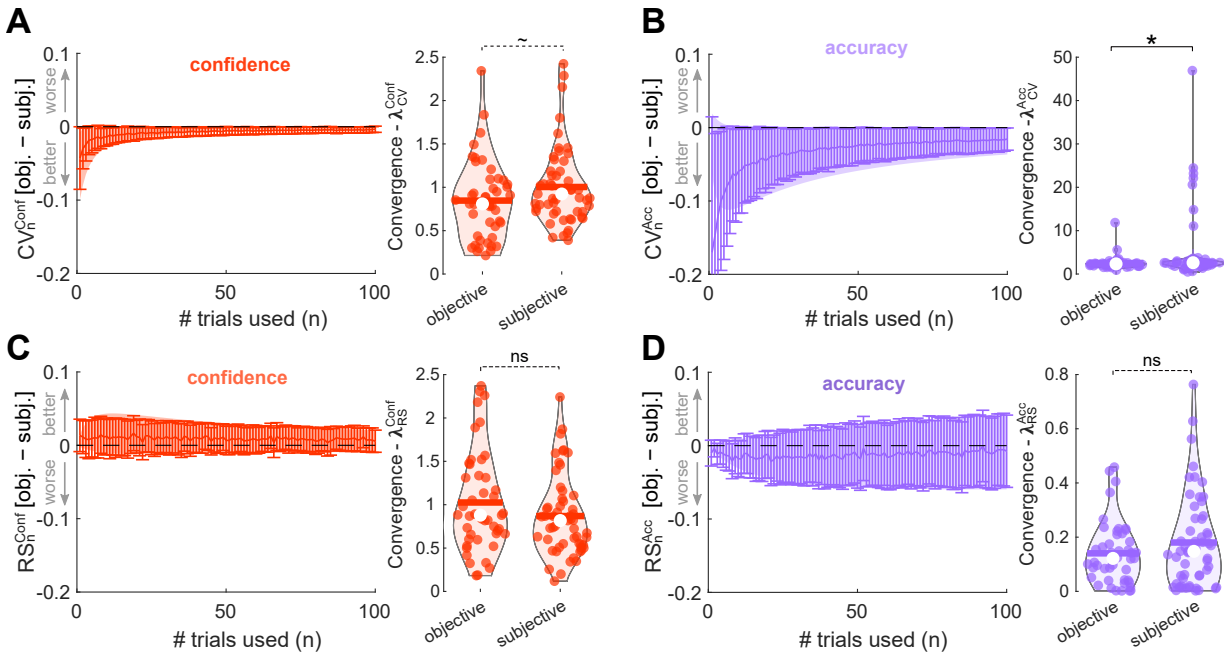
Note. For both accuracy (*left-hand side*) and confidence (*right-hand side*), each matrix reports the Spearman correlations between measures of reliability (within-individual in the *top panel*, CV_n; between-individuals in the *bottom panel*, RS_n) computed using different number of trials, n. *Significance levels*: *** 10%, ** 5%, *** 1%

Figure C: Variations in reliability induced by staircase procedures: robustness to study selection



Note. Replication of Figure 6 on the entire sample. The figure reports the distribution (average value and 95% CI) of the contrast between studies implementing a staircase procedure to the cognitive task ($N = 56$ studies) and those which don't ($N = 47$), regarding the reliability of within-individual confidence (**Panel A**) and accuracy (**Panel B**), as well as between-individuals confidence (**Panel C**) and accuracy (**Panel D**). In each panel, the additional plots on the right-hand side provide the empirical distribution of the convergence parameter for the corresponding measure in each group.

Figure D: Variations in reliability induced by the nature of the scale: robustness to study selection



Note. Replication of Figure 7 on the entire sample. Distribution (average value and 95% CI) of the contrast between studies relying on an objective scale ($N = 45$ studies) and those relying on a subjective scale ($N = 55$) , regarding the reliability of within-individual confidence (**Panel A**) and accuracy (**Panel B**), as well as between-individuals confidence (**Panel C**) and accuracy (**Panel D**). In each panel, the additional plots on the right-hand side provide the empirical distribution of the convergence parameter for the corresponding measure in each group.

B Heterogeneity analysis on the convergence parameters

In order to disaggregate reliability according to the characteristics of the task elicitation, we use the set of convergence parameters estimated from (1) as a summary statistic of study-specific dynamics in reliability and regress these parameters on a set of covariates documenting design characteristics. We estimate separate models for each of the four vectors of convergence parameters in $[\boldsymbol{\lambda}_{\text{CV}}^{Conf}; \boldsymbol{\lambda}_{\text{CV}}^{Acc}; \boldsymbol{\lambda}_{\text{RS}}^{Conf}; \boldsymbol{\lambda}_{\text{RS}}^{Acc}]$. The results from Gamma Generalized Linear Models with robust standard errors are presented in Tables A—D.

Table A: Gamma regressions of $\boldsymbol{\lambda}_{\text{CV}}^{Conf}$ on design characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
# of participants	0.002*	0.002**	0.001***	0.001***	0.001**	0.001**
	(0.096)	(0.030)	(0.006)	(0.001)	(0.014)	(0.031)
150-300 Trials	-0.287***	-0.307***	-0.247***	-0.193***	-0.175***	-0.167***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
300-500 Trials	-0.164**	-0.133*	-0.126**	-0.099**	-0.101**	-0.092*
	(0.028)	(0.083)	(0.033)	(0.023)	(0.029)	(0.072)
500-700 Trials	-0.241***	-0.221***	-0.207***	-0.193***	-0.179***	-0.174***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Memory		0.108	0.189***	0.189***	0.151**	0.159**
		(0.103)	(0.008)	(0.004)	(0.030)	(0.034)
Motor		-0.192**	-0.316***	-0.599***	-0.302***	-0.319***
		(0.030)	(0.001)	(0.000)	(0.000)	(0.000)
Perception		-0.149**	-0.123**	-0.146***	-0.148***	-0.142**
		(0.012)	(0.039)	(0.007)	(0.006)	(0.011)
Staircase			0.227***	0.197***	0.204***	0.209***
			(0.000)	(0.000)	(0.000)	(0.000)
Binary scale				0.050	0.193*	0.184*
				(0.548)	(0.052)	(0.055)
Discrete scale				-0.222***	-0.073	-0.076
				(0.000)	(0.388)	(0.338)
Objective scale					0.082	0.072
					(0.188)	(0.245)
Feedback						0.043
						(0.563)
Constant	0.382***	0.511***	0.387***	0.576***	0.416***	0.402***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	103	103	103	103	100	95

Note. Estimated coefficients (and p-values in parenthesis, computed using robust standard errors) from Gamma Generalized Linear Models of the relation between design characteristics and the estimated value of the convergence parameter $\boldsymbol{\lambda}_{\text{CV}}^{Conf}$. The reference study features more than 700 trials of a task belonging to a mixed domain, measures confidence on a subjective and continuous scale and does not not implement either a staircase procedure or feedback. Changes in the number of observations across columns are due to missing data in the variables that are added, see Section 2 above. *Significance levels*: *** 10%, ** 5%, *** 1%.

Table B: Gamma regressions of $\boldsymbol{\lambda}_{\text{CV}}^{Acc}$ on design characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
# of participants	0.001**	0.001	0.001*	0.001*	0.001*	0.001
	(0.039)	(0.118)	(0.098)	(0.069)	(0.098)	(0.147)
150-300 Trials	0.118	0.134	0.134	0.142	0.166	0.152
	(0.305)	(0.375)	(0.370)	(0.334)	(0.261)	(0.277)
300-500 Trials	0.146	0.167	0.180	0.184	0.223*	0.252*
	(0.344)	(0.192)	(0.145)	(0.149)	(0.091)	(0.052)
500-700 Trials	0.155	0.192	0.208	0.201	0.246	0.266
	(0.304)	(0.236)	(0.188)	(0.208)	(0.135)	(0.104)
Memory		0.200	0.143	0.131	0.075	0.141
		(0.193)	(0.433)	(0.463)	(0.719)	(0.507)
Motor		-0.242	-0.183	-0.179	-0.162	-0.161
		(0.157)	(0.294)	(0.320)	(0.375)	(0.390)
Perception		0.224*	0.245*	0.227*	0.201	0.214
		(0.070)	(0.058)	(0.088)	(0.209)	(0.188)
Staircase			-0.134	-0.139	-0.168	-0.169
			(0.256)	(0.210)	(0.170)	(0.163)
Binary scale				0.288	0.328	0.336
				(0.196)	(0.232)	(0.236)
Discrete scale				0.006	0.032	0.058
				(0.967)	(0.865)	(0.754)
Objective scale					0.100	0.064
					(0.524)	(0.693)
Feedback						0.154
						(0.384)
Constant	0.921***	0.764***	0.830***	0.825***	0.793***	0.739***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.004)
Observations	103	103	103	103	100	95

Note. Estimated coefficients (and p-values in parenthesis, computed using robust standard errors) from Gamma Generalized Linear Models of the relation between design characteristics and the estimated value of the convergence parameter $\boldsymbol{\lambda}_{\text{CV}}^{Acc}$. The reference study features more than 700 trials of a task belonging to a mixed domain, measures confidence on a subjective and continuous scale and does not not implement either a staircase procedure or feedback. Changes in the number of observations across columns are due to missing data in the variables that are added, see Section 2 above. *Significance levels*: *** 10%, ** 5%, *** 1%.

Table C: Gamma regressions of $\boldsymbol{\lambda}_{\text{RS}}^{Conf}$ on design characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
# of participants	0.035	0.048	0.054	0.042	0.023	0.110
	(0.565)	(0.463)	(0.466)	(0.509)	(0.693)	(0.132)
150-300 Trials	-35.189*	-28.028	-24.129	-27.925	-23.557	-16.872
	(0.076)	(0.199)	(0.252)	(0.190)	(0.320)	(0.490)
300-500 Trials	-48.437**	-43.699**	-42.679**	-47.278***	-43.137**	-32.998
	(0.018)	(0.026)	(0.022)	(0.009)	(0.039)	(0.133)
500-700 Trials	-32.123	-30.740	-30.278	-32.769*	-28.689	-18.064
	(0.131)	(0.126)	(0.127)	(0.089)	(0.190)	(0.434)
Memory		-57.581**	-47.903*	-42.638	-47.592	-39.483
		(0.043)	(0.090)	(0.119)	(0.106)	(0.140)
Motor		42.882	28.390	19.761	24.552	13.151
		(0.456)	(0.623)	(0.730)	(0.671)	(0.806)
Perception		-49.002*	-49.322*	-47.065*	-47.872*	-41.892*
		(0.071)	(0.064)	(0.073)	(0.081)	(0.093)
Staircase			28.540***	34.921***	34.190***	43.097***
			(0.009)	(0.002)	(0.004)	(0.002)
Binary scale				-47.308***	-41.361***	-49.405***
				(0.000)	(0.007)	(0.005)
Discrete scale				-12.383	-6.716	-5.711
				(0.201)	(0.601)	(0.641)
Objective scale					14.526	6.785
					(0.147)	(0.481)
Feedback						58.690**
						(0.024)
Constant	90.195***	131.318***	115.716***	126.820***	117.198***	97.597***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Observations	103	103	103	103	100	95

Note. Estimated coefficients (and p-values in parenthesis, computed using robust standard errors) from Gamma Generalized Linear Models of the relation between design characteristics and the estimated value of the convergence parameter $\boldsymbol{\lambda}_{\text{RS}}^{Conf}$. The reference study features more than 700 trials of a task belonging to a mixed domain, measures confidence on a subjective and continuous scale and does not not implement either a staircase procedure or feedback. Changes in the number of observations across columns are due to missing data in the variables that are added, see Section 2 above. *Significance levels*: *** 10%, ** 5%, *** 1%.

Table D: Gamma regressions of $\boldsymbol{\lambda}_{\text{RS}}^{Acc}$ on design characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
# of participants	-0.010***	-0.007*	-0.007	-0.006	-0.007	-0.005
	(0.004)	(0.067)	(0.106)	(0.137)	(0.123)	(0.397)
150-300 Trials	1.666	1.135	1.207	1.626	1.320	1.962
	(0.288)	(0.543)	(0.524)	(0.363)	(0.466)	(0.332)
300-500 Trials	2.126	2.277	2.185	2.526	2.290	2.439
	(0.289)	(0.251)	(0.270)	(0.145)	(0.232)	(0.286)
500-700 Trials	-0.305	-0.342	-0.398	-0.438	-0.634	-0.483
	(0.844)	(0.820)	(0.787)	(0.727)	(0.637)	(0.753)
Memory		-2.830	-2.242	-2.759	-2.610	-1.748
		(0.252)	(0.380)	(0.279)	(0.309)	(0.503)
Motor		0.320	-0.400	0.091	0.276	-0.469
		(0.906)	(0.882)	(0.974)	(0.921)	(0.868)
Perception		-4.145*	-4.323*	-4.650*	-4.420*	-3.880
		(0.096)	(0.078)	(0.060)	(0.075)	(0.114)
Staircase			1.475	1.068	1.120	1.874
			(0.222)	(0.330)	(0.348)	(0.142)
Binary scale				24.453***	24.307***	23.344***
				(0.000)	(0.000)	(0.000)
Discrete scale				0.702	0.826	0.394
				(0.547)	(0.506)	(0.750)
Objective scale					0.197	-0.340
					(0.848)	(0.760)
Feedback						2.005
						(0.526)
Constant	10.159***	13.574***	12.852***	12.274***	12.234***	11.638***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	103	103	103	103	100	95

Note. Estimated coefficients (and p-values in parenthesis, computed using robust standard errors) from Gamma Generalized Linear Models of the relation between design characteristics and the estimated value of the convergence parameter $\boldsymbol{\lambda}_{\text{RS}}^{Acc}$. The reference study features more than 700 trials of a task belonging to a mixed domain, measures confidence on a subjective and continuous scale and does not not implement either a staircase procedure or feedback. Changes in the number of observations across columns are due to missing data in the variables that are added, see Section 2 above. *Significance levels*: *** 10%, ** 5%, *** 1%.