



The Confidence Database

Dobromir Rahnev, Kobe Desender, Alan Lee, William Adler, David Aguilar-Lleyda, Başak Akdoğan, Polina Arbuzova, Lauren Atlas, Fuat Balci, Ji Won Bang, et al.

► To cite this version:

Dobromir Rahnev, Kobe Desender, Alan Lee, William Adler, David Aguilar-Lleyda, et al.. The Confidence Database. *Nature Human Behaviour*, 2020, 4 (3), pp.317-325. 10.1038/s41562-019-0813-1 . hal-02958766

HAL Id: hal-02958766

<https://hal.science/hal-02958766>

Submitted on 22 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 The Confidence Database

2
3 Dobromir Rahnev (0000-0002-5265-2559)^{1*}, Kobe Desender^{2,3}, Alan L. F. Lee (0000-0001-6722-6889)⁴,
4 William T. Adler⁵, David Aguilar-Lleyda (0000-0001-6963-4069)⁶, Başak Akdoğan (0000-0002-6115-
5 9897)⁷, Polina Arbuzova (0000-0002-4917-9633)^{8,9,10}, Lauren Y. Atlas^{11,12,13}, Fuat Balci (0000-0003-
6 3390-9352)¹⁴, Ji Won Bang¹⁵, Indrit Bègue (0000-0001-5350-4454)¹⁶, Damian P. Birney¹⁷, Timothy F.
7 Brady (0000-0001-5924-5211)¹⁸, Joshua Calder-Travis (0000-0003-3764-2042)¹⁹, Andrey
8 Chetverikov^{20,21}, Torin K. Clark(0000-0002-9345-9712)²², Karen Davranche²³, Rachel N. Denison²⁴, Troy
9 C. Dildine^{11,25}, Kit S. Double²⁶, Yalçın A. Duyan¹⁴, Nathan Faivre (0000-0001-6011-4921)²⁷, Kaitlyn
10 Fallow (0000-0002-7178-9403)²⁸, Elisa Filevich^{8,9,10}, Thibault Gajdos²³, Regan M. Gallagher (0000-
11 0003-2600-8273)^{29,30,31}, Vincent de Gardelle³², Sabina Gherman^{33,34}, Nadia Haddara¹, Marine
12 Hainguerlot³⁵, Tzu-Yu Hsu (0000-0001-6157-9114)³⁶, Xiao Hu³⁷, Iñaki Iturrate (0000-0001-7781-
13 7826)³⁸, Matt Jaquiere (0000-0003-3599-1580)¹⁹, Justin Kantner³⁹, Marcin Koculak (0000-0002-8839-
14 2365)⁴⁰, Mahiko Konishi (0000-0001-8212-9964)⁴¹, Christina Koß^{8,9,10}, Peter D. Kvam⁴², Sze Chai Kwok
15 (0000-0002-7439-1193)^{43,44,45}, Maël Lebreton (0000-0002-2071-4890)⁴⁶, Karolina M. Lempert⁴⁷, Chien
16 Ming Lo^{36,48}, Liang Luo³⁷, Brian Maniscalco⁴⁹, Antonio Martin³⁶, Sébastien Massoni (0000-0001-6980-
17 7505)⁵¹, Julian Matthews (0000-0002-1895-3403)^{52,53}, Audrey Mazancieux⁵⁴, Daniel M. Merfeld⁵⁵,
18 Denis O’Hora (0000-0003-3776-6782)⁵⁶, Eleanor R. Palser^{57,58,59}, Borysław Paulewicz⁴⁰, Michael
19 Pereira (0000-0003-0778-674X)⁶⁰, Caroline Peters^{8,9,10}, Marios G. Philiastides (0000-0002-7683-
20 3506)³³, Gerit Pfuhl (0000-0002-3271-6447)⁶¹, Fernanda Prieto⁶², Manuel Rausch (0000-0002-5805-
21 5544)⁶³, Samuel Recht (0000-0003-0317-2994)⁶⁴, Gabriel Reyes⁶², Marion Rouault (0000-0001-6586-
22 3788)⁶⁵, Jérôme Sackur^{65,66}, Saeedeh Sadeghi (0000-0002-0668-8147)⁶⁷, Jason Samaha⁶⁸, Tricia X.F.
23 Seow (0000-0002-5930-8929)⁶⁹, Medha Shekhar¹, Maxine T. Sherman (0000-0002-0954-7040)^{70,71},
24 Marta Siedlecka (0000-0002-0688-4500)⁴⁰, Zuzanna Skóra (0000-0001-6018-8031)⁴⁰, Chen Song
25 (0000-0002-5418-5747)⁷², David Soto (0000-0003-0205-7513)^{73,74}, Sai Sun⁷⁵, Jeroen J.A. van Boxtel
26 (0000-0003-2643-0474)⁷⁶, Shuo Wang⁷⁷, Christoph T. Weidemann (0000-0002-4280-2744)⁷⁸, Gabriel
27 Weindel²³, Michał Wierzchoń (0000-0002-7347-2696)⁴⁰, Xinming Xu⁴³, Qun Ye⁴³, Jiwon Yeon¹, Futing
28 Zou⁴³, Ariel Zylberberg⁷⁹

30 ¹ School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

31 ² Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-
32 Eppendorf, Hamburg, Germany

33 ³ Department of Experimental Psychology, Ghent University, Ghent, Belgium

34 ⁴ Department of Applied Psychology and Wofoo Joseph Lee Consulting and Counselling Psychology
35 Research Centre, Lingnan University, Hong Kong

36 ⁵ Washington, DC, USA

37 ⁶ Centre d’Économie de la Sorbonne (CNRS & Université Paris 1 Panthéon-Sorbonne), Paris, France

38 ⁷ Department of Psychology, Columbia University, New York, NY, USA

39 ⁸ Bernstein Center for Computational Neuroscience, Berlin, Germany

40 ⁹ Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Germany

41 ¹⁰ Institute of Psychology, Humboldt Universität zu Berlin, Germany

42 ¹¹ National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda,
 43 MD, USA
 44 ¹² National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
 45 ¹³ National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD, USA
 46 ¹⁴ Department of Psychology, Koç University, Istanbul, Turkey
 47 ¹⁵ Department of Ophthalmology, School of Medicine, New York University, NY, USA
 48 ¹⁶ Department of Psychiatry and Mental health, University hospitals of Geneva and University of
 49 Geneva, Geneva, Switzerland
 50 ¹⁷ School of Psychology, University of Sydney, Sydney, Australia
 51 ¹⁸ Department of Psychology, University of California, San Diego, La Jolla, CA, USA
 52 ¹⁹ Department of Experimental Psychology, University of Oxford, Oxford, UK
 53 ²⁰ Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, The
 54 Netherlands
 55 ²¹ Cognitive Research Lab, Russian Academy of National Economy and Public Administration, Moscow,
 56 Russia
 57 ²² Smead Aerospace Engineering Sciences, University of Colorado, Boulder, CO, USA
 58 ²³ Aix Marseille University, CNRS, LPC, Marseille, France
 59 ²⁴ Department of Psychology and Center for Neural Science, New York University, New York, NY,
 60 USA
 61 ²⁵ Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden
 62 ²⁶ Department of Education, University of Oxford, Oxford, UK
 63 ²⁷ Laboratoire de Psychologie et Neurocognition, Université Grenoble Alpes, Grenoble, France
 64 ²⁸ Department of Psychology, University of Victoria, Victoria, BC, Canada
 65 ²⁹ School of Psychology, University of Queensland, Brisbane, Australia
 66 ³⁰ Department of Experimental & Applied Psychology, Vrije Universiteit, Amsterdam, Netherlands
 67 ³¹ School of Psychological Sciences, Monash University, Melbourne, Australia
 68 ³² Paris School of Economics and CNRS, Paris, France
 69 ³³ Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK
 70 ³⁴ Feinstein Institute for Medical Research, Manhasset, NY, USA
 71 ³⁵ Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands
 72 ³⁶ Graduate Institute of Mind, Brain, and Consciousness, Taipei Medical University, Taipei, Taiwan
 73 ³⁷ Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal
 74 University, Beijing, China
 75 ³⁸ National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda,
 76 MD, USA
 77 ³⁹ Department of Psychology, California State University, Northridge, Northridge, CA, USA
 78 ⁴⁰ Consciousness Lab, Institute of Psychology, Jagiellonian University, Kraków, Poland
 79 ⁴¹ Laboratoire de Sciences Cognitives et de Psycholinguistique, Dept d'Etudes Cognitives, ENS, PSL
 80 University, EHESS, CNRS, Paris, France
 81 ⁴² Department of Psychology, University of Florida, Gainesville, FL, USA
 82 ⁴³ Shanghai Key Laboratory of Brain Functional Genomics, Key Laboratory of Brain Functional
 83 Genomics Ministry of Education, School of Psychology and Cognitive Science, East China Normal
 84 University, Shanghai, China
 85 ⁴⁴ Shanghai Key Laboratory of Magnetic Resonance, East China Normal University, Shanghai, China

86 ⁴⁵ NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai, China
 87 ⁴⁶ Swiss Center for Affective Science and LaBNIC, Department of Basic Neuroscience, University of
 88 Geneva, Geneva, Switzerland
 89 ⁴⁷ Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA
 90 ⁴⁸ Brain and Consciousness Research Centre, TMU Shuang-Ho Hospital, New Taipei City, Taiwan
 91 ⁴⁹ Department of Bioengineering, University of California Riverside, CA, USA
 92 ⁵⁰ Graduate Institute of Mind, Brain, and Consciousness, Taipei Medical University, Taipei, Taiwan
 93 ⁵¹ Université de Lorraine, Université de Strasbourg, CNRS, BETA, Nancy, France
 94 ⁵² Philosophy Department, Monash University, Australia
 95 ⁵³ School of Psychological Sciences, Monash University, Australia
 96 ⁵⁴ Laboratoire de Psychologie et Neurocognition, Université Grenoble Alpes, Grenoble, France
 97 ⁵⁵ Otolaryngology-Head and Neck Surgery, The Ohio State University, Columbus, OH, USA
 98 ⁵⁶ School of Psychology, National University of Ireland Galway, Galway, Ireland
 99 ⁵⁷ Department of Neurology, University of California San Francisco, San Francisco, California, USA
 100 ⁵⁸ Psychology and Language Sciences, University College London, London, UK
 101 ⁵⁹ Department of Neurology, University College London, London, UK
 102 ⁶⁰ Laboratory of Cognitive Neuroscience, Ecole Polytechnique Fédérale de Lausanne, Switzerland
 103 ⁶¹ Department of Psychology, UiT the Arctic University of Norway, Norway
 104 ⁶² Faculty of Psychology, Universidad del Desarrollo, Santiago, Chile.
 105 ⁶³ Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany
 106 ⁶⁴ Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure -
 107 PSL University, CNRS, France
 108 ⁶⁵ Département d'études cognitives, École normale supérieure - PSL University, CNRS, EHESS, INSERM,
 109 Paris, France
 110 ⁶⁶ École Polytechnique, Palaiseau, France.
 111 ⁶⁷ Department of Human Development, Cornell University, Ithaca, NY, USA
 112 ⁶⁸ Department of Psychology, University of California, Santa Cruz, CA, USA
 113 ⁶⁹ School of Psychology, Trinity College Dublin, Dublin, Ireland
 114 ⁷⁰ Sackler Centre for Consciousness Science, Brighton, UK
 115 ⁷¹ Brighton and Sussex Medical School, University of Sussex, Brighton, UK
 116 ⁷² Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff University, Cardiff,
 117 UK
 118 ⁷³ Basque Center on Cognition, Brain and Language, San Sebastian, Spain
 119 ⁷⁴ Ikerbasque, Basque Foundation for Science, Bilbao, Spain
 120 ⁷⁵ Division of Biology and Biological Engineering/Computation and Neural Systems, California Institute
 121 of Technology, Pasadena, CA, USA
 122 ⁷⁶ Discipline of Psychology, University of Canberra, Bruce, ACT, Australia
 123 ⁷⁷ Department of Chemical and Biomedical Engineering and Rockefeller Neuroscience Institute, West
 124 Virginia University, Morgantown, USA
 125 ⁷⁸ Department of Psychology, Swansea University, Swansea, UK
 126 ⁷⁹ Mind Brain Behavior Institute, Columbia University, New York, USA
 127
 128 * Corresponding author: Dobromir Rahnev (rahnev@psych.gatech.edu)

Abstract

Understanding how people rate their confidence is critical for characterizing 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 analyzed in multiple software packages. Each dataset is further accompanied by an explanation regarding the nature of the collected data. At the time of publication, the Confidence Database (available at osf.io/s46pr) contained 145 datasets with data from over 8,700 participants and almost 4 million trials. The database will remain open for new submissions indefinitely and is expected to continue to grow. We show the usefulness of this large collection of datasets in four different analyses that provide precise estimation for several foundational confidence-related effects.

Main

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 instance, participants may first decide whether a probe stimulus belongs to a previously learned study list or not. A confidence rating, in this case, could involve the participants' second-order judgment regarding how sure they are about the accuracy of the decision made in that trial (i.e., accuracy of the first-order task performance). Such second-order judgments reflect people's ability to introspect and can be dissociated from the first-order judgment¹. Confidence ratings tend to

150 correlate strongly with accuracy, response speed, and brain activity distinguishing old and new
151 probes² suggesting that they reflect relevant internal states.

152

153 The question of how humans (or other animals) evaluate their own decisions has always been an
154 important topic in psychology, and the use of confidence ratings dates back to the early days of
155 experimental psychology³. In addition, confidence has been used as a tool to, among many other
156 things, determine the number of distinct memory retrieval processes⁴, reveal distortions of visual
157 awareness⁵, understand the factors that guide learning⁶, assess the reliability of eyewitness
158 testimony⁷, test theories of sensory processing⁸ and decision-making^{9,10}, help estimate the fit of
159 parameters of the psychometric function more efficiently¹¹, and characterize various psychiatric
160 conditions¹². The wide application of confidence makes it a fundamental measure in psychological
161 research.

162

163 However, despite the widespread use of confidence ratings, scientific progress has been slowed by
164 the traditional unavailability of previously collected data. In the current system, testing a new idea
165 often requires scientists to spend months or years gathering the relevant data. The substantial cost in
166 time and money associated with new data collection has undoubtedly led to many new ideas simply
167 being abandoned without ever being examined empirically. This is especially unfortunate given that
168 these ideas could likely have been tested using the dozens of datasets already collected by other
169 scientists.

170

171 Typically, when data re-use takes place, it is within a lab or a small scientific group -- that often
172 restricts itself to very specific paradigms -- which potentially limits the formation of a broader
173 understanding of confidence across a wider range of tasks and participants. Therefore, another
174 important advantage of data re-use lies in the diversity of experimental tasks, set-ups, and
175 participants offered by compiling datasets from different labs and different populations.

176

177 Although data sharing can speed up scientific progress considerably, fields devoted to understanding
178 human behavior unfortunately have cultures of not sharing data^{13,14}. For example, Wicherts et al.¹⁵
179 documented their painstaking and ultimately unsuccessful endeavor to obtain behavioral data for re-
180 analysis; despite persistent efforts, the authors were able to obtain just 25.7% of datasets the authors
181 claimed to be available for re-analysis. Nevertheless, recent efforts towards increased openness have
182 started to shift the culture considerably and more and more authors post their data in online
183 depositories^{16,17}.

184

185 There are, however, several challenges involved in secondary analyses of data, even when such data
186 have been made freely available. First, the file type may not be usable or clear for some researchers.
187 For example, sharing files in proprietary formats may limit other researcher's ability to access them
188 (e.g., if reading the file requires software that is not freely or easily obtainable). Second, even if the
189 data can be readily imported and used, important information about the data may not have been
190 included. Third, researchers who need data from a large number of studies have to spend a
191 considerable amount of time finding individual datasets, familiarizing themselves with how each

192 dataset is structured, and organizing all datasets into a common format for analysis. Finally, given the
193 size of the literature, it can be difficult to even determine which papers contain relevant data.

194

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

202

203 **Details on the database**

204 The Confidence Database is hosted on the Open Science Framework (OSF) website (osf.io/s46pr).

205 Each dataset is represented by two files – a data file in .csv format and a readme file in .txt format.

206

207 The majority of data files contain the following fields: participant index, stimulus, response,
208 confidence, response time of the decision, and response time of the confidence rating. Depending on
209 the specific design of each study, these fields can be slightly different (e.g., if there are two stimuli on
210 each trial or confidence and decision are given with a single button press). Further, many datasets
211 include additional fields needed to fully describe the nature of the collected data.

212

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

219
220 The Confidence Database includes a wide variety of studies. Individual datasets recruit different
221 populations (e.g., healthy or patient populations), focus on different fields of study (e.g., perception,
222 memory, motor control, decision making), employ different confidence scales (e.g., binary, n-point
223 scales, continuous scales, wagering), use different types of tasks (e.g., binary judgements vs.
224 continuous estimation tasks), and collect confidence at different times (e.g., after or simultaneous
225 with the decision). Figure 1 gives a broad overview of the types of datasets included in the database
226 at the time of publication. This variety ensures that future re-analyses can address a large number of
227 scientific questions and test them based on multiple methods of evaluating one’s own primary task
228 performance.

229
230 Importantly, the database will remain open for new submissions indefinitely. Instructions for new
231 submissions are made available on the OSF page of the database. Carefully formatted .csv and .txt
232 files that follow the submission instructions can be e-mailed to confidence.database@gmail.com.
233 They will be checked for quality and then uploaded with the rest of the database.

234

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

At the time of publication, the Confidence Database contained 145 datasets, bringing together 8,787 participants, for 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 thus placing the data in the public domain. As with the re-use of any other data, publications that result from such re-analysis should cite the current paper, as well as the listed citation for each of the datasets that were re-analyzed. We highly 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 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 response times; and building theories of the response times associated with confidence ratings. Further, the database can also be used to test hypotheses unrelated to confidence due to the inclusion of choice, accuracy, and response time. Different studies can re-use a few relevant datasets (maybe even a single one) or simultaneously analyze a large set of the available datasets thus achieving substantially higher power than typical individual studies.

Below we present results from four different example analyses in order to demonstrate the potential utility and versatility of the database. These analyses are designed to take advantage of a large proportion of the available data, thus resulting in very large sample sizes. Annotated codes for running these analyses are freely available at the OSF page of the database (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.

Analysis 1: How confidence is related to choice and confidence response times (RTs)

One of the best known properties of confidence ratings is that they correlate negatively with choice RT^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 a replicable quantitative science requires that researchers, among other things, “adopt estimation thinking and avoid dichotomous thinking”¹⁹. Precise estimation, though, requires very large sample sizes and any individual study is usually not large enough to allow for accuracy in estimation. The Confidence Database thus provides a unique opportunity to estimate with unprecedented precision the strength of foundational effects such as the negative correlation between confidence and choice RT, thus informing theories that rely on these effects. Further, the database allows for investigations of lesser studied relationships such as between confidence and confidence RT.

Using the data from the Confidence Database, we thus investigated the precise strength of the correlation of confidence with both choice and confidence RT. We first selected all datasets where choice and confidence RTs were reported. Note that some datasets featured designs where the choice and confidence were made with a single button press -- such datasets were excluded from the current analyses. In addition, we excluded individual participants who only used a single level of confidence because it is impossible to correlate confidence and RT for such subjects, 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 cleanup. This step is important as contributors are encouraged to include all participants and trials from an experiment even if some participants or trials were excluded from data analyses in the original publications. Specifically, we

301 excluded all trials without a confidence rating (such trials typically came from studies that included a
302 deadline for the confidence response), all trials without choice RT (typically due to a deadline on the
303 main decision), and all trials with confidence and/or choice RTs slower than 5 seconds (the results
304 remained very similar if a threshold of 3 or 10 seconds was used instead). These exclusion criteria
305 resulted in removing 7.3% of the data. In addition, for each participant, we excluded all choice and
306 confidence RTs differing by more than 3 standard deviations from the mean (resulting in the removal
307 of additional 1.8% of the data).

308
309 We then correlated, for each participant, the confidence ratings with choice RTs. We found that the
310 average correlation across participants was $r = -.24$ ($t(4088) = -71.09$, $p < 2.2e-16$, $d = 1.11$). The very
311 large sample size allowed us to estimate the average correlation with a very high degree of precision:
312 the 99.9% confidence interval for the average correlation value was $[-.25, -.23]$, which should be
313 considered as a medium-to-large effect²⁰. At the same time, it is important to emphasize that the high
314 precision in estimating the average correlation does not imply a lack of variability between individual
315 participants. Indeed, we observed very high individual variability ($SD = .21$), which we visualize by
316 plotting all individual correlation values and corresponding density functions in the form of raincloud
317 plots²¹ (Figure 2A). Still, the effect size is large enough that power analyses indicate that a sample size
318 as small as $N=9$ provides >80% power and a sample size of $N=13$ provides >95% power to detect this
319 effect (at $\alpha = .05$).

320
321 We next performed the same analyses for the correlation between confidence and confidence RT. We
322 found that the average correlation across participants was $r = -.07$, $SD = .24$ ($t(4088) = -18.77$, $p <$

2.2e-16, $d = .29$) with a 99.9% confidence interval for the average correlation value of $[-.08, -.06]$. This effect should be considered as “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%, one requires a sample size of $N=93$; for power of 95%, $N=152$ is needed.

It should be noted that existing models of confidence generation (e.g. ²²) predict a lack of any association between confidence and confidence RT (but see ²³). The small but reliable negative correlation thus raises the question about what is causing 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 (for a similar argument in the case of attention, see 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, $r(4087) = .20$, $p < 2.2e-16$ (Figure 2B). Future research should investigate whether this correlation is due to variability in individual participants or variability at the level of the datasets.

Analysis 2: Serial dependence in confidence RT

344 It is well known that perceptual choices²⁵, confidence judgments²⁶, and choice RTs²⁷ are subject to
345 serial dependence. Such findings have been used to make fundamental claims about the nature of
346 perceptual processing such as that the visual system forms a “continuity field” over space and
347 time^{28,29}. The presence of serial dependence can thus help reveal the underlying mechanisms of
348 perception and cognition. However, to the best of our knowledge, the presence of serial dependence
349 has never been investigated for one of the most important components of confidence generation:
350 confidence RT. Therefore, determining whether serial dependence exists for confidence, and if so,
351 estimating precisely its effect size, can therefore provide important insight about the nature of
352 confidence generation.

353
354 To address this question, we considered the data from the Confidence Database. We analyzed all
355 datasets in which confidence was provided with a separate button press from the primary decision
356 and that reported confidence RT. In total, 82 datasets were included, comprising 4,474 participants.
357 Data cleanup was performed as in the previous analysis. Specifically, we removed all trials without
358 confidence RT and all trials with confidence RT slower than 5 seconds (results remained very similar if
359 a threshold of 3 or 10 seconds was used instead), both on the current trial and up to seven trials back,
360 because we wanted to investigate serial dependence up to lag-7 (this excluded a total of 4.3% of the
361 data). Further, as before, we excluded, separately for each participant, all confidence RTs differing by
362 more than 3 standard deviations from the mean (thus excluding additional 9.6% of the data).

363
364 We performed a mixed regression analysis predicting confidence RT with fixed effects for the recent
365 trial history up to seven trials back²⁵ and random intercepts for each participant. Degrees of freedom

were estimated using Satterthwaite's approximation, as implemented in the lmerTest package³⁰. We found evidence for strong autocorrelation in confidence RT. Specifically, there was a very large lag-1 autocorrelation ($b = 1.346$, $t(1299601) = 153.6$, $p < 2.2e-16$; Figure 3). The strength of the autocorrelation dropped sharply for higher lags but remained significantly positive until at least lag-7 (all p 's $< 2.2e-16$).

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 speed up of 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 on subsequent trials.

Analysis 3: 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 decisions³¹. In other words, humans have positive metacognitive sensitivity³², meaning that higher levels of confidence predict better performance. However, it is not uncommon that individual participants fail to show the typically observed positive metacognitive sensitivity. Until now, such cases have been difficult to investigate because they occur infrequently within a given dataset.

387 Using the Confidence Database, we estimated the prevalence of negative metacognitive sensitivity
388 and investigated its causes. We analyzed all datasets that contained the variables confidence and
389 accuracy. In total, 71 datasets were included, comprising of 4,768 participants. We excluded studies
390 on subjective difficulty, because these investigate the relation between confidence and performance
391 *within* correct trials. We further excluded participants who only reported a single level of confidence
392 (since it is impossible to estimate metacognitive sensitivity for such participants), studies with a
393 continuous measure of accuracy, and participants for whom more than 90% of the data were
394 excluded (which occurred for six participants from a study with very high confidence RTs).
395 Metacognitive sensitivity was computed using a logistic regression predicting accuracy by normalized
396 confidence ratings. This measure of metacognition has a number of undesirable properties³² but
397 reliably indicates whether metacognitive sensitivity is positive or negative.

398
399 We found that, across all participants, the average beta value from the logistic regression was .096,
400 $SD = .064$, ($t(4767) = 104.01$, $p < 2.2e-16$, $d = 1.5$; Figure 4A), thus indicating that metacognitive
401 sensitivity was reliably positive in the group. However, 293 of the participants (6.1% of all
402 participants) had a negative beta value, indicating the potential presence of negative metacognitive
403 sensitivity.

404
405 We next explored why such negative coefficients may occur for these 293 participants. We reasoned
406 that the majority of the cases of estimated negative metacognitive sensitivity could be due to several
407 factors unrelated to the true metacognitive sensitivity of each participant. First, the negative beta
408 values could simply be due to misestimation stemming from relatively small sample sizes. Even

409 though the number of trials per participant did not correlate with participants' beta coefficient
 410 ($r(4766) = -.021, p = .143$; Figure 4B), 9.9% of all participants with negative beta value completed less
 411 than 50 trials in total. Second, a positive relationship between confidence and accuracy can be
 412 expected only if performance is above chance (if performance is at chance, this may indicate that
 413 there is no reliable signal that could be used by the metacognitive system, although see^{33,34}). We did
 414 indeed observe a correlation between the beta values and average accuracy ($r(4766) = .203, p < 2.2e-$
 415 16 , Figure 4C) with 19.4% of all participants with negative beta values having an accuracy of less than
 416 55%. Third, for those datasets including choice RT or confidence RT, we calculated the overall median
 417 choice/confidence RTs and correlated these with the beta coefficients (one dataset was excluded
 418 here, because the primary task was to complete Raven's progressive matrices and therefore choice
 419 and confidence RTs were within the range of minutes rather than seconds). Again, we observed
 420 significant correlations between betas and choice RTs ($r(3076) = -.083, p = 3.6e-06$, Figure 4D) and
 421 between betas and confidence RTs ($r(2191) = .071, p = 0.0009$, Figure 4E), but the magnitude of these
 422 correlations was very small and only 2.3% and 2.4% of participants with negative betas had median
 423 choice or confidence RT of less than 200 ms, respectively. Finally, we reasoned that beta coefficients
 424 could be misestimated if a very large proportion of confidence judgments were the same. Therefore,
 425 we computed the proportion of the most common confidence rating for each participant ($M=37.9\%$,
 426 $SD = .22$). We did not observe a significant correlation between the proportion of the most common
 427 confidence rating and the beta values ($r(4766) = -.025, p = .086$, Figure 4F), and only 5.4% of all
 428 participants with negative betas only used a single confidence rating for more than 95% of the time.

429

Overall, 96 participants from the 293 with negative beta values (32.7%) completed less than 50 trials, had 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 beta values despite the absence of any of these factors (note that for 55 of these participants, no RT information was provided, so a few of them could have had overly fast choice or confidence RT). This result raises the question about the underlying causes of the negative beta values. Follow-up studies could focus on these subjects and determine whether there is anything different about them or the tasks that they completed.

Analysis 4: Confidence scales used in perception and memory studies

One of the strengths of the Confidence Database is that it allows for investigations on how specific effects depend on factors that differ from study to study. For example, for any of the analyses above, one could ask how the results depend on factors like the domain of study (i.e., perception, memory, cognitive, etc.), confidence scale used (e.g., n-point vs. continuous), whether confidence was provided simultaneously with the decision, the number of trials per participant, etc. These questions can reveal some of the mechanisms behind confidence generation, such as, for example, 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 meta-science 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 (e.g., a 4-point scale where 1 = lowest confidence, 4 = highest confidence). Such scales typically have a fixed stimulus-response mapping so that a given button always indicates the same level of

confidence (though variable stimulus-response mappings are still possible). Likert scales can also have different number of options. Comparatively fewer studies use continuous scales (e.g., a 0-100 scale where 0 = lowest confidence, 100 = highest confidence). Such scales typically do not have a fixed stimulus-response mapping and responses are often given using a mouse click rather than a button press (though it is possible to use a keyboard in such cases too).

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; see Figure 1). We categorized each dataset from these two domains as employing a 2-point, 3-point, 4-point, 5-point, 6-point, 7-to-11-point, or a continuous confidence scale (we combined the 7- to 11-point scales into a single category because of the low number of datasets with such scales). Finally, we computed the percent 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. Most 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 just 16% of the time (14 out of 89 datasets) with the difference in proportions being significant ($Z = -3.49, p = 0.0005$; Figure 5). On the other hand, a much lower percent of memory datasets (4%, 1 out of 27 datasets) used a continuous scale compared to perception studies (33%, 29 out of 89 datasets; $Z = 3.002, p = 0.003$). Both comparisons remained significant at the .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 's > 0.2 before Bonferroni correction).

These results suggest the presence of 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. Since 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 behavioral sciences

It is a sad reality that “most of the data generated by humanity’s previous scientific endeavors 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 lack of interpretable metadata describing 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 similar databases in other subfields of behavioral science and beyond. Many subfields of psychology produce data that can be fully summarized in a single file using a common format and thus can 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 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 (i.e., favoring significant findings) by simply including datasets that also show null effects. Importantly, it is critical that sharing data is done ethically and that participant anonymity is not compromised^{39–41}. We have followed these principles in assembling the Confidence Database: All datasets have received IRB approvals by the relevant local committees (these can be found in the original publications), all participants have provided informed consent, and all available data are de-identified.

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 rate⁴³, and its accuracy 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 make a case for a more universal effort for data sharing. We hope the construction

and maintenance of the Confidence Database will help 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 (though 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. Therefore, 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 behavioral 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 behavioral 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.

Data availability

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

538 **Code availability**

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

References

1. Mamassian, P. Visual Confidence. *Annu. Rev. Vis. Sci.* **2**, annurev-vision-111815-114630 (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**, 352–74 (1995).
5. Azzopardi, P. & Cowey, A. Is blindsight like normal, near-threshold vision? *Proc. Natl. Acad. Sci.* **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. Public Interes.* **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.* jn.00318.2015 (2016). doi:10.1152/jn.00318.2015

12. David, A. S., Bedford, N., Wiffen, B. & Gilleen, J. Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 1379–90 (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**, (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.* 251524591984868 (2019). doi:10.1177/2515245919848684
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).

584 23. Moran, R., Teodorescu, A. R. & Usher, M. Post choice information integration as a causal
 585 determinant of confidence: Novel data and a computational account. *Cogn. Psychol.* **78**, 99–
 586 147 (2015).

587 24. Nikolov, S., Rahnev, D. & Lau, H. Probabilistic model of onset detection explains paradoxes in
 588 human time perception. *Front. Psychol.* **1**, 37 (2010).

589 25. Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and
 590 alters serial choice bias. *Nat. Commun.* **8**, 14637 (2017).

591 26. Rahnev, D., Koizumi, A., McCurdy, L. Y., D’Esposito, M. & Lau, H. Confidence Leak in Perceptual
 592 Decision Making. *Psychol. Sci.* **26**, 1664–1680 (2015).

593 27. Laming, D. Autocorrelation of choice-reaction times. *Acta Psychol. (Amst)*. **43**, 381–412 (1979).

594 28. Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nat. Neurosci.* **17**, 738–43
 595 (2014).

596 29. Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K. & Whitney, D. Serial dependence in
 597 position occurs at the time of perception. *Psychon. Bull. Rev.* **25**, 2245–2253 (2018).

598 30. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed
 599 Effects Models. *J. Stat. Softw.* **82**, 1–26 (2017).

600 31. Metcalfe, J. & Shimamura, A. P. *Metacognition: Knowing about Knowing*. (MIT Press, 1994).

601 32. Fleming, S. M. & Lau, H. How to measure metacognition. *Front. Hum. Neurosci.* **8**, (2014).

602 33. Rosenthal, C. R. R., Andrews, S. K. K., Antoniadou, C. A. A., Kennard, C. & Soto, D. Learning and
 603 recognition of a non-conscious sequence of events in human primary visual cortex. *Curr. Biol.*
 604 **26**, 834–841 (2016).

605 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. *arXiv* (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).

628 **Acknowledgements**

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

633 **Author contributions**

634 The Confidence Database was conceived and organized by D.R. who also drafted the paper. Analyses
635 were performed by K.D., A.L., and D.R. All contributors at the time of publication are listed as authors
636 in alphabetical order except for the first three authors. All authors also edited and approved the final
637 version of the manuscript.

639 **Competing interests**

640 The authors declare no competing interests.

642 **Figure legends**

643 **Figure 1. Datasets currently in the Confidence Database.** Pie charts showing the number of datasets
644 split by category, publication year, number of participants, number of trials per participant, type of
645 judgment, and rating scale. The label “Multiple” in the first pie chart indicates that the same
646 participants completed tasks from more than one category. The maximum number of participants
647 was 589 and the maximum trials per participant was 4,320 (“variable” indicates that different
648 participants completed different number of trials).

Figure 2. Correlating confidence with choice and confidence RT. (A) We found a medium-to-large negative correlation ($r = -.24, p < 2.2e-16, n = 4,089$) between confidence and choice RT, as well as a small negative correlation ($r = -.07, p < 2.2e-16, n = 4,089$) between confidence and confidence RT. Box shows the median and the interquartile (25-75%) range, whereas the whiskers show the 2-98% range. (B) The strength of the two correlations in panel A were themselves correlated across subjects ($r = .23, p < 2.2e-16, n = 4,089$).

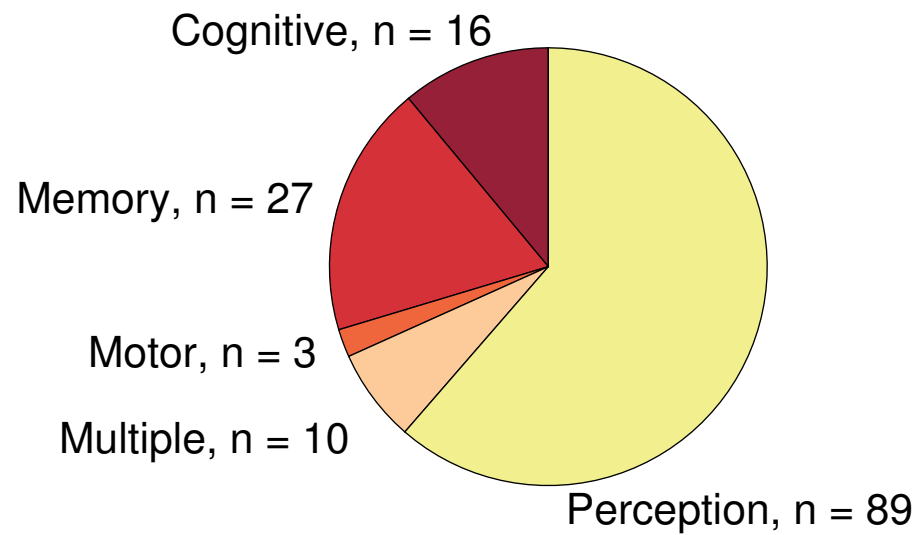
Figure 3. Serial dependence in confidence RT. We observed a large lag-1 autocorrelation ($b = 1.346, t(1299601) = 153.6, p < 2.2e-16, n = 4,474$). The autocorrelation decreased for higher lags but remained significant up to lag-7 (all p 's $< 2.2e-16, n = 4,474$). Error bars indicate SEM. Individual datapoints are not shown because the plots are based on the results of a mixed model analysis.

Figure 4. The prevalence of estimates of negative metacognitive sensitivity. (A) Individual beta values and beta values density plot for the observed relationship between confidence and accuracy. Box shows the median and the interquartile (25-75%) range, whereas the whiskers show the 2-98% range. (B-F) Scatter plots, including lines of best fit, for the relationships between the beta value for 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 where the most common confidence judgment was given (F).

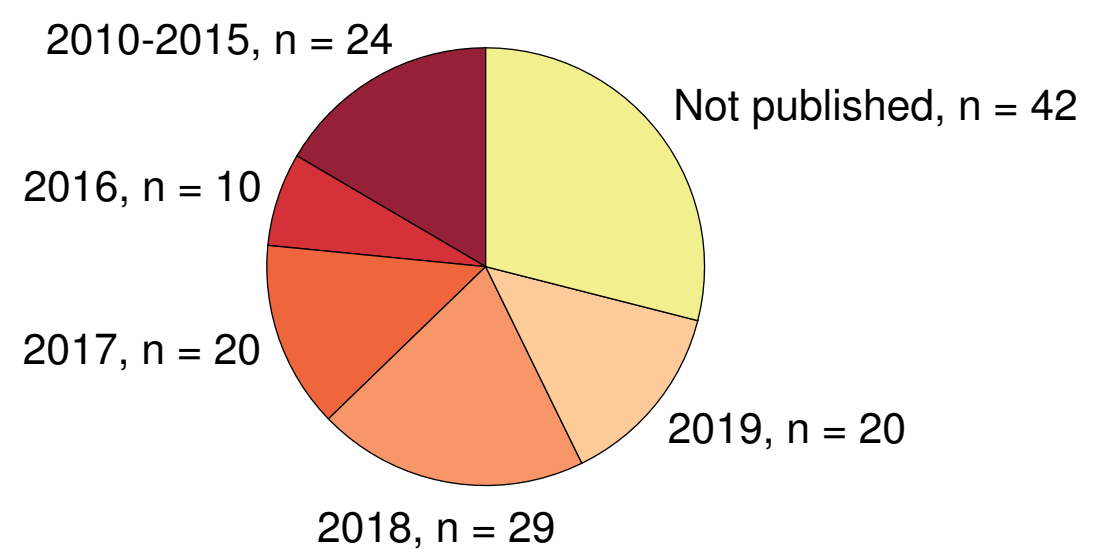
Figure 5. Confidence scale use for perception and memory studies. The percent of 2-point, 3-point, 4-point, 5-point, 6-point, 7-to-11-point, and continuous confidence scales were plotted separately for

672 perception and memory datasets. We combined the 7- to 11-point scales because of the low number
673 of datasets with such scales. The two domains differed in how often they employed 3-point and
674 continuous scales.

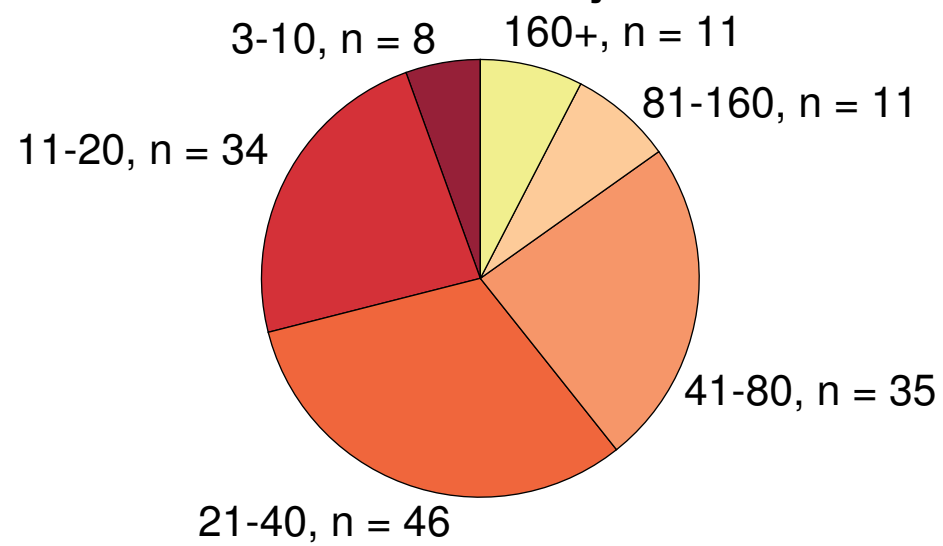
Category



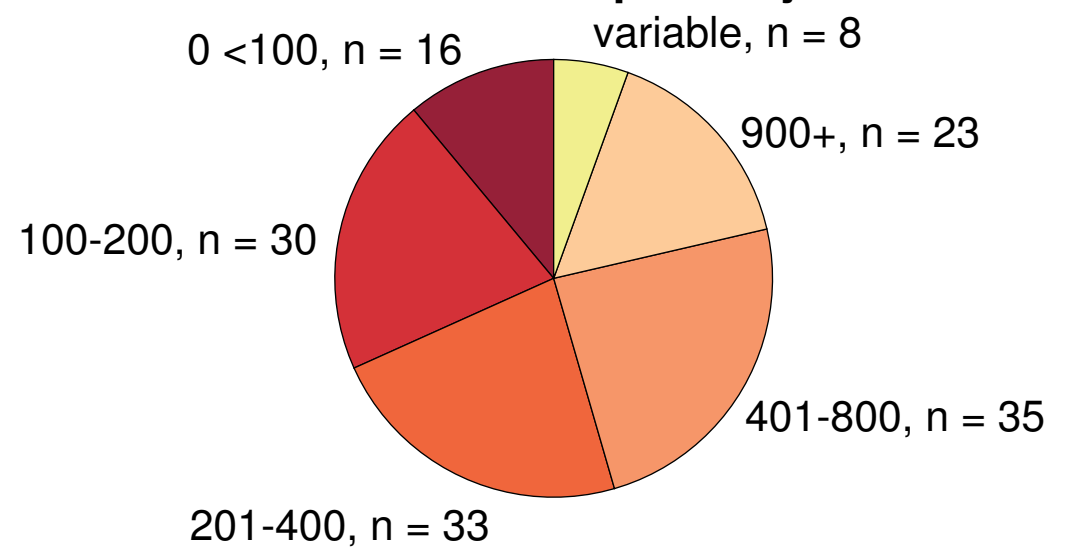
Publication year



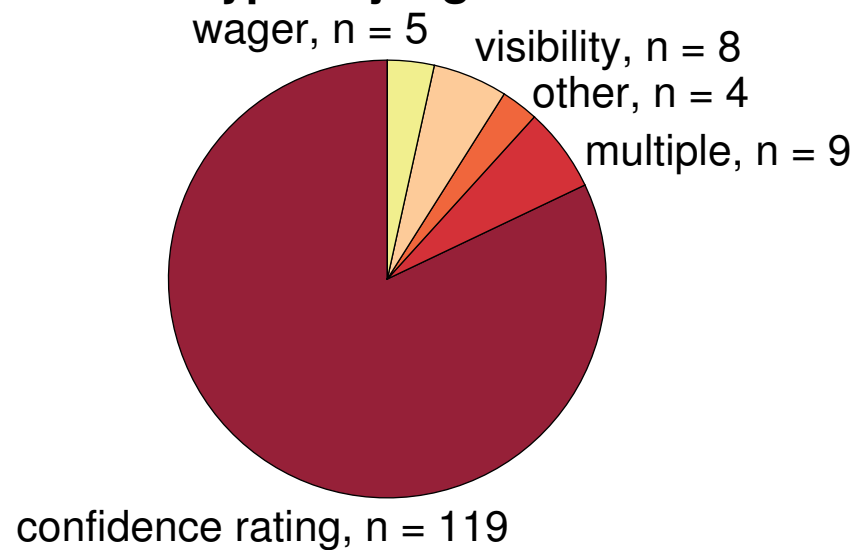
Number of subjects



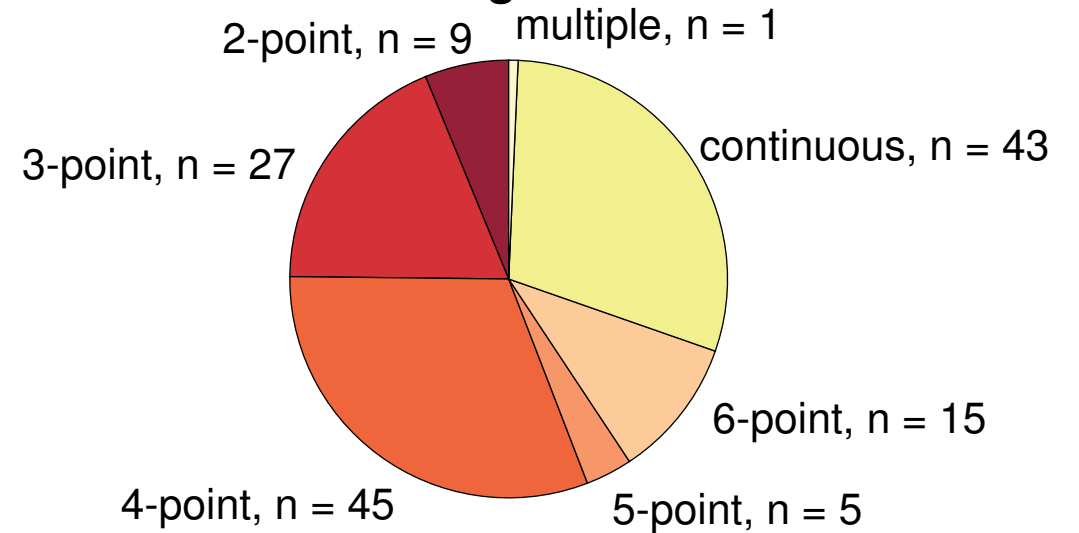
Number of trials per subject



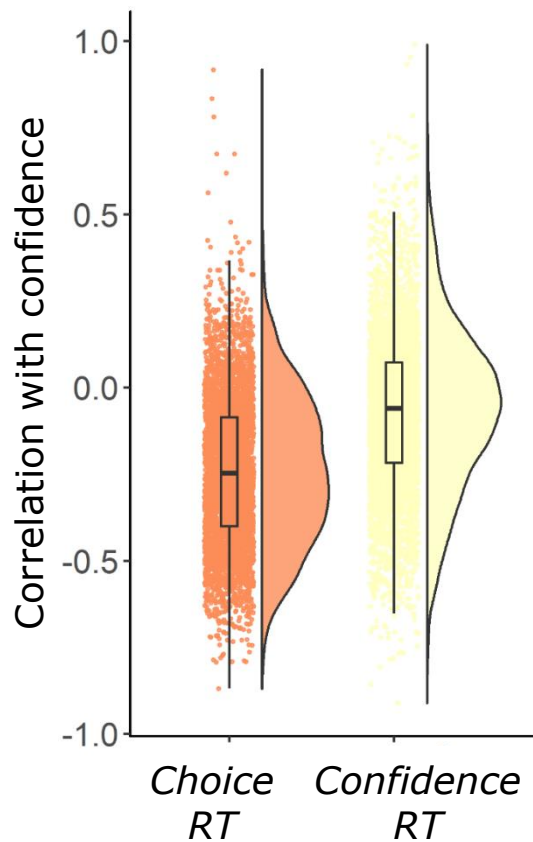
Type of judgment



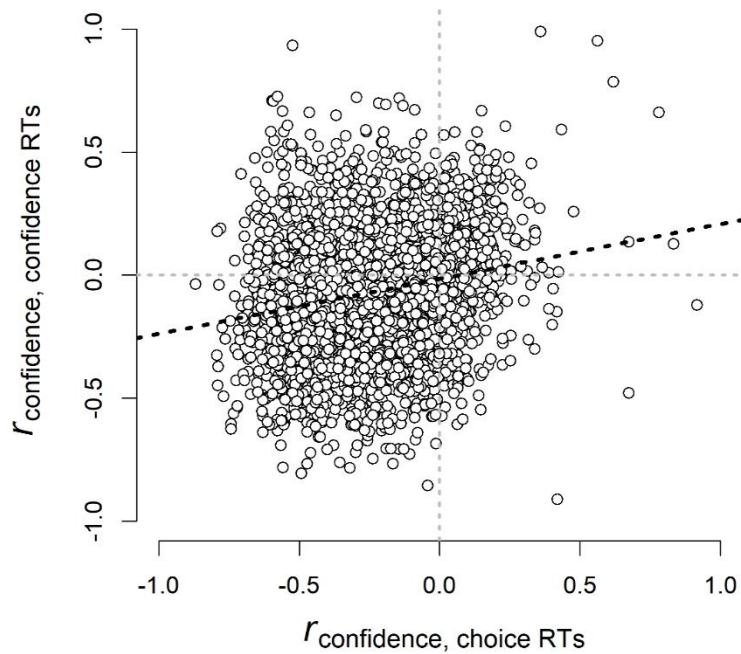
Rating scale

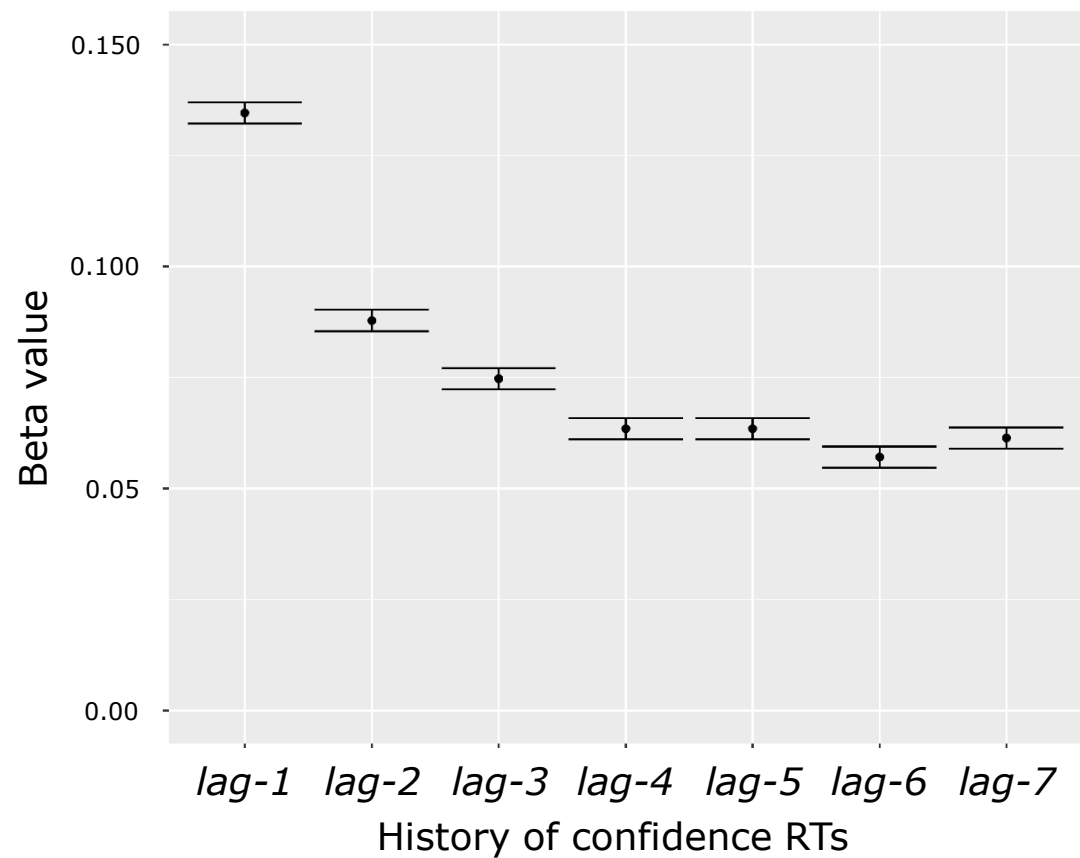


A.

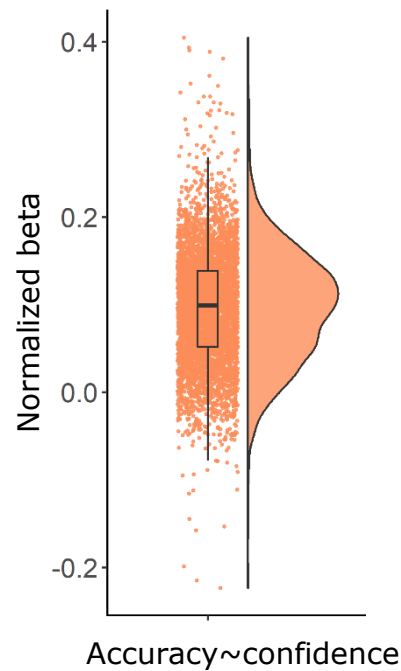


B.

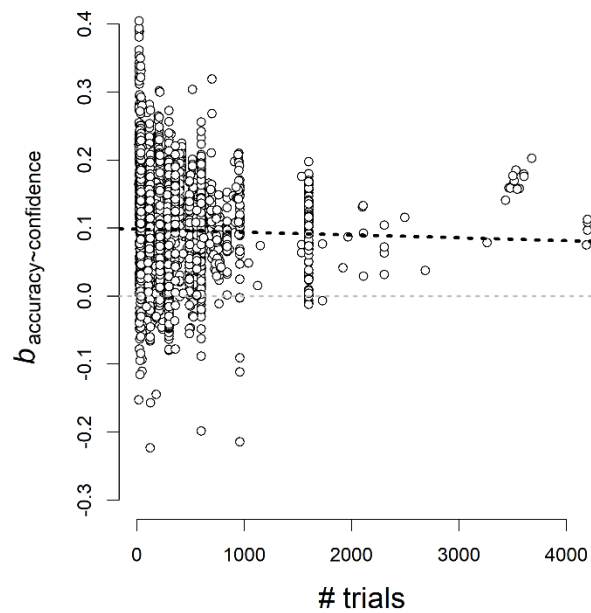




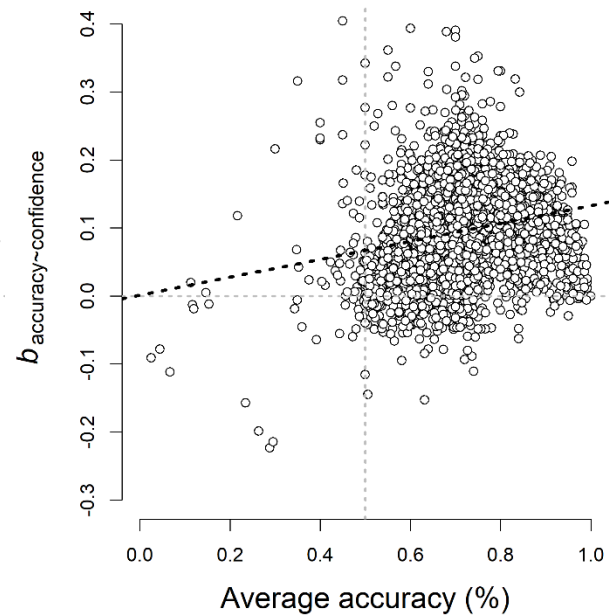
A.



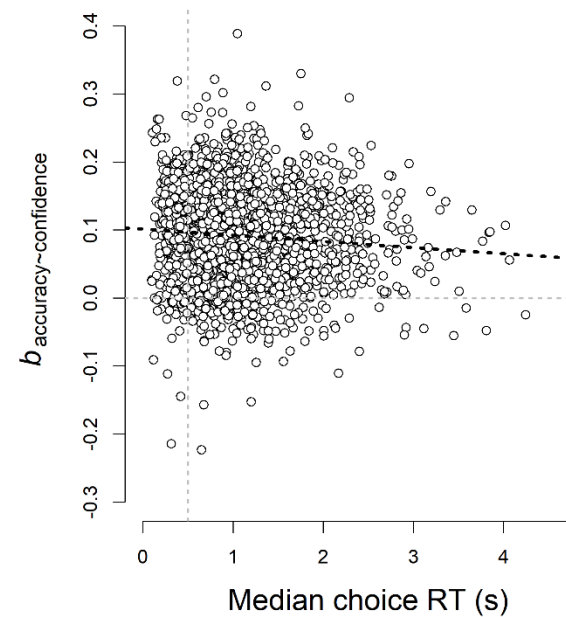
B.



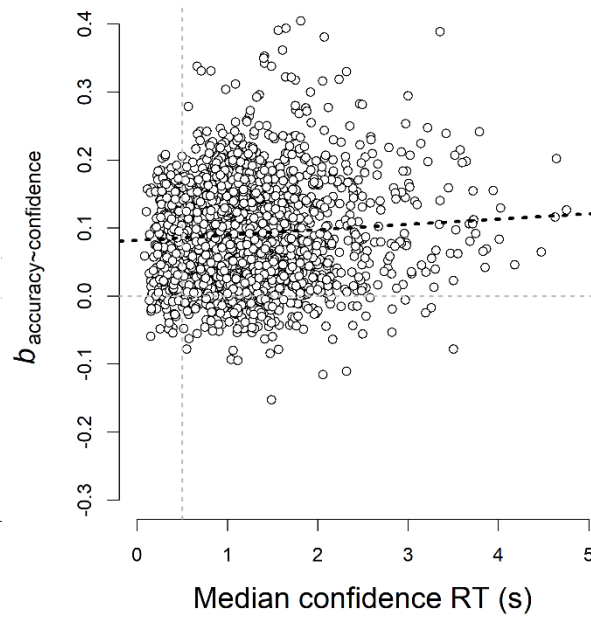
C.



D.



E.



F.

