Psychonomic Bulletin & Review (2022) 29:1405–1413
https://doi.org/10.3758/s13423-022-02063-7

BRIEF REPORT

Across-subject correlation between confidence and accuracy: A meta-analysis of the Confidence Database

Sunny Jin[1] · **Paul Verhaeghen**[1] · **Dobromir Rahnev**[1]

Accepted: 22 January 2022 / Published online: 7 February 2022

Abstract
If one friend confidently tells us to buy Product A while another friend thinks that Product B is better but is not confident, we may go with the advice of our confident friend. Should we? The relationship between people's confidence and accuracy has been of great interest in many fields, especially in high-stakes situations like eyewitness testimony. However, there is still little consensus about how much we should trust someone's overall confidence level. Here, we examine the across-subject relationship between average accuracy and average confidence in 213 unique datasets from the Confidence Database. This approach allows us to empirically address this issue with unprecedented statistical power and check for the presence of various moderators. We find an across-subject correlation between average accuracy and average confidence of $R = .22$. Importantly, this relationship is much stronger for memory than for perception tasks ("domain effect"), as well as for confidence scales with fewer points ("granularity effect"). These results show that we should take one's confidence seriously (and perhaps buy Product A) and suggest several factors that moderate the relative consistency of how people make confidence judgments.

Keywords Confidence · Metacognition · Perceptual decision making · Memory · calibration · Bias

Introduction

Expressing an appropriate level of confidence is of utmost importance in many facets of human life. From the individual's perspective, one's confidence helps determine whether to commit to a decision or gather more information (Desender et al., 2018) and possibly seek advice from others (Pescetelli & Yeung, 2021). Our confidence also helps us determine whether we have sufficient expertise in a particular domain or if more learning is required (Dautriche et al., 2021).

Critically, confidence is equally essential in the interpersonal domain. For example, the confidence level in a diagnosis expressed by medical personnel can substantially affect how patients receive the diagnosis (Yang & Thompson, 2010). Similarly, the confidence expressed in eyewitness testimonies can have an immense impact on the outcome of a trial (Tenney et al., 2008). Perhaps due to the enormous stakes of such testimonies, there is a lively debate on how trustworthy confidence ratings of eyewitnesses are (Berkowitz et al., 2020; Juslin et al., 1996; Loftus & Greenspan, 2017; Penrod & Cutler, 1995; Wixted & Wells, 2017). As these examples demonstrate, understanding how much we can trust the confidence expressed by others has important implications for many domains as disparate as law, medicine, and education.

Yet the question of how the confidence expressed by one person relates to the confidence expressed by another remains ill-understood. For example, imagine a situation where one friend confidently tells us to buy Product A while another friend thinks that Product B is better but is not confident. How should we act? Can we meaningfully compare the confidence of two different people? Addressing this fundamental question requires answering at least two separate questions: (1) How strong is the across-subject relationship between confidence and accuracy, and (2) what factors moderate the strength of this relationship.

The first question regarding the strength of the across-subject relationship between confidence and accuracy has been primarily studied by eliciting a single or only a few decisions from each participant. While this approach closely mimics many real-world situations (e.g., eyewitness testimony), it sheds virtually no light on the question of how well

✉ Dobromir Rahnev
rahnev@psych.gatech.edu

[1] School of Psychology, Georgia Institute of Technology, 654 Cherry Str. NW, Atlanta, GA 30332, USA

people's overall confidence reflects their overall accuracy on a task that involves many decisions. This is because a single decision is either correct or incorrect, but, in reality, people vary on their underlying performance ability continuously (e.g., 70% vs. 80% accurate on a two-choice task). Therefore, the question of how much we should trust confident vs. nonconfident people should also be addressed by examining situations where each participant completes many trials that involve both a primary decision and a confidence judgment so that their overall accuracy and confidence can be assessed.

The second question regarding the factors that moderate the across-subject confidence–accuracy relationship has also remained largely unexplored. The reason is that few studies to date have been able to compare data that vary in critical dimensions such as the domain of study (e.g., memory or perception) or how confidence is elicited (e.g., whether it is given with or after the primary decision). Therefore, we know virtually nothing about the factors that increase or decrease the across-person relationship between confidence and accuracy.

Here, we addressed both of these questions with unprecedented statistical power by taking advantage of the recently published Confidence Database (Rahnev et al., 2020). This database includes many datasets featuring confidence ratings derived primarily from traditional laboratory studies where each participant completes hundreds of trials. Importantly, the datasets vary in many aspects, including the domain of study (e.g., perception or memory), the confidence scale used, and the presence of trial-by-trial feedback. This variability allowed us to examine further the moderators that make overall confidence more or less predictive of overall accuracy. To anticipate, we found that average confidence and average accuracy correlate at $R = .22$, with higher correlations emerging in memory studies ("domain effect") and with confidence scales that have fewer points ("granularity effect").

Methods

Background on the Confidence Database

The data for analysis was taken from the Confidence Database (Rahnev et al., 2020). The Confidence Database is a large collection of datasets that include confidence ratings (available at https://osf.io/s46pr/), with the data being provided by dozens of labs across 16 countries on five continents. Each dataset contains the data from a separate experiment. Most of these experiments represent standard lab-based studies, though several datasets come from data collected online.

The complete database was downloaded in October 2019 (note that several more datasets have been added since then) and featured 145 separate datasets. The individual datasets vary in many dimensions (Fig. 1). First, the individual studies come from several domains, including perceptual, memory, cognitive, and motor tasks. Most perception tasks involve discrimination between two stimulus categories: left-tilted vs. right-tilted gratings or dot motion that moves to the left vs. right. Most memory tasks involve the standard procedure of studying a list of items (e.g., words, pictures) and providing old/new judgments with confidence during a later recognition test. However, while these standard designs account for most datasets, many specific datasets feature unique designs that depart from the typical tasks described above. Second, the individual datasets vary considerably in their sizes, with the number of participants ranging from 4 to 589 (mean = 42.9, $SD = 76.1$) and the number of trials per participant ranging from 20 to 5,307 (mean = 507.8, $SD = 760.9$). Third, the datasets vary on the scale used for collecting confidence ratings, with 3-point, 4-point, and continuous scales being the most common. Finally, the datasets vary on many other dimensions, including whether confidence was collected simultaneously or after the primary decision, whether trial-by-trial feedback was provided, and whether the task required participants to discriminate between several options or to estimate a continuous quantity (e.g., the length of a time interval).

Most studies elicited a confidence rating by asking a general question such as "How confident are you?" Some studies only labeled the extremes of the confidence scale (e.g., "not confident at all" and "extremely confident"). In contrast, others labeled each of the provided options (e.g., "not confident," "somewhat confident," "very confident"). As is customary for lab-based research, participants typically received training on the specific task during a few practice blocks. The data from these practice blocks were usually not included in the provided datasets.

One desirable feature of the studies in the Confidence Database is that none of these studies were collected with the main purpose of identifying the correlation between average confidence and average accuracy. Indeed, each dataset in the Confidence Database features a complementary text file with information that includes fields for both the "Experiment goal" and "Main result." Examining these fields confirmed that investigating the correlation between average accuracy and average confidence (or related analysis) was never listed as the primary goal or result for any dataset.

It should be noted that while the studies represented in the Confidence Database represent a large and diverse slice of the literature, they only make up a small fraction of all studies that collected confidence ratings (which may number in the hundreds of thousands). As such, the current study did not attempt to comprehensively cover the relevant literature,

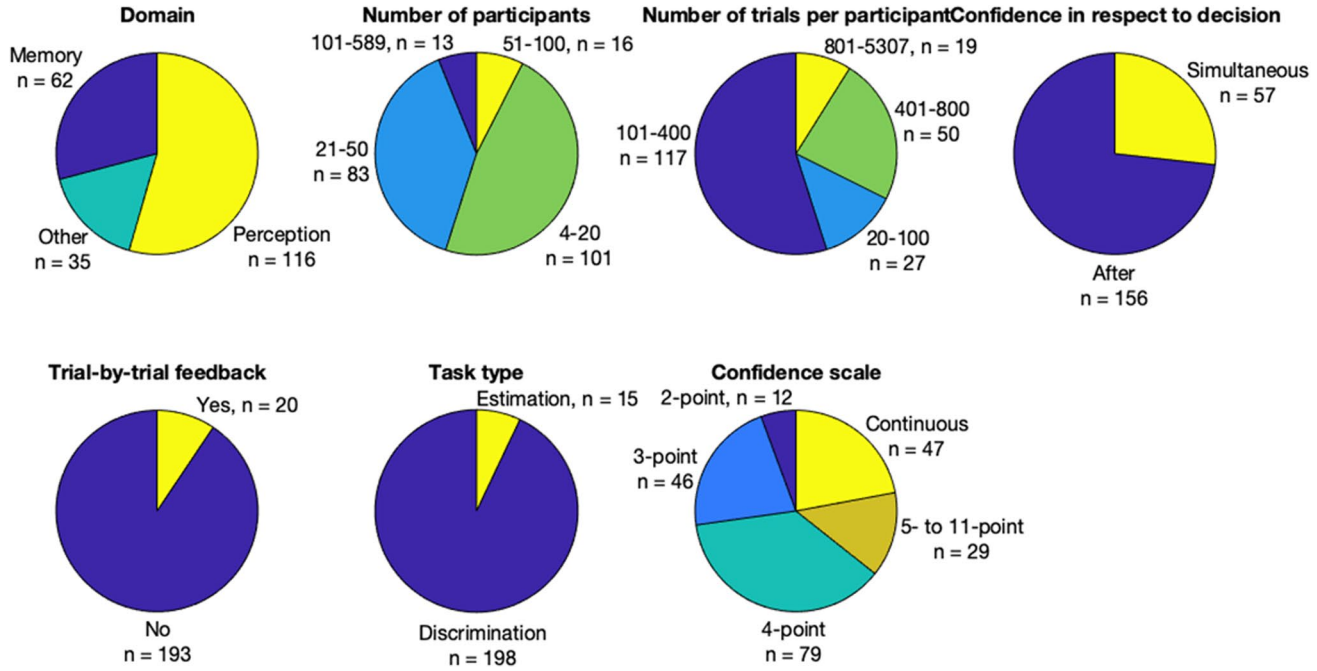

Fig. 1 Dataset information. Pie charts with details regarding the different datasets. The "Other" domain includes cognitive and motor tasks. The vast majority of datasets were relatively small, with an average size of 42.9 participants. On the other hand, each participant performed many trials (average = 507.8 trials per participant). Most datasets collected confidence after the decision, did not provide trial-by-trial feedback, and employed discrimination tasks (where participants chose an answer among, typically, two alternatives). There was a lot of variability in the confidence scale used, with 3-point, 4-point, and continuous scales being the most frequent

unlike many previous meta-analyses in the literature. Therefore, we emphasize that the current study constitutes a meta-analysis of the Confidence Database specifically and does not cover the broader literature.

Data selection

The complete database was downloaded in October 2019 (note that several more datasets have been added to it after that date) and featured 145 separate datasets. We excluded three of the original datasets. The first dataset ("Dildine_unpub") was excluded because it used a task where objective performance could not be computed. The reason was that participants rated their subjective perception of pain, but there was no "objective" pain level that they should have reported, thus making it impossible to compute participants' accuracy on the task. The second dataset ("Duyan_unpub_logos") was excluded because it used a task that produced chance-level performance in all participants making any variations in accuracy levels across subjects essentially random. Finally, the third dataset ("Zylberberg_2016") was excluded because it had only three participants (for the purposes of conducting a meta-analysis, we required that each dataset has at least four participants; see below). All analyses were thus performed on the remaining 142 original datasets.

Several datasets included the data from the initial training that includes practice and often stimulus calibration prior to the main experiment, but most datasets did not include these data. Therefore, for consistency, we excluded the training data from all datasets. Due to the vast differences between the tasks used in different datasets, it was impossible to devise standardized exclusion criteria for individual participants. Therefore, all participants from each dataset were included in the analyses.

Data preprocessing

While most datasets featured a single task completed by all participants, this was not universally true. In fact, several datasets featured different tasks (or conditions), with each task completed by a separate subset of participants. However, task differences between participants could potentially bias the relationship between average accuracy and average confidence across the whole group. Therefore, to avoid such biases, we manually examined each dataset and split each original dataset that features multiple tasks (or conditions) completed by separate subsets of participants into independent datasets. Subsequently, we use the term "original dataset" to refer to the original datasets found in the Confidence Database and the term "dataset" to refer to the datasets obtained after the splitting process. This process identified

34 original datasets with multiple tasks (19 with two tasks each, seven with three tasks each, three with four tasks each, three with six tasks each, one with eight tasks, and one with 10 tasks). Details on these original datasets and how they were split are provided in the Supplementary Methods. This process resulted in a total of 215 datasets. However, two of these new datasets only included three participants and were therefore excluded, thus leaving us with a total of 213 individual datasets for all remaining analyses.

Analyses

For each of the 213 final datasets, we computed each participant's average accuracy and confidence. Two hundred out of the 213 datasets featured discrimination tasks where participants chose one response among several (typically two) options. For such cases, accuracy was computed based on whether a response was the same as the correct answer. The remaining 14 datasets, however, featured estimation tasks where, for example, participants had to give a numerical guess regarding the duration or orientation of a continuous stimulus. In such cases, accuracy was computed as the deviation between the response and the correct answer. However, since low deviation values correspond to better performance, for these datasets, "accuracy" was defined as the deviation between the response and the correct answer multiplied by −1, thus ensuring that higher accuracy corresponds to higher performance.

Once the average accuracy and average confidence were computed for each participant in a dataset, we performed a Pearson correlation between these two quantities. We then z-transformed the resulting R-values before conducting further statistical analyses (though, for display purposes, we plot the original R-values in all figures).

The primary analysis consisted of a meta-analysis across all 213 z-transformed R-values. The meta-analysis weighted each z-value based on its variance, z_{var}, which is equal to $1/(n-3)$, where n is the number of participants. This value is only defined for $n \geq 4$, which necessitated the exclusion of all datasets with $n \leq 3$.

Additional analyses investigated whether different experimental factors moderated the relationship between average confidence and average accuracy. Based on the information provided as part of the Confidence Database, we were able to identify six factors for which we could find relevant information for every dataset. The first factor was the granularity of the confidence scale used. Confidence scales had 2, 3, 4, 5, 6, 9, 11 points or were continuous. Therefore, in separate analyses, we compared the continuous scales against all others, as well as the scales with five or more points against the scales with four or fewer points. The second factor was the domain of study—that is, whether the study used a memory task, a perception task, or "other." The last category featured 35 datasets from three different types of tasks: cognitive (21 datasets), mixed (that is, datasets that include tasks from multiple domains, 11 datasets), and motor (3 datasets). We combined these different categories because we did not have sufficient power to examine them separately. Nonetheless, in control analyses, we further split the "other" category into its constituent parts, but that did not affect any of the main results or reveal significant effects associated with the smaller categories. The third factor was the timing of the confidence judgments, that is, whether the confidence rating was given simultaneously with the decision or after the decision. The fourth factor was the presence of trial-by-trial feedback (present or absent). The fifth factor was the type of task (discrimination or estimation). Finally, the sixth factor was the average number of trials per participant. We performed a mixed-effects meta-analysis with all six factors, using the *metafor* package in R (Viechtbauer, 2010). *Metafor* is currently one of the most popular packages for meta-analysis; it includes functions for fitting fixed-effects, random-effects, and mixed models and allows for the inclusion of moderator variables in these models (Lortie & Filazzola, 2020; Viechtbauer, 2010).

We assessed the inter-study variation in effect sizes via the Q (Hedges & Olkin, 1985) and I^2 (Cooper, 2017) statistics. We further checked for publication bias by performing a standard Egger's regression test for funnel plot asymmetry (Egger et al., 1997).

Data and code

All data and codes for preprocessing and analysis are available at https://osf.io/kpe75/.

Results

We investigated the across-subject correlation between average confidence and accuracy using 213 separate datasets extracted from the Confidence Database (Rahnev et al., 2020). Details regarding the datasets are available in Fig. 1. There were 9,132 total participants and 3,896,543 total trials used for estimating the correlations.

We first computed the meta-analytic average correlation across all 213 datasets. We found a significant correlation of small to moderate size ($R = .22$, $p < .0001$, 95% CI [.18, .27]). Further, we found heterogeneity among the datasets ($I^2 = 67.83\%$) that was significantly above chance (Q (212) $= 649.62$, $p < .0001$), suggesting that one or several moderators may further determine the strength of the correlation between average accuracy and average confidence. Repeating the analyses by ignoring sample size (via a simple t-test) produced very similar results ($R = .20$, $t(212) = 9.24$, $p = 2.7\times10^{-17}$, Cohen's $d = .63$). An additional meta-analysis

performed only on the 113 datasets with a sample size $n > 20$ also produced very similar results ($R = .21$, $p < .0001$, 95% CI [.16, .26]).

To our knowledge, none of the data in the Confidence Database were collected with the main purpose of identifying the correlation between average confidence and average accuracy. Therefore, we expected little to no publication bias for the datasets in the current study. Indeed, Fig. 2 reveals no clear asymmetry, which would indicate publication bias. Egger's regression test for funnel plot asymmetry confirmed the lack of asymmetry ($z = .49$, $p = .62$).

Having established the existence of a significant correlation between average accuracy and average confidence, we examined whether this relationship depends on any moderators. To this end, we performed a mixed-effects meta-analysis with the following factors: the granularity of the confidence scale used (discrete vs. continuous), the domain of study (perception vs. memory vs. other), the timing of the confidence judgments (with vs. after the decision), the presence of trial-by-trial feedback (present vs. absent), the type of task (discrimination vs. estimation), and the average number of trials per participant.

We found a significant effect of the granularity of the confidence scale with less granular scales (i.e., scales with fewer points), resulting in higher R-values (Fig. 3a and Table 1). We call this the "granularity effect." Specifically, continuous scales produced R-values that were about three times lower compared with discrete scales ($R_{\text{continuous}} = .08$, $R_{\text{discrete}} = .26$; $z = 2.60$, $p = .0092$). Since there were many fewer continuous than discrete confidence scales, we performed a control analysis that minimized the difference in the size of both sets of studies. To do so, we compared the datasets with 2-, 3-, and 4-points confidence scales to the datasets with scales with at least 5 points and still found the same results ($R_{\text{2-to-4-point scale}} = .12$, $R_{\text{5+-point scale}} = .28$; $z = 2.92$, $p = .0035$). Further, a non-parametric Spearman correlation revealed a significantly negative relationship between the number of points on the confidence scale and R-values (rho $= -.21$, $p = .002$). These results suggest that less granular confidence scales increase the correlation between average accuracy and average confidence.

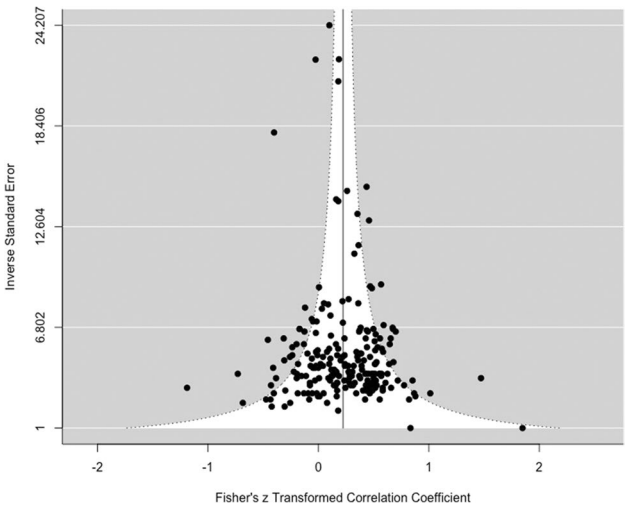

Fig. 2 Funnel plot. The z-transformed strength of the correlation between average accuracy and average confidence is plotted against the inversed standard error. Each circle is a separate dataset

In addition, we found that the domain of study is also a strong moderator of the strength of the correlation between average accuracy and average confidence (Fig. 3b; we call this "domain effect"). Specifically, perception studies produced the lowest R-value ($R = .13$), which was significantly lower than in memory studies ($R = .35$, $z = 3.04$, $p = .002$) and marginally lower than in studies from other domains ($R = .25$, $z = 1.80$, $p = .07$). Studies from memory and other domains were not significantly different ($z = -1.11$, $p = .27$).

No other moderator was found to be significant including the timing of the confidence judgment (with or after the primary decision, $z = .79$, $p = .43$; Fig. 3c), the existence of trial-by-trial feedback ($z = 1.19$, $p = .24$; Fig. 3d), the type of task (discrimination or estimation, $z = 1.32$, $p = .19$; Fig. 3e), or the mean number of trials per participant in a dataset ($z = .60$, $p = .55$).

Discussion

We investigated how much one should trust the judgment of a confident person compared with that of someone who is not confident. To address this question, we performed a meta-analysis on 213 datasets derived from the Confidence Database. We computed the average accuracy and average confidence for each participant for each dataset and then correlated these values across subjects. We found that this correlation was significantly positive and was of moderate size ($R = .22$). Further, the strength of the relationship was moderated by both the domain of the study and the granularity of the confidence scale used. These results begin to reveal not just the overall strength of the across-subject relationship between accuracy and confidence but also the moderators of this relationship.

It is important to appreciate precisely what drives the correlation between average accuracy and average confidence. This correlation reflects the *relative calibration* of confidence ratings across participants. In other words, it shows whether people's confidence ratings are well-calibrated relative to each other and is largely unrelated to the metacognitive ability of the individual participants. For example, it is possible to achieve high relative calibration even if there's a strong over- or under-confidence bias in

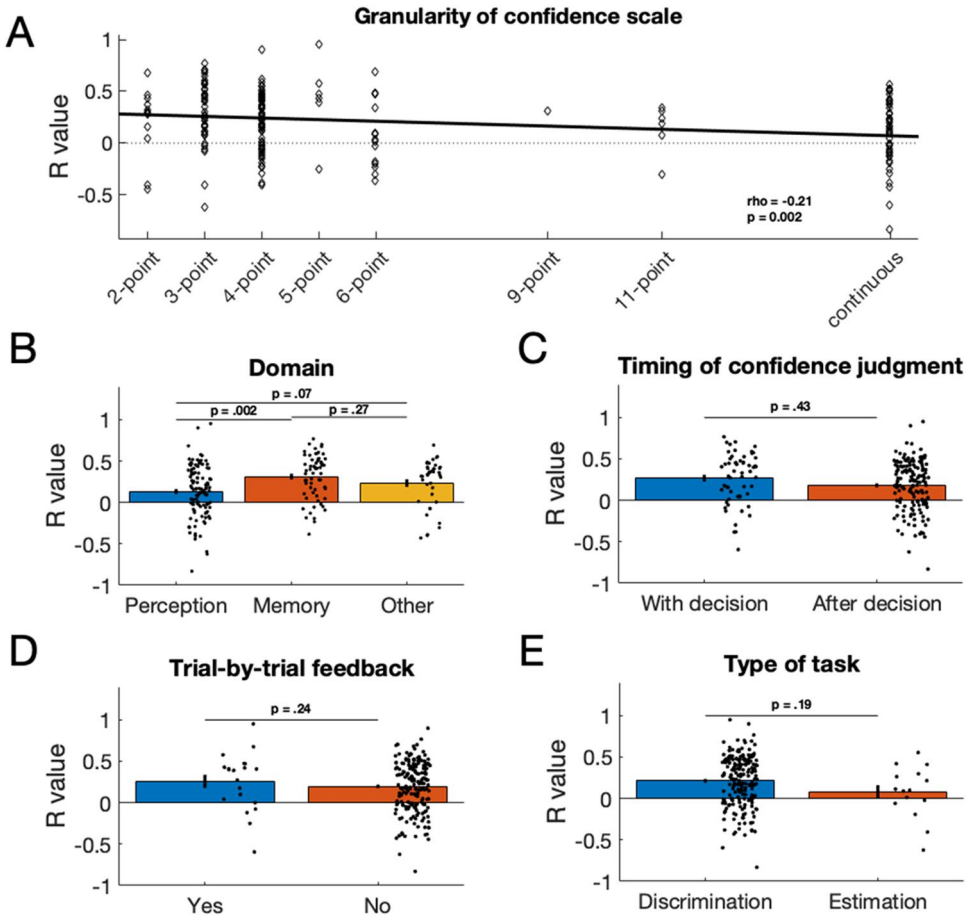

Fig. 3 Moderators of the relationship between average accuracy and average confidence. The strength of the correlation (*R*-value) between average accuracy and average confidence was (**a**) higher for less granular confidence scales (that is, scales with fewer points), (**b**) highest for memory tasks and lowest for perception tasks, and not significantly different for datasets that differed on (**c**) the timing of the confidence judgment, (**d**) the presence of trial-by-trial feedback, and (**e**) the type of task. Error bars show *SEM*, and each diamond/circle represents one dataset. The *p*-values are derived from a (non-meta-analytic) Spearman correlation in (**a**) and mixed-effects meta-analyses in (**b–e**)

the whole group as long as this bias is consistent across participants (Fleming & Lau, 2014). Conversely, within-subject confidence ratings may be maximally informative (i.e., they could show no metacognitive inefficiency; Shekhar & Rahnev, 2021a, 2021b), but different over- or under-confidence bias for different people could still result in low across-subject correlations. Therefore, the *R*-values examined here should not be interpreted as showing "how informative confidence is" in general, but rather how well people's confidence ratings are calibrated relative to other people in the group. Returning to our example from the Abstract, if two friends give us conflicting advice, going with the advice of the more confident friend relies on the assumption that the confidence ratings expressed by these two friends are well-calibrated relative to each other. Our findings indeed confirm that, in the absence of additional information, following the advice of the more confident friend is the best strategy.

A related question is how to judge the magnitude of the meta-analytic correlation of $R = .22$. Traditionally, this correlation would be considered to be somewhere between small ($R = .1$) and medium ($R = .3$; Cohen, 1988). However, this characterization has been criticized as nonsensical and has been reportedly disavowed by Cohen himself (Funder & Ozer, 2019). Instead, Funder and Ozer proposed a classification according to which $R = .2$ "indicates a medium effect that is of some explanatory and practical use even in the short run," $R = .3$ "indicates a large effect that is potentially powerful in both the short and the long run," and $R > .4$ indicates a "very large effect size" that, in the context of psychological research, is likely to be an overestimation. They argue that this classification matches intuitively understood correlations such as the effectiveness of antihistamines on sneezing ($R = .14$), the higher weight of men compared with that of women ($R = .26$), and the lower average annual temperatures at higher elevations ($R = .34$). In this context,

Table 1 Moderators of the relationship between average accuracy and average confidence

	k	*R*	Lower limit of 95% CI	Upper limit of 95% CI
Overall*	213	.22*	.18	.26
Domain				
Memory[a]	62	.35*	.28	.42
Perception[b]	116	.13*	.06	.20
Other[a]	35	.25*	.14	.35
Trial-by-trial feedback				
Yes[a]	20	.26*	.09	.42
No[a]	193	.22*	.17	.26
Timing of confidence judgment				
With decision[a]	57	.31*	.22	.40
After decision[a]	156	.19*	.14	.23
Granularity of confidence scale				
Discrete[a]	166	.26*	.22	.30
Continuous[b]	47	.08	-.03	.19
Type of task				
Discrimination[a]	198	.23*	.19	.27
Estimation[a]	15	.09	-.10	.28

Note. Asterisks denote significant effects at $p < .05$. Average effect sizes with different superscripts (e.g., "a" and "b") are significantly different from each other (except for the comparison between Perception and Other, which is only marginally significant). The bolded moderators are significant; *k* = number of datasets; CI = confidence interval

the observation of $R = .22$ should be thought of as a medium effect, while the correlation of $R = .35$ for memory studies should be seen somewhere between a large and a very large effect size. In other words, our results should be interpreted as showing overall good relative calibration of confidence that becomes especially strong in memory tasks.

The current paper investigated the relationship between average accuracy and average confidence obtained over the course of the same task. Nevertheless, our results align with the extensive literature on the relationship between objective and subjective ability estimates. For example, a meta-synthesis of 22 published meta-analyses of the relationship between self-evaluations and objective performance measures found an average correlation of $R = .29$, with 18 meta-analyses reporting a correlation between .19 and .39 (Zell & Krizan, 2014). Similarly, there has been a lot of work on the reliability of confidence judgments in eyewitness testimony, with one meta-analysis (Sporer et al., 1995) finding an average correlation between the accuracy and confidence of identification of $R = .29$. Therefore, it appears that the correlation between accuracy on a laboratory task and average confidence obtained on a trial-by-trial basis in the same task is largely consistent with the relationship between objective and subjective performance across a variety of fields.

Our meta-analysis of the Confidence Database revealed two significant moderators of the relationship between average accuracy and average confidence. First, we observed a domain effect such that the correlation between confidence and accuracy was about three times stronger for memory studies ($R = .35$) than perception studies ($R = .13$). We suspect that this difference may reflect participants' familiarity with the two types of tasks. Indeed, most people have plenty of experience with communicating their certainty in faint and unreliable memories, but few have sufficient experience with expressing confidence for perceptual stimuli near the psychophysical threshold. As such, it is natural for people's confidence to be well-calibrated relative to other people's confidence for memory but not perception tasks.

Second, we observed a granularity effect such that the granularity of the confidence scale was a significant moderator of the across-subject confidence–accuracy relationship. Specifically, we found that less granular confidence scales (that is, scales with few options) resulted in higher confidence–accuracy correlations than more granular scales (that is, scales with many options or continuous responses). The difference was substantial, with continuous scales producing over three times lower correlation ($R = .08$) than discrete scales ($R = .26$). Interestingly, a qualitatively similar effect appeared when comparing estimation tasks (where the primary response is continuous, $R = .09$) with discrimination tasks (where participants choose from several discrete options, $R = .23$), though this latter effect was not statistically significant. One may at first think the granularity effect (in both the confidence scale and type of task) is simply a statistical artifact that arises from performing Pearson correlation on continuous vs. discrete quantities. However, it should be emphasized that we correlated *average* confidence with *average* accuracy; both of these quantities are essentially continuous regardless of the granularity of the confidence scale on individual trials. Instead, we favor a different interpretation. As we already emphasized, the across-subject strength of the correlation between confidence and accuracy primarily reflects the similarity among participants' confidence biases. Specifically, similar biases across all participants are likely to result in high correlations, whereas divergent biases across participants are likely to result in low correlations. Therefore, the low correlations observed for continuous confidence scales could stem from such scales being interpreted differently by different people, thus creating more divergent biases. Conversely, a binary low/high confidence scale may result in greater consistency across subjects, increasing the size of the across-subject correlations. Nevertheless, this interpretation remains speculative, and more direct tests are needed to confirm it.

We failed to find other significant moderators of the relationship between average accuracy and average confidence. Nevertheless, it should be noted that we had relatively little power for three of the moderators: the presence of trial-by-trial feedback, the type of task, and the timing of the confidence judgment. There were numerical differences in the *R*-values in all three cases with higher correlation coefficients observed in the presence of trial-by-trial feedback, discrimination (as opposed to estimation) tasks, and confidence ratings given simultaneously with (as opposed to after) the decision. The numerical difference in the context of trial-by-trial feedback is consistent with the findings of a recent large study that found a significant increase in the confidence-accuracy correlation in the presence of trial-by-trial feedback (Haddara & Rahnev, 2022). It is, therefore, possible that some of the factors that were not significant in the current study may nonetheless be important for the size of the correlation between average accuracy and average confidence. In other words, the present study cannot confirm or disprove the possibility that these variables moderate the size of the across-subject confidence–accuracy correlation, and the current null results should be interpreted with caution.

An important limitation of the current study is that we used a convenience sample of studies deposited in the Confidence Database that may not accurately represent the overall literature. Although our funnel plot analyses showed no evidence of publication bias, it remains possible that at least some of our results are driven by a selection bias.

In conclusion, we performed a meta-analysis of the Confidence Database on the relationship between average accuracy and average confidence in standard laboratory tasks. We found a moderately strong relationship ($R = .22$) moderated by the domain of the study and the granularity of the confidence scale. These findings provide important clues regarding how people interpret confidence ratings and how consistent these interpretations are across participants.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-022-02063-7.

Acknowledgments This work was supported by the National Institute of Health (award: R01MH119189) and the Office of Naval Research (award: N00014-20-1-2622).

References

Berkowitz, S. R., Garrett, B. L., Fenn, K. M., & Loftus, E. F. (2020). Convicting with confidence? Why we should not over-rely on eyewitness confidence. *Memory*. https://doi.org/10.1080/09658211.2020.1849308

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). SAGE.

Dautriche, I., Rabagliati, H., & Smith, K. (2021). Subjective confidence influences word learning in a cross-situational statistical learning task. *Journal of Memory and Language*, 121, 104277. https://doi.org/10.1016/j.jml.2021.104277

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

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

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168. https://doi.org/10.1177/2515245919847202

Haddara, N., & Rahnev, D. (2022). The impact of feedback on perceptual decision making and metacognition: Reduction in bias but no change in sensitivity. *Psychological Science*. (in press)

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*(5), 1304–1316. https://doi.org/10.1037/0278-7393.22.5.1304

Loftus, E. F., & Greenspan, R. L. (2017). If I'm Certain, Is It True? Accuracy and Confidence in Eyewitness Memory. *Psychological Science in the Public Interest*, *18*(1), 1–2. https://doi.org/10.1177/1529100617699241

Lortie, C. J., & Filazzola, A. (2020). A contrast of meta and metafor packages for meta-analyses in R. *Ecology and Evolution*, *10*(20), 10916–10921. https://doi.org/10.1002/ece3.6747

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*(4), 817–845. https://doi.org/10.1037/1076-8971.1.4.817

Pescetelli, N., & Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General*, *150*(3), 507–526. https://doi.org/10.1037/xge0000960

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., ... Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Shekhar, M., & Rahnev, D. (2021a). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Shekhar, M., & Rahnev, D. (2021b). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45–70. https://doi.org/10.1037/rev0000249

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*(3), 315–327. https://doi.org/10.1037/0033-2909.118.3.315

Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, *44*(5), 1368–1375. https://doi.org/10.1016/j.jesp.2008.04.006

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(1), 1–48. https://doi.org/10.18637/JSS.V036.I03

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Yang, H., & Thompson, C. (2010). Nurses' risk assessment judgements: A confidence calibration study. *Journal of Advanced Nursing*, *66*(12), 2751–2760. https://doi.org/10.1111/j.1365-2648.2010.05437.x

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. https://doi.org/10.1177/1745691613518075

Open practices statement The data and materials for all experiments are available (https://osf.io/kpe75/). The analyses performed were not preregistered .