nature human behaviour

RESOURCE

https://doi.org/10.1038/s41562-019-0813-1

The Confidence Database

Dobromir Rahnev[1*], Kobe Desender[2,3], Alan L. F. Lee[4], William T. Adler[5], David Aguilar-Lleyda[6], Başak Akdoğan[7], Polina Arbuzova[8,9,10], Lauren Y. Atlas[11,12,13], Fuat Balcı[14], Ji Won Bang[15], Indrit Bègue[16], Damian P. Birney[17], Timothy F. Brady[18], Joshua Calder-Travis[19], Andrey Chetverikov[20], Torin K. Clark[21], Karen Davranche[22], Rachel N. Denison[23], Troy C. Dildine[11,24], Kit S. Double[25], Yalçın A. Duyan[14], Nathan Faivre[26], Kaitlyn Fallow[27], Elisa Filevich[8,9,10], Thibault Gajdos[22], Regan M. Gallagher[28,29,30], Vincent de Gardelle[31], Sabina Gherman[32,33], Nadia Haddara[1], Marine Hainguerlot[34], Tzu-Yu Hsu[35], Xiao Hu[36], Iñaki Iturrate[37], Matt Jaquiery[19], Justin Kantner[38], Marcin Koculak[39], Mahiko Konishi[40], Christina Koß[8,10], Peter D. Kvam[41], Sze Chai Kwok[42,43,44], Maël Lebreton[45], Karolina M. Lempert[46], Chien Ming Lo[35,47], Liang Luo[36], Brian Maniscalco[48], Antonio Martin[35], Sébastien Massoni[49], Julian Matthews[30,50], Audrey Mazancieux[26], Daniel M. Merfeld[51], Denis O'Hora[52], Eleanor R. Palser[53,54,55], Borysław Paulewicz[56], Michael Pereira[57], Caroline Peters[8,9,10], Marios G. Philiastides[32], Gerit Pfuhl[58], Fernanda Prieto[59], Manuel Rausch[60], Samuel Recht[61], Gabriel Reyes[59], Marion Rouault[62], Jérôme Sackur[62,63], Saeedeh Sadeghi[64], Jason Samaha[65], Tricia X. F. Seow[66], Medha Shekhar[1], Maxine T. Sherman[67,68], Marta Siedlecka[39], Zuzanna Skóra[39], Chen Song[69], David Soto[70,71], Sai Sun[72], Jeroen J. A. van Boxtel[30,73], Shuo Wang[74], Christoph T. Weidemann[75], Gabriel Weindel[22], Michał Wierzchoń[39], Xinming Xu[42], Qun Ye[42], Jiwon Yeon[1], Futing Zou[42] and Ariel Zylberberg[76]

Understanding how people rate their confidence is critical for the characterization of a wide range of perceptual, memory, motor and cognitive processes. To enable the continued exploration of these processes, we created a large database of confidence studies spanning a broad set of paradigms, participant populations and fields of study. The data from each study are structured in a common, easy-to-use format that can be easily imported and analysed using multiple software packages. Each dataset is accompanied by an explanation regarding the nature of the collected data. At the time of publication, the Confidence Database (which is available at https://osf.io/s46pr/) contained 145 datasets with data from more than 8,700 participants and almost 4 million trials. The database will remain open for new submissions indefinitely and is expected to continue to grow. Here we show the usefulness of this large collection of datasets in four different analyses that provide precise estimations of several foundational confidence-related effects.

Researchers from a wide range of fields use ratings of confidence to provide fundamental insights about the mind. Confidence ratings are subjective ratings regarding one's first-order task performance. For example, participants may first decide whether a probe stimulus belongs to a previously learned study list or not. In this case, a confidence rating could involve the second-order judgement of the participants regarding how sure they are about the accuracy of the decision made in that trial (that is, the accuracy of the first-order task performance). Such second-order judgements reflect the ability of people to introspect and can be dissociated from the first-order judgement[1]. Confidence ratings tend to correlate strongly with accuracy, response speed and brain activity distinguishing old and new probes[2], suggesting that they reflect relevant internal states.

The question of how humans (or other animals) evaluate their own decisions has always been an important topic in psychology, and the use of confidence ratings dates back to the early days of experimental psychology[3]. Among many other things, confidence has been used as a tool to determine the number of distinct memory retrieval processes[4], reveal distortions of visual awareness[5], understand the factors that guide learning[6], assess the reliability of eyewitness testimony[7], test theories of sensory processing[8] and decision-making[9,10], help estimate the fit of parameters of the psychometric function more efficiently[11] and characterize various psychiatric conditions[12]. The wide application of confidence makes it a fundamental measure in psychological research.

However, despite the widespread use of confidence ratings, scientific progress has been slowed by the traditional unavailability of

A full list of affiliations appears at the end of the paper.

previously collected data. In the current system, testing a new idea often requires scientists to spend months or years gathering the relevant data. The substantial cost, in time and money, associated with the collection of new data has undoubtedly led to many new ideas being abandoned without ever being examined empirically. This is especially unfortunate given that these ideas could probably have been tested using the dozens of datasets that have previously been collected by other scientists.

When data re-use takes place, it is typically within a laboratory or a small scientific group that often restrict themselves to very specific paradigms; this potentially limits the formation of a broader understanding of confidence across a wider range of tasks and participants. Therefore, another important advantage of data re-use lies in the diversity of experimental tasks, set-ups and participants offered by compiling datasets from different labs and different populations.

Although data sharing can accelerate scientific progress considerably, fields devoted to understanding human behaviour unfortunately have cultures of not sharing data[13,14]. For example, Wicherts et al.[15] documented their painstaking and ultimately unsuccessful endeavour to obtain behavioural data for re-analysis; despite persistent efforts, Wicherts et al. were able to obtain only 25.7% of datasets that the original authors had claimed were available for re-analysis. Nevertheless, recent efforts to increase openness have started to shift the culture considerably, and more and more authors post their data in online depositories[16,17].

There are, however, several challenges involved in secondary analyses of data, even when such data have been made freely available. First, the file type may not be usable or clear for some researchers. For example, sharing files in proprietary formats may limit the ability of other researchers to access them (for example, if reading the file requires software that is not freely or easily obtainable). Second, even if the data can be readily imported and used, important information about the data may not have been included. Third, researchers who need data from a large number of studies have to spend a considerable amount of time finding individual datasets, familiarizing themselves with how each dataset is structured and organizing of the all datasets into a common format for analysis. Finally, given the size of the literature, it can be difficult to determine which papers contain relevant data.

Here we report on a large-scale effort to create a database of confidence studies that addresses all of the problems described above. The database uses an open standardized format (.csv) that can be easily imported into any software program used for analysis. The individual datasets are formatted using the same general set of guidelines making it less likely that critical components of the datasets are not included and ensuring that data re-use is much less time consuming. Finally, creating a single collection of confidence datasets makes it much easier and faster to find datasets that could be re-used to test new ideas or models.

Details of the database

The Confidence Database is hosted on the Open Science Framework (OSF) website (https://osf.io/s46pr/). Each dataset is represented by two files—a data file in .csv format and a readme file in .txt format.

The majority of data files contain the following fields: participant index, stimulus, response, confidence, response time (RT) of the decision and RT of the confidence rating. Depending on the specific design of each study, these fields can be slightly different (for example, if there are two stimuli on each trial, or confidence and decision are given with a single button press). Furthermore, many datasets include additional fields that are required to fully describe the nature of the collected data.

The readme files contain essential information about the contributor, corresponding published paper (if the dataset is published and current status of the project if not), stimuli used, confidence scale and experimental manipulations. Other information, including the original purpose of the study, the main findings and the location of data collection, are also often included. In general, the readme files provide a quick reference regarding the nature of each dataset and describe details that could be needed for future re-analyses.

The Confidence Database includes a wide variety of studies. Individual datasets recruit different populations (such as healthy or patient populations), focus on different fields of study (such as perception, memory, motor control and decision making), use different confidence scales (such as binary, *n*-point scales, continuous scales and wagering), employ different types of tasks (such as binary judgements versus continuous estimation tasks) and collect confidence at different times (for example, after or simultaneous with the decision). Figure 1 provides a broad overview of the types of datasets that are included in the Confidence Database at the time of publication. This variety ensures that future re-analyses can address a large number of scientific questions and test them on the basis of multiple methods of evaluating one's own primary-task performance.

Importantly, the Confidence Database will remain open for new submissions indefinitely. Instructions for new submissions are provided on the OSF page of the database. Carefully formatted .csv and .txt files that follow the submission instructions can be e-mailed to the Confidence Database (confidence.database@gmail.com). They will be checked for quality and then uploaded with the rest of the database.

Finally, to facilitate searching the database, a spreadsheet with basic information regarding each study will be maintained (a link to this can be found on the OSF page). The spreadsheet includes information about a number of different details regarding each dataset, such as the field of study (for example, perception or memory), authors, corresponding publication, number of participants and trials, and the type of confidence scale.

At the time of publication, the Confidence Database contained 145 datasets, bringing together 8,787 participants and a total of 3,955,802 individual trials. The data were collected mostly in laboratory experiments (from 18 different countries over five continents) but also in online experiments. Despite its already large size, the database still contains only a small fraction of the available data on confidence and is expected to continue to grow. We encourage researchers who already make their data available to also submit their data to the Confidence Database. This would make their data easier to discover and re-use, and would multiply the impact of their research.

Anyone is encouraged to download and re-use the data from the database. The database is shared under the most permissive CC0 license and, therefore, places the data in the public domain. As with the re-use of any other data, publications that result from such re-analysis should cite this paper, as well as the listed citation for each of the datasets that were re-analysed. We strongly encourage the preregistration of future secondary analyses and refer readers who wish to perform such analyses to an excellent discussion of this process, including preregistration templates, by Weston et al.[18] (the templates are available at https://osf.io/x4gzt).

Example uses of the Confidence Database

The Confidence Database can be used for a variety of purposes, such as developing and testing new models of confidence generation; comparing confidence across different cognitive domains, rating scales and populations; determining the nature of metacognitive deficits that accompany psychiatric disorders; characterizing the relationship between confidence, accuracy and RTs; and building theories of the RTs associated with confidence ratings. Furthermore, the database can also be used to test hypotheses that are unrelated to confidence due to the inclusion of choice, accuracy and RT. Different studies can re-use a few relevant datasets (or a single dataset) or simultaneously analyse a large set of the available datasets and can therefore achieve substantially higher power than typical individual studies.

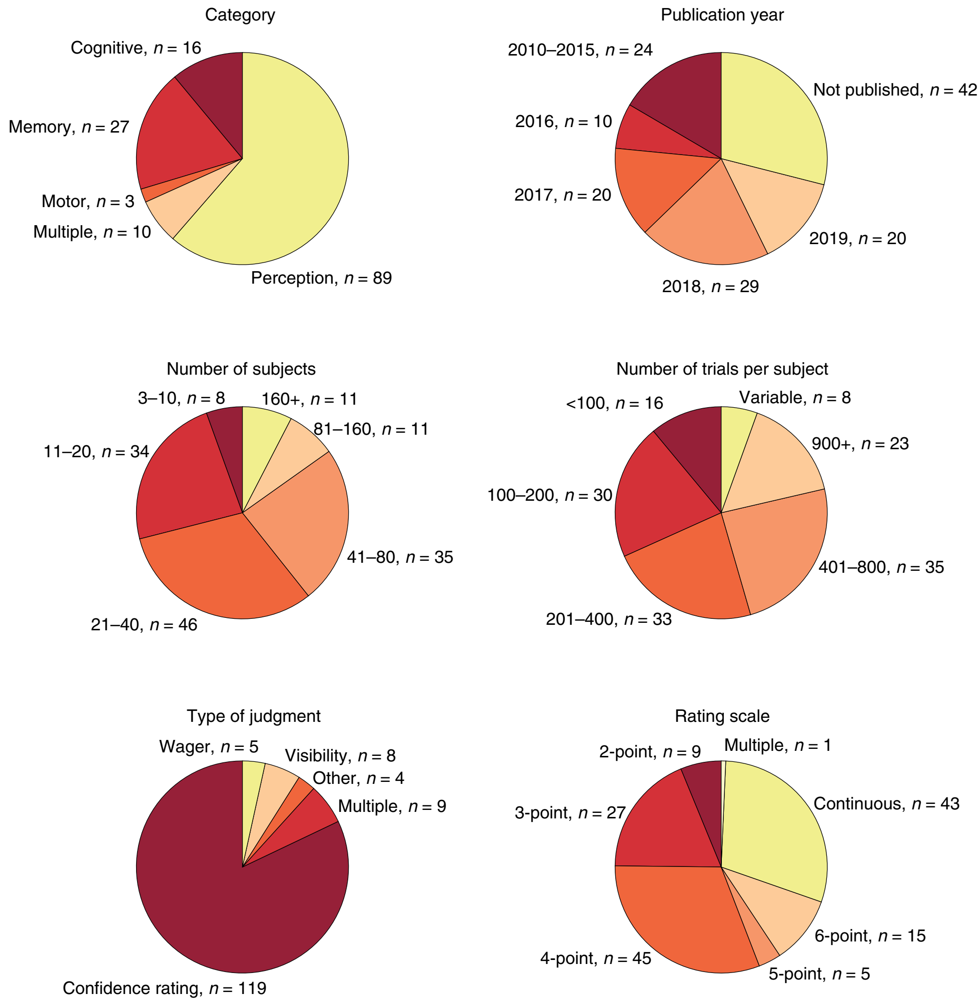

Fig. 1 | Datasets in the Confidence Database at the time of publication. The number of datasets split by category, publication year, number of participants, number of trials per participant, type of judgement and rating scale. 'Multiple' in the top-left chart indicates that the same participants completed tasks from more than one category. The maximum number of participants was 589 and the maximum trials per participant was 4,320. 'Variable' in the middle chart on the right indicates that different participants completed different numbers of trials.

The results of four different example analyses that demonstrate the potential utility and versatility of the database are shown below. These analyses were designed to take advantage of a large proportion of the available data, resulting in very large sample sizes. Annotated codes for running these analyses are freely available at the OSF page of the database (https://osf.io/s46pr/). We note that these codes can be used by researchers as a starting point for future analyses. All statistical tests are two-tailed and their assumptions were verified. Measurements were taken from distinct samples.

How confidence is related to choice and confidence RTs. One of the best-known properties of confidence ratings is that they correlate negatively with choice RTs[2]. However, despite its importance, this finding is virtually always treated as the outcome of a binary null-hypothesis significance test, which does not reveal the strength of the effect. At the same time, it is becoming widely recognized that building replicable quantitative science requires that researchers, among other things, "adopt estimation thinking and avoid dichotomous thinking"[19]. However, precise estimation requires very large sample sizes and any individual study is usually not large enough to allow for accuracy in estimation. The Confidence Database provides a unique opportunity to estimate, with great precision, the strength of foundational effects such as the negative correlation between confidence and choice RT and, therefore, to inform theories that rely on these effects. The database also enables investigations of lesser-studied relationships, such as the relationship between confidence and confidence RT.

Using the data from the Confidence Database, we therefore investigated the precise strength of the correlation of confidence with both choice and confidence RT. We first selected all of the datasets in which choice and confidence RTs were reported. Note that some datasets featured designs whereby the choice and confidence were made through a single button press—such datasets were excluded from these analyses. Furthermore, we excluded individual participants who used only a single level of confidence, because it is impossible to correlate confidence and RT for such individuals, and participants for whom more than 90% of the data were excluded (which occurred for six participants from a study with very high confidence RTs; see below). In total, the final analyses were based on 4,089 participants from 76 different datasets.

Before conducting the main analyses, we performed basic data clean-up. This step is important as contributors are encouraged to include all of the participants and trials from an experiment even if some of the participants or trials were excluded from data analyses in the original publications. Specifically, we excluded all of the trials without a confidence rating (such trials typically came from studies that included a deadline for the confidence response), all of the trials without choice RT (typically due to a deadline on the main decision) and all of the trials with confidence and/or choice RTs slower than 5 s (the results remained very similar if a threshold of 3 s or 10 s was used instead). These exclusion criteria resulted in the removal of 7.3% of the data. Moreover, for each participant, we excluded all choice and confidence RTs that differed by more than 3 s.d. from the mean (resulting in the removal of an additional 1.8% of the data).

For each participant, we then correlated the confidence ratings with choice RTs. We found that the average correlation across participants was Pearson's $r = -0.24$ ($t_{4,088} = -71.09$, $P < 2.2 \times 10^{-16}$, Cohen's $d = 1.11$). The very large sample size enabled us to estimate the average correlation with a very high degree of precision—the 99.9% confidence interval (CI) for the average correlation value was −0.25 to −0.23, which should be considered to be a medium-to-large effect[20]. At the same time, it is important to emphasize that the high precision in estimating the average correlation does not imply a lack of variability between individual participants. Indeed, we observed very high individual variability (s.d. = 0.21), which we visualized by plotting all of the individual correlation values and corresponding density functions in the form of raincloud plots[21] (Fig. 2a). However, the effect size is large enough that power analyses indicate that a sample size as small as $n = 9$ provides greater than 80% power and a sample size of $n = 13$ provides greater than 95% power to detect this effect (at $\alpha = 0.05$).

We next performed the same analyses for the correlation between confidence and confidence RT. We found that the average correlation across participants was $r = -0.07$, s.d. = 0.24 ($t_{4,088} = -18.77$, $P < 2.2 \times 10^{-16}$, $d = 0.29$) with a 99% CI for the average correlation value of −0.08 to −0.06. This effect should be considered to be "very small for the explanation of single events but potentially consequential in the not-very-long run"[20]. The small but reliable negative association between confidence and confidence RT would have been particularly difficult to detect with a small sample size. Indeed, a study with a sample size of 33 (the median sample size of the studies in the Confidence Database) would have only 37% power of detecting this effect. To achieve power of 80%, a sample size of $n = 93$ is required; for power of 95%, $n = 152$ is needed.

Note that existing models of confidence generation[22] predict a lack of any association between confidence and confidence RT (but see ref. [23]). The small but reliable negative correlation therefore raises the question of what causes this negative association. One possibility is that participants are faster to give high confidence ratings because a strong decision-related signal can propagate faster to neural circuits that generate the confidence response (in the case of attention, a similar argument was described previously[24]) but further research is needed to directly test this hypothesis.

Finally, we also found that the strength of the correlation between confidence and confidence RT was itself correlated with the strength of the correlation between confidence and choice RT ($r_{4,087} = 0.20$, $P < 2.2 \times 10^{-16}$, 99% CI = 0.16–0.24; Fig. 2b). Future research should investigate whether this correlation is due to variability in individual participants or variability at the level of the datasets.

Serial dependence in confidence RT. It is well known that perceptual choices[25], confidence judgements[26] and choice RTs[27] are subject to serial dependence. Such findings have been used to make fundamental claims about the nature of perceptual processing such as that the visual system forms a 'continuity field' through space and time[28,29]. The presence of serial dependence can therefore help to reveal the underlying mechanisms of perception and cognition. However, to the best of our knowledge, the presence of serial dependence has never been investigated for one of the most important components of confidence generation—confidence RT. Determining whether serial dependence exists for confidence RT and, if so, estimating its effect size precisely can therefore provide important insights about the nature of confidence generation.

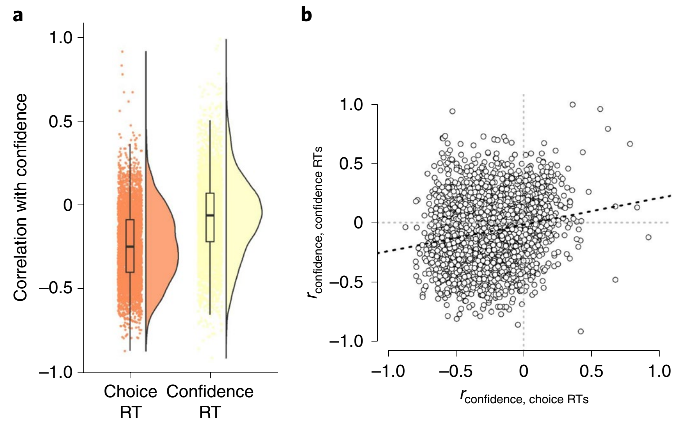

Fig. 2 | Correlating confidence with choice and confidence RT. a, We found a medium-to-large negative correlation ($r = -0.24$, $P < 2.2 \times 10^{-16}$, $n = 4{,}089$) between confidence and choice RT, as well as a small negative correlation ($r = -0.07$, $P < 2.2 \times 10^{-16}$, $n = 4{,}089$) between confidence and confidence RT. The boxes show the median and the interquartile (25–75%) range, and the whiskers show the 2–98% range. **b**, The strength of the two correlations in **a** were themselves correlated across individuals ($r = 0.20$, $P < 2.2 \times 10^{-16}$, $n = 4{,}089$).

To address this question, we considered the data from the Confidence Database. We analysed all of the datasets in which confidence was provided with a separate button press from the primary decision and that reported confidence RT. In total, 82 datasets were included, comprising 4,474 participants. Data clean-up was performed as described for the analysis presented above. Specifically, we removed all of the trials without confidence RT and all of the trials with confidence RT slower than 5 s (results remained very similar if a threshold of 3 s or 10 s was used instead), both on the current trial and up to seven trials back, because we wanted to investigate serial dependence up to lag-7 (this excluded a total of 4.3% of the data). Furthermore, we also excluded, separately for each participant, all confidence RTs that differed by more than 3 s.d. from the mean (excluding an additional 9.6% of the data).

We performed a mixed regression analysis predicting confidence RTs with fixed effects for the recent trial history up to seven trials back[25] and random intercepts for each participant. Degrees of freedom were estimated using Satterthwaite's approximation, implemented using the lmerTest package[30]. We found evidence for strong autocorrelation in confidence RT. Specifically, there was a large lag-1 autocorrelation ($\beta = 1.346$, $t_{1,299,601} = 153.6$, $P < 2.2 \times 10^{-16}$, $d = 0.27$; Fig. 3). The strength of the autocorrelation dropped sharply for higher lags but remained significantly positive until at least lag-7 (all P values $< 2.2 \times 10^{-16}$).

These results suggest the existence of serial dependence in confidence RT. However, it remains unclear whether previous trials have a causal effect on the current trial. For example, some of the observed autocorrelation may be due to a general decrease in confidence RTs over the course of each experiment. To address this question, future studies should experimentally manipulate the speed of the confidence ratings on some trials and explore whether such manipulations affect the confidence RT during subsequent trials.

Negative metacognitive sensitivity. Many studies have shown that humans and other animals have the metacognitive ability to use confidence ratings to judge the accuracy of their own decisions[31]. In other words, humans have positive metacognitive sensitivity[32], meaning that higher levels of confidence predict better performance. However, it is not uncommon that individual participants

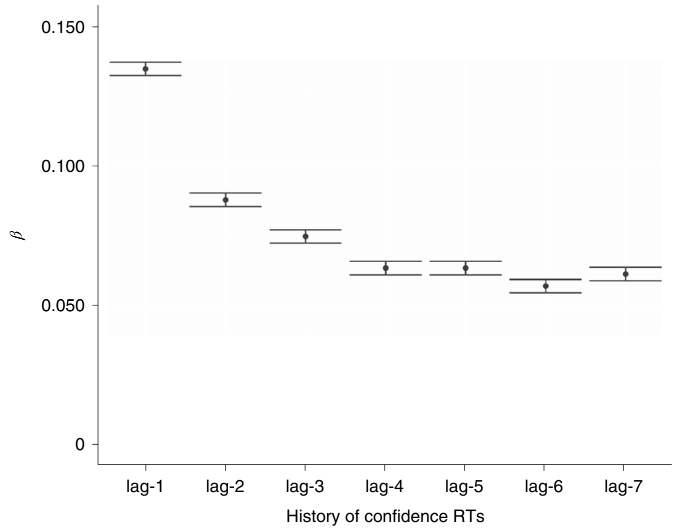

Fig. 3 | Serial dependence in confidence RT. We observed a large lag-1 autocorrelation ($\beta = 1.346$, $t_{1,299,601} = 153.6$, $P < 2.2 \times 10^{-16}$, $n = 4,474$). The autocorrelation decreased for higher lags but remained significant up to lag-7 (all P values $< 2.2 \times 10^{-16}$, $n = 4,474$). Data are mean ± s.e.m. Individual datapoints are not shown because the plots are based on the results of a mixed-model analysis.

fail to show the typically observed positive metacognitive sensitivity. To date, such cases have been difficult to investigate because they occur infrequently within a given dataset.

Using the Confidence Database, we estimated the prevalence of negative metacognitive sensitivity and investigated its causes. We analysed all of the datasets that contained the variables confidence and accuracy. In total, 71 datasets were included, comprising of 4,768 participants. We excluded studies on subjective difficulty, because these investigate the relationship between confidence and performance within correct trials. We further excluded participants who reported only a single level of confidence (as it is impossible to estimate metacognitive sensitivity for such participants), studies with a continuous measure of accuracy and participants for whom more than 90% of the data were excluded (which occurred for six participants from a study with very high confidence RTs). Metacognitive sensitivity was computed using a logistic regression that predicted accuracy using normalized confidence ratings. This measure of metacognition has a number of undesirable properties[32], but reliably indicates whether metacognitive sensitivity is positive or negative.

We found that, across all of the participants, the average β value from the logistic regression was 0.096, s.d. = 0.064, ($t_{4,767} = 104.01$, $P < 2.2 \times 10^{-16}$, $d = 1.5$; Fig. 4a), indicating that metacognitive sensitivity was reliably positive in the group. However, 293 of the participants (6.1% out of all of the participants) had a negative β value, indicating the potential presence of negative metacognitive sensitivity.

We next examined why such negative coefficients may occur for these 293 participants. We reasoned that the majority of the cases of estimated negative metacognitive sensitivity could be due to several factors that were unrelated to the true metacognitive sensitivity of each participant. First, the negative β values could simply be due to misestimation due to relatively small sample sizes. Even though the number of trials per participant did not correlate with the β coefficient of participants ($r_{4,766} = -0.021$, $P = 0.143$, 99% CI = −0.25 to −0.17; Fig. 4b), 9.9% of all participants with negative β values completed less than 50 trials in total. Second, a positive relationship between confidence and accuracy can be expected only if performance is above chance (if performance is at chance level, this may indicate that there is no reliable signal that could be used by the metacognitive system, although some previous studies have suggested that positive metacognition may be present even in such cases[33,34]). We did indeed observe a correlation between the β values and average accuracy ($r_{4,766} = 0.203$, $P < 2.2 \times 10^{-16}$, 99% CI = 0.17–0.24; Fig. 4c) with 19.4% of all participants with negative β values having an accuracy of less than 55%. Third, for the datasets that included choice RT or confidence RT, we calculated the overall median choice/confidence RTs and correlated these with the β coefficients (one dataset was excluded here because the primary task was to complete Raven's progressive matrices and, therefore, choice and confidence RTs were in the range of minutes rather than seconds). Again, we observed significant correlations between β values and choice RTs ($r_{3,076} = -0.083$, $P = 3.6 \times 10^{-6}$, 99% CI = −0.13 to −0.04; Fig. 4d) and between β values and confidence RTs ($r_{2,191} = 0.071$, $P = 0.0009$, 99% CI = 0.02–0.13; Fig. 4e), but the magnitude of these correlations was very small and only 2.3% and 2.4% of participants with negative β values had median choice or confidence RT of less than 200 ms, respectively. Finally, we reasoned that β coefficients could be misestimated if a very large proportion of confidence judgements were the same. We therefore computed the proportion of the most common confidence rating for each participant (mean = 37.9%, s.d. = 0.22). We did not observe a significant correlation between the proportion of the most common confidence rating and the β values ($r_{4,766} = -0.025$, $P = 0.086$, 99% CI = 0.05–0.12; Fig. 4f), and only 5.4% of all participants with negative β values used only a single confidence rating for more than 95% of the time.

Overall, 96 out of 293 participants with negative β values (32.7%) completed less than 50 trials, had an overall accuracy of less than 55% or used the same confidence response on more than 95% of all trials. This means that 197 participants had negative β values despite the absence of any of these factors (note that, for 55 of these participants, no RT information was provided and, therefore, a few of them could have had overly fast choice or confidence RTs). This result raises the question about the underlying causes of the negative β values. Follow-up studies could focus on these individuals and determine whether there is anything different about them or the tasks that they completed.

Confidence scales used in perception and memory studies. One of the strengths of the Confidence Database is that it enables investigations into how specific effects depend on factors that differ from study to study. For example, for any of the analyses described above, one could ask how the results depend on factors such as the domain of study (that is, perception, memory or cognitive), confidence scale used (for example, n-point versus continuous), whether confidence was provided simultaneously with the decision or the number of trials per participant. These questions can reveal some of the mechanisms behind confidence generation, such as whether metacognition is a domain-specific or domain-general process[35,36].

Here we took advantage of this feature of the Confidence Database to ask a metascience question: does the type of confidence scale researchers use depend on the subfield that they work in? Confidence ratings are typically given in one of two ways. The majority of studies use a discrete Likert scale (for example, a 4-point scale where 1 is lowest confidence and 4 is highest confidence). Such scales typically have a fixed stimulus–response mapping so that a given button always indicates the same level of confidence (although variable stimulus–response mappings are still possible). Likert scales can also have a different number of options. Comparatively fewer studies use continuous scales (for example, a 0–100 scale where 0 is lowest confidence and 100 is highest confidence). Such scales typically do not have a fixed stimulus–response mapping and

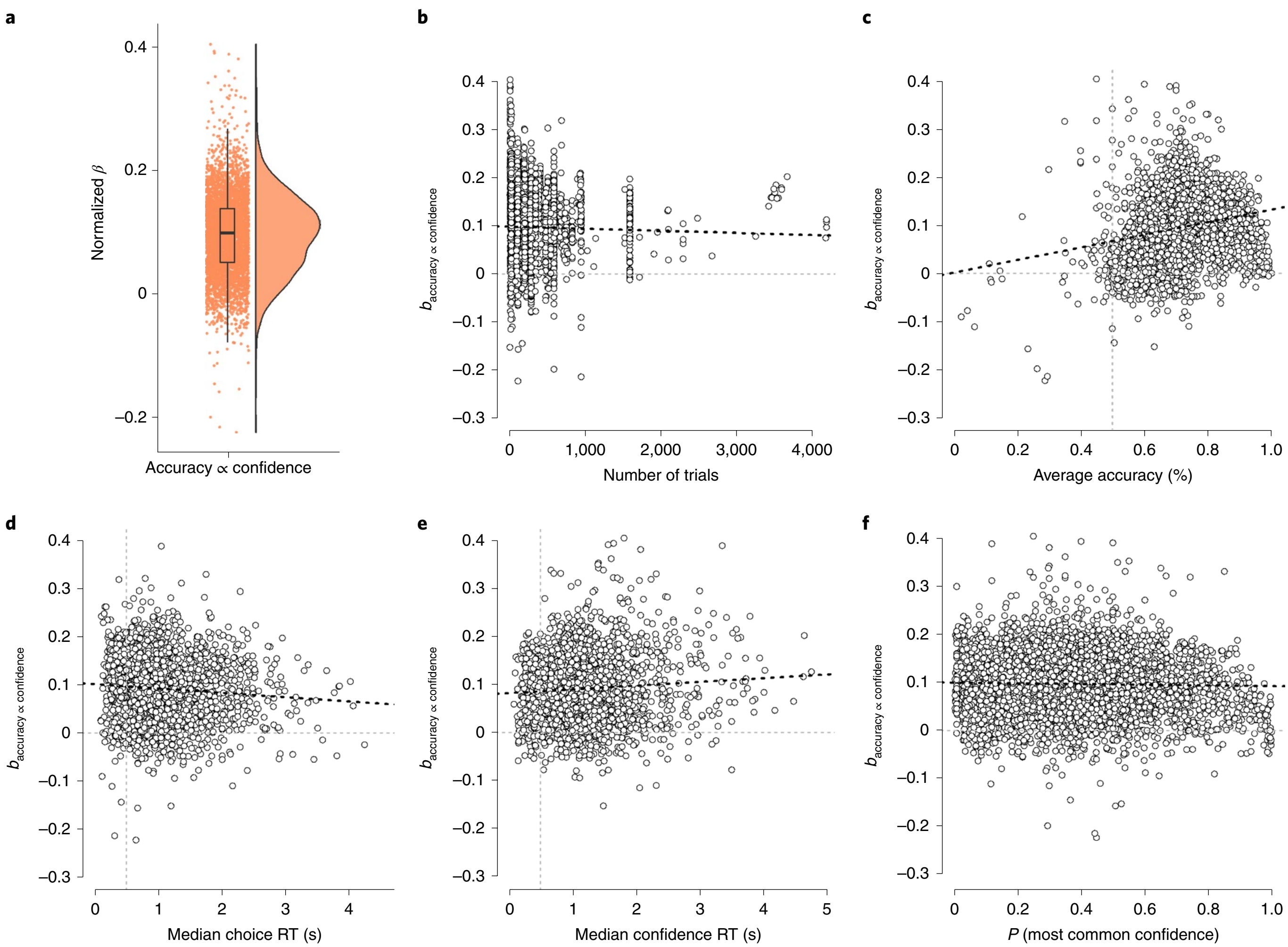

Fig. 4 | The prevalence of estimates of negative metacognitive sensitivity. a, Individual β values and β-value density plot for the observed relationship between confidence and accuracy. Box shows the median and the interquartile (25–75%) range and the whiskers show the 2–98% range. **b–f**, The relationships, including lines of best fit, between the β value for the confidence–accuracy relationship and the number of trials (**b**), average accuracy (**c**), median choice RT (**d**), median confidence RT (**e**) and the proportion of trials in which the most common confidence judgement was given (**f**).

responses are often given using a mouse click rather than a button press (although it is also possible to use a keyboard in such cases).

We focused on the domains of perception and memory because these were the only two domains with a sufficient number of datasets in the database (89 datasets for perception and 27 datasets for memory; all other domains had at most 16 datasets; Fig. 1). We categorized each dataset from these two domains as using a 2-point, 3-point, 4-point, 5-point, 6-point, 7-to-11-point or a continuous confidence scale (we combined the 7-point to 11-point scales into a single category owing to the low number of datasets with such scales). Finally, we computed the percentage of datasets with each of the confidence scales separately for the perception and memory domains.

We found that there were several systematic differences between the two domains. Notably, memory studies used a 3-point confidence scale 48% of the time (13 out of 27 datasets), whereas perception studies used a 3-point confidence scale only 16% of the time (14 out of 89 datasets) with the difference in proportions being significant ($Z = -3.49$, $P = 0.0005$; Fig. 5). On the other hand, a much lower percentage of memory datasets (4%, 1 out of 27 datasets) used a continuous scale compared with perception studies (33%, 29 out of 89 datasets; $Z = 3.002$, $P = 0.003$). Both comparisons remained significant at the 0.05 level after Bonferroni correction for multiple comparisons was applied. We did not find any difference between perception and memory studies for the rest of the confidence scale types (all P values > 0.2 before Bonferroni correction).

These results suggest that there are systematic differences in how confidence is collected in perception and memory studies with most pronounced differences in the use of 3-point and continuous scales. As it is unclear why perception and memory research would benefit from the use of different confidence scales, these findings may point to a lack of sufficient cross-talk between the two fields. Future research should first confirm the presence of such differences using an unbiased sample of published studies and then trace the origin of these differences.

Data sharing in the behavioural sciences

It is a sad reality that "most of the data generated by humanity's previous scientific endeavours is now irrecoverably lost"[13]. Data are lost due to outdated file formats; researchers changing universities, leaving academia or becoming deceased; websites becoming defunct; and a lack of interpretable metadata that describe the raw data. It is unlikely that much of the data not already uploaded to websites dedicated to data preservation will remain available for future research several decades from now.

We hope that the Confidence Database will contribute to substantially increased data preservation and serve as an example for

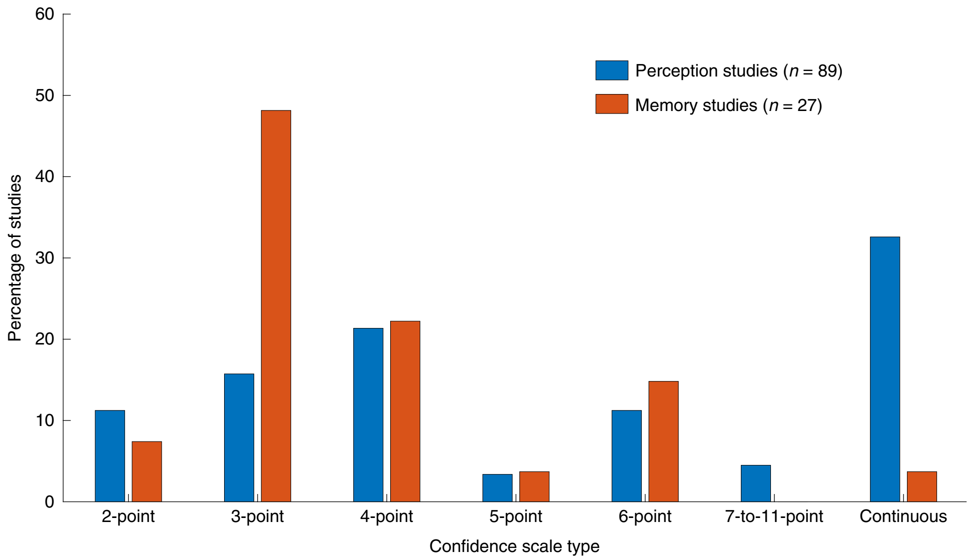

Fig. 5 | The use of confidence scales in perception and memory studies. The percentages of 2-point, 3-point, 4-point, 5-point, 6-point, 7-to-11-point and continuous confidence scales were plotted separately for the perception and memory datasets. We combined the 7-point to 11-point scales because of the low numbers of datasets with such scales. The two domains differed in how often they used 3-point and continuous scales.

similar databases in other subfields of behavioural science and beyond. Many subfields of psychology produce data that can be fully summarized in a single file using a common format and can therefore be easily shared. The mere existence of such a database in a given field may encourage data sharing by facilitating the process of preparing and uploading data; indeed, a lack of easy options for data sharing is among the important factors preventing researchers from sharing their data[37,38]. A popular database can also provide the benefit of the extra visibility afforded to the studies in it. Databases could serve as invaluable tools for meta-analyses and as a means to minimize false-positive rates that may originate from low-powered studies and publication bias (that is, favouring significant findings) by simply including datasets that also show null effects. Importantly, it is critical that sharing data is performed ethically and that participant anonymity is not compromised[39–41]. We have followed these principles in assembling the Confidence Database—all of the datasets have received IRB approvals by the relevant local committees (these can be found in the original publications), all of the participants have provided informed consent and all available data are deidentified.

Facilitation of data sharing would benefit from determining the factors that prevent researchers from exercising this important practice as part of their dissemination efforts. One of these factors could be the notion that researchers who spent resources to collect the original dataset should have priority over others in re-using their own data[37,42]. We argue that sharing data can have positive consequences for individual researchers by increasing the visibility of their research, the citation rate[43] and the accuracy of that research by enabling meta-analysis. Another set of factors are those that deter researchers from using shared data in open repositories. One of those factors is the belief that utilizing shared data could limit the impact of the work. Milham et al.[44] addressed such issues by demonstrating that manuscripts using shared data can, in fact, result in impactful papers in cognitive neuroscience and made a case for a more universal effort for data sharing. We hope that the construction and maintenance of the Confidence Database will help to address some of these issues in the domain of confidence research.

Finally, it is important to consider the limitations of the Confidence Database and similar future databases. First, the quality of such databases is determined by the quality of the individual studies; amassing large quantities of unreliable data would be of little use. Second, the datasets included are unlikely to be an unbiased sample of the literature (although the literature as a whole is unlikely to be an unbiased sample of all possible studies). Third, in standardizing the data format across various datasets, some of the richness of each dataset is lost. Thus, in addition to contributing to field-wide databases, we encourage researchers to also share their raw data in a separate repository.

Conclusion

The traditional unavailability of data in the behavioural sciences is beginning to change. An increasing number of funding agencies now require data sharing and individual researchers often post their data even in the absence of official mandates to do so. The Confidence Database represents a large-scale attempt to create a common database in a subfield of behavioural research. We believe that this effort will have a large and immediate effect on confidence research and will become the blueprint for many other field-specific databases.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The Confidence Database is available at https://osf.io/s46pr/.

Code availability

Codes reproducing all analyses in this paper are available at https://osf.io/s46pr/.

Received: 6 August 2019; Accepted: 11 December 2019;
Published online: 3 February 2020

References

1. Mamassian, P. Visual confidence. *Annu. Rev. Vis. Sci.* **2**, 459–481 (2016).
2. Weidemann, C. T. & Kahana, M. J. Assessing recognition memory using confidence ratings and response times. *R. Soc. Open Sci.* **3**, 150670 (2016).
3. Peirce, C. S. & Jastrow, J. On small differences in sensation. *Mem. Natl Acad. Sci.* **3**, 75–83 (1884).
4. Ratcliff, R., Van Zandt, T. & McKoon, G. Process dissociation, single-process theories, and recognition memory. *J. Exp. Psychol. Gen.* **124**, 352–374 (1995).

5. Azzopardi, P. & Cowey, A. Is blindsight like normal, near-threshold vision? *Proc. Natl Acad. Sci. USA* **94**, 14190–14194 (1997).
6. Robey, A. M., Dougherty, M. R. & Buttaccio, D. R. Making retrospective confidence judgments improves learners' ability to decide what *not* to study. *Psychol. Sci.* **28**, 1683–1693 (2017).
7. Wixted, J. T. & Wells, G. L. The relationship between eyewitness confidence and identification accuracy: a new synthesis. *Psychol. Sci. Publ. Int.* **18**, 10–65 (2017).
8. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*. (John Wiley & Sons Ltd, 1966).
9. Mueller, S. T. & Weidemann, C. T. Decision noise: an explanation for observed violations of signal detection theory. *Psychon. Bull. Rev.* **15**, 465–494 (2008).
10. Balakrishnan, J. D. & Ratcliff, R. Testing models of decision making using confidence ratings in classification. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 615–633 (1996).
11. Yi, Y. & Merfeld, D. M. A quantitative confidence signal detection model: 1. Fitting psychometric functions. *J. Neurophysiol.* **115**, 1932–1945 (2016).
12. David, A. S., Bedford, N., Wiffen, B. & Gilleen, J. Failures of metacognition and lack of insight in neuropsychiatric disorders. *Proc. R. Soc. B* **367**, 1379–1390 (2012).
13. Hardwicke, T. E. & Ioannidis, J. P. A. Populating the data ark: an attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLoS One* **13**, e0201856 (2018).
14. Vines, T. H. et al. The Availability of research data declines rapidly with article age. *Curr. Biol.* **24**, 94–97 (2014).
15. Wicherts, J. M., Borsboom, D., Kats, J. & Molenaar, D. The poor availability of psychological research data for reanalysis. *Am. Psychol.* **61**, 726–728 (2006).
16. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
17. Nelson, L. D., Simmons, J. & Simonsohn, U. Psychology's renaissance. *Annu. Rev. Psychol.* **69**, 511–534 (2018).
18. Weston, S. J., Ritchie, S. J., Rohrer, J. M. & Przybylski, A. K. Recommendations for increasing the transparency of analysis of preexisting data sets. *Adv. Methods Pract. Psychol. Sci*. **2**, 214–227 (2019).
19. Cumming, G. The new statistics: why and how. *Psychol. Sci.* **25**, 7–29 (2014).
20. Funder, D. C. & Ozer, D. J. Evaluating effect size in psychological research: sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* **2**, 156–168 (2019).
21. Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R. & Kievit, R. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res.* **4**, 63 (2019).
22. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864–901 (2010).
23. Moran, R., Teodorescu, A. R. & Usher, M. Post choice information integration as a causal determinant of confidence: novel data and a computational account. *Cogn. Psychol.* **78**, 99–147 (2015).
24. Nikolov, S., Rahnev, D. & Lau, H. Probabilistic model of onset detection explains paradoxes in human time perception. *Front. Psychol.* **1**, 37 (2010).
25. Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat. Commun.* **8**, 14637 (2017).
26. Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M. & Lau, H. Confidence leak in perceptual decision making. *Psychol. Sci.* **26**, 1664–1680 (2015).
27. Laming, D. Autocorrelation of choice-reaction times. *Acta Psychol.* **43**, 381–412 (1979).
28. Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nat. Neurosci.* **17**, 738–743 (2014).
29. Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K. & Whitney, D. Serial dependence in position occurs at the time of perception. *Psychon. Bull. Rev.* **25**, 2245–2253 (2018).
30. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
31. Metcalfe, J. & Shimamura, A. P. *Metacognition: Knowing About Knowing*. (MIT Press, 1994).
32. Fleming, S. M. & Lau, H. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
33. Rosenthal, C. R. R., Andrews, S. K. K., Antoniades, C. A. A., Kennard, C. & Soto, D. Learning and recognition of a non-conscious sequence of events in human primary visual cortex. *Curr. Biol.* **26**, 834–841 (2016).
34. Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D. & Seth, A. K. Blind insight: metacognitive discrimination despite chance task performance. *Psychol. Sci.* **25**, 2199–2208 (2014).
35. Faivre, N., Filevich, E., Solovey, G., Kühn, S. & Blanke, O. Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognition. *J. Neurosci.* **38**, 263–277 (2018).
36. Morales, J., Lau, H. & Fleming, S. M. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J. Neurosci.* **38**, 3534–3546 (2018).
37. Houtkoop, B. L. et al. Data sharing in psychology: a survey on barriers and preconditions. *Adv. Methods Pract. Psychol. Sci.* **1**, 70–85 (2018).
38. King, G. An introduction to the dataverse network as an infrastructure for data sharing. *Sociol. Methods Res.* **36**, 173–199 (2007).
39. Alter, G. & Gonzalez, R. Responsible practices for data sharing. *Am. Psychol.* **73**, 146–156 (2018).
40. Martone, M. E., Garcia-Castro, A. & VandenBos, G. R. Data sharing in psychology. *Am. Psychol.* **73**, 111–125 (2018).
41. Mello, M. M. et al. Preparing for responsible sharing of clinical trial data. *N. Engl. J. Med.* **369**, 1651–1658 (2013).
42. Tenopir, C. et al. Data sharing by scientists: practices and perceptions. *PLoS One* **6**, e21101 (2011).
43. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. & McGillivray, B. The citation advantage of linking publications to research data. Preprint at *arXiv* https://arxiv.org/abs/1907.02565 (2019).
44. Milham, M. P. et al. Assessment of the impact of shared brain imaging data on the scientific literature. *Nat. Commun.* **9**, 2818 (2018).

Acknowledgements

The organization of the Confidence Database was supported by the National Institute of Mental Health under award number R56MH119189 to D.R. The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

The Confidence Database was conceived and organized by D.R., who also drafted the paper. K. Desender, A.L.F.L. and D.R. performed the analyses. D.R., K. Davranche, A.L.F.L., W.T.A., D.A.-L., B.A., P.A., L.Y.A., F.B., J.W.B., I.B., D.P.B., T.F.B., J.C.-T., A.C., T.K.C., K.S.D., R.N.D., T.C.D., K.S.D., Y.A.D., N.F., K.F., E.F., T.G., R.M.G., V.d.G., S.G., N.H., M.H., T.-Y.H., X.H., I.I., M.J., J.K., M. Koculak, M. Konishi, C.K., P.D.K., S.C.K., M.L., K.M.L., C.M.L., L.L., B.M., A. Martin, S.M., J.M., A. Mazancieux, D.M.M., D.O., E.R.P., B.P., M.P., C.P., M.G.P., G.P., F.P., M. Rausch, S.R., G.R., M. Rouault, J. Sackur, S. Sadeghi, J. Samaha, T.X.F.S., M. Shekhar, M.T.S., M. Siedlecka, Z.S., C.S., D.S., S. Sun, J.J.A.v.B., S.W., C.T.W., G.W., M.W., X.X., Q.Y., J.Y., F.Z. and A.Z. contributed to the database. All of the contributors at the time of publication are listed as authors in alphabetical order except for the first three authors. All of the authors edited and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0813-1.

Correspondence and requests for materials should be addressed to D.R.

Peer review information Primary Handling Editor: Marike Schiffer.

[1]School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA. [2]Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [3]Department of Experimental Psychology, Ghent University, Ghent, Belgium. [4]Department of Applied Psychology and Wofoo Joseph Lee Consulting and Counselling Psychology Research Centre, Lingnan University, Tuen Mun, Hong Kong. [5]Center for Neural Science, New York University, New York, NY, USA. [6]Centre d'Économie de la Sorbonne, CNRS & Université Paris 1 Panthéon-Sorbonne, Paris, France. [7]Department of Psychology, Columbia University, New York, NY, USA. [8]Bernstein Center for Computational Neuroscience, Berlin, Germany. [9]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany. [10]Institute of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany. [11]National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda, MD, USA. [12]National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA. [13]National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD, USA. [14]Department of Psychology, Koç University, Istanbul, Turkey. [15]Department of Ophthalmology, New York University (NYU) School of Medicine, NYU Langone Health, New York, NY, USA. [16]Department of Psychiatry and Mental Health, University Hospitals of Geneva and University of Geneva, Geneva, Switzerland. [17]School of Psychology, University of Sydney, Sydney, New South Wales, Australia. [18]Department of Psychology, University of California, San Diego, La Jolla, CA, USA. [19]Department of Experimental Psychology, University of Oxford, Oxford, UK. [20]Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, the Netherlands. [21]Smead Aerospace Engineering Sciences, University of Colorado, Boulder, CO, USA. [22]Aix Marseille University, CNRS, LPC, Marseille, France. [23]Department of Psychology and Center for Neural Science, New York University, New York, NY, USA. [24]Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden. [25]Department of Education, University of Oxford, Oxford, UK. [26]Laboratoire de Psychologie et Neurocognition, Université Grenoble Alpes, Grenoble, France. [27]Department of Psychology, University of Victoria, Victoria, British Columbia, Canada. [28]School of Psychology, University of Queensland, Brisbane, Queensland, Australia. [29]Department of Experimental & Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. [30]School of Psychological Sciences, Monash University, Melbourne, Victoria, Australia. [31]Paris School of Economics and CNRS, Paris, France. [32]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [33]Feinstein Institute for Medical Research, Manhasset, NY, USA. [34]Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands. [35]Graduate Institute of Mind, Brain, and Consciousness, Taipei Medical University, Taipei, Taiwan. [36]Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China. [37]National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. [38]Department of Psycholgoy, California State University, Northridge, CA, USA. [39]Consciousness Lab, Institute of Psychology, Jagiellonian University, Krakow, Poland. [40]Laboratoire de Sciences Cognitives et de Psycholinguistique, Department d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS, Paris, France. [41]Department of Psychology, University of Florida, Gainesville, FL, USA. [42]Shanghai Key Laboratory of Brain Functional Genomics, Key Laboratory of Brain Functional Genomics Ministry of Education, School of Psychology and Cognitive Science, East China Normal University, Shanghai, China. [43]Shanghai Key Laboratory of Magnetic Resonance, East China Normal University, Shanghai, China. [44]NYU-ECNU Institute of Brain and Cognitive Science, NYU Shanghai, Shanghai, China. [45]Swiss Center for Affective Science and LaBNIC, Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland. [46]Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA. [47]Brain and Consciousness Research Centre, TMU Shuang-Ho Hospital, New Taipei City, Taiwan. [48]Department of Bioengineering, University of California, Riverside, Riverside, CA, USA. [49]Université de Lorraine, Université de Strasbourg, CNRS, BETA, Nancy, France. [50]Philosophy Department, Monash University, Monash, Victoria, Australia. [51]Otolaryngology-Head and Neck Surgery, The Ohio State University, Columbus, OH, USA. [52]School of Psychology, National University of Ireland Galway, Galway, Ireland. [53]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. [54]Psychology and Language Sciences, University College Londo, London, UK. [55]Institute of Neurology, University College London, London, UK. [56]SWPS University of Social Sciences and Humanities, Katowice Faculty of Psychology, Katowice, Poland. [57]Laboratory of Cognitive Neuroscience, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [58]Department of Psychology, UiT the Arctic University of Norway, Tromso, Norway. [59]Faculty of Psychology, Universidad del Desarrollo, Santiago, Chile. [60]Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany. [61]Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École normale supérieure—PSL University, CNRS, Paris, France. [62]Département d'Études Cognitives, École Normale Supérieure—PSL University, CNRS, EHESS, INSERM, Paris, France. [63]École Polytechnique, Palaiseau, France. [64]Department of Human Development, Cornell University, Ithaca, NY, USA. [65]Department of Psychology, University of California, Santa Cruz, Santa Cruz, CA, USA. [66]School of Psychology, Trinity College Dublin, Dublin, Ireland. [67]Sackler Centre for Consciousness Science, Brighton, UK. [68]Brighton and Sussex Medical School, University of Sussex, Brighton, UK. [69]Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff University, Cardiff, UK. [70]Basque Center on Cognition, Brain and Language, San Sebastian, Spain. [71]Ikerbasque, Basque Foundation for Science, Bilbao, Spain. [72]Divisions of Biology and Biological Engineering and Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA. [73]Discipline of Psychology, University of Canberra, Canberra, Australian Capital Territory, Australia. [74]Department of Chemical and Biomedical Engineering and Rockefeller Neuroscience Institute, West Virginia University, Morgantown, WV, USA. [75]Department of Psychology, Swansea University, Swansea, UK. [76]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA. *e-mail: rahnev@psych.gatech.edu

nature research

Corresponding author(s): Dobromir Rahnev

Last updated by author(s): Nov 5, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Individual datasets included in the Confidence Database typically use standard data collection methods and those are described in detail in the individual publications for each dataset.
Data analysis	Codes for all data analyses are provided.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are available at osf.io/s46pr.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Data are quantitative and include stimulus, response, reaction time, and confidence.
Research sample	Each dataset has a different sample of participants. Details regarding each sample can be found in the original publications associated with each dataset.
Sampling strategy	Details about the sampling strategy of individual studies can be found in the original publications associated with each dataset.
Data collection	Details about the data collection strategy of individual studies can be found in the original publications associated with each dataset.
Timing	Information about when data were collected is present in the individual readme files on the OSF website.
Data exclusions	Detailed information about data exclusions in the analyses that we report is present in the manuscript.
Non-participation	Details about non-participation in each individual dataset can be found in the original publications associated with each dataset.
Randomization	Details about randomization for each individual dataset can be found in the original publications associated with each dataset. The majority of datasets did not include multiple groups and thus random assignment was not needed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Each dataset has a different sample of participants. Details regarding each sample can be found in the original publications associated with each dataset.
Recruitment	Information about when data were collected is present in the individual readme files on the OSF website.
Ethics oversight	Each dataset was approved by a corresponding IRB committee that is identified in the paper associated with each dataset.

Note that full information on the approval of the study protocol must also be provided in the manuscript.