Investigating the experimental contexts in which people with high confidence have high accuracy

Senior Thesis Presented to
The Academic Faculty

by
Sunny Jin

Research Option for
the Neuroscience Degree in the
College of Sciences

Approved by:

Dr. Dobromir Rahnev, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Paul Verhaeghen
School of Psychology
Georgia Institute of Technology

Dr. Mary Holder
School of Neuroscience
Georgia Institute of Technology

TABLE OF CONTENTS

ABSTRACT

Metacognitive ability describes one's ability to discern the accuracy of one's previous decisions by attributing high confidence values to correct decisions and low values for incorrect decisions. This study aimed to determine the experimental conditions under which confidence-accuracy correlations are the strongest, indicating an overall greater metacognitive ability of subjects testing under those conditions than in others. To this end, this study investigated confidence-accuracy correlation across subjects for 143 cognitive neuroscience experiments from the Confidence Database on Open Science Framework. Using their respective across-subject correlations, this study determined for these experiments whether their correlation strengths differed between each other by their unique experimental design characteristics. This was done in an effort to investigate how these characteristics influence the strength of the correlation obtained and thus subjects' metacognition. A significant, positive mean correlation was found from all subsets, following the general trend in confidence-accuracy correlation. It was also found that correlations between experiments of different categories, specifically perception and memory, are significantly different, with mixed-type experiments having the highest correlations. There was a significant effect of confidence range on confidence-accuracy correlation, but no significant effect of feedback, number of subjects, minimum trials per subject, maximum trials per subject, number of tasks times conditions, or number of difficulty levels. Future studies are needed to further investigate the effects of the design characteristics for which this study could not find a significant difference. By finding the right combinations of design characteristics for good metacognition, these combinations could be translated and applied to real-world settings in which high confidence-accuracy calibration is desired.

INTRODUCTION

Metacognitive ability describes one's ability to discern the accuracy of one's previous decisions by attributing high confidence values to correct decisions and low values for incorrect decisions. In subjects with good metacognitive ability, accuracy and confidence values should increase and decrease together (Sanders, Hangya, & Kepecs, 2016). Confidence calibration describes this relationship and thus serves as a measure of metacognitive ability. It assesses the extent to which accuracy values and confidence values are proportional to each other: 100% of the decisions with 100 confidence (0-100 scale) should be correct, 90% of the decisions with 90 confidence should be correct, 80% for 80 confidence, and so on (Luna & Martín-Luengo, 2012). In a perfectly calibrated human subject, confidence and accuracy values should be equal and linearly increase together, modeled by a line of best fit with a slope of one (Bjorkman, Juslin, & Winman, 1993). A related measure of metacognitive ability is confidence-accuracy correlation, with higher positive correlations between accuracy and confidence demonstrating greater ability.

Confidence calibration and related measures have relevant, real-world implications in many societal domains, such as determining social influence in team decisions, the reduction of errors in medical diagnosis, and the believability of eye-witness testimonies in court (Zarnoth & Sniezek, 1997; Yang & Thompson, 2010; Tenney, Spellman, & MacCoun, 2008). In team decisions, social influence is augmented in intellectual tasks when an individual's accuracy and confidence levels align, giving him or her more leverage in the group decision-making process (Zarnoth & Sniezek, 1997). In a clinical setting, experienced nurses tend to be overconfident in the accuracy of their diagnoses, hinting at the possibility of overconfidence with inaccuracy fueling unaddressed medical errors (Yang & Thompson, 2010). When witnesses' calibration values are not disclosed to subjects, eye-witness testimonies of highly confident witnesses are

favored by subjects over less confident witnesses, even if both accounts are incorrect (Tenney, Spellman, & MacCoun, 2008). These applications show that understanding calibration is important for the assignment of credibility in situations where correct judgment is sought.

Although trusted to be an indicator of accuracy, confidence may not always correspond to one's accuracy level. Deviations from the line of perfect calibration, with the y-axis as accuracy and x-axis as confidence, can be seen when the data points fall above the line (underconfidence) or below (overconfidence). Underconfidence describes lower confidence levels than would be appropriate for a subject's accuracy, and overconfidence with higher levels than would be appropriate. Many different experimental factors may affect the reported confidence calibration, such as task difficulty, feedback, task type, and confidence reporting method used. In an effort to study confidence calibration, many studies have employed sensory discrimination, general knowledge, and memory-recall tasks in order to research how confidence and accuracy relate in decision-making. Task difficulty is found to influence confidence calibration, with subjects exhibiting the "hard-easy effect," in which they were overconfident in hard trials and underconfident in easy trials (Lichtenstein & Fischhoff, 1977; Baranski & Petrusic, 1999). Decision accuracy feedback is shown to help subjects get closer to perfect calibration in a globally hard task (more hard trials than easy) but further in the globally easy task (Petrusic & Baranski, 1997). Confidence is better calibrated in general knowledge than in eye-witness memory-recall tasks (Luna & Martín-Luengo, 2012). One study suggests a pervasive underconfidence bias in sensory discrimination tasks as an inherent result of the configuration of the sensory system (Bjorkman, Juslin, & Winman, 1993). Out of three main confidence reporting techniques (perceptual awareness scale, confidence rating, and post-decision wager), perceptual awareness scale was associated with the highest performance-awareness correlation (Sandberg,

Timmermans, Overgaard, & Cleeremans, 2010). On the other hand, subjects reporting with post-decision wagers were susceptible to loss aversion strategies that result in altered reports deviating from internal confidence levels (Fleming & Dolan, 2010). The findings of these previous studies suggest that confidence-accuracy correlation may be influenced by experimental task type and design choices, such as whether the task is perceptual or memory-related, or the reporting technique chosen to record confidence values.

Previous literature highlights the significance and relevance of confidence calibration to the real world, across many societal domains, such as medicine, politics, law, and industry. It also reveals the susceptibility of confidence calibration being altered or influenced by experimental design characteristics. Building upon this existing literature, this study investigated the effects of other experimental design choices, such as number of conditions/manipulations, on the related calibration measure of confidence-accuracy correlation across subjects. This study aimed to determine the experimental design conditions under which confidence-accuracy correlations are the strongest, indicating an overall greater metacognitive ability of subjects testing under those conditions than in others. To this end, this study investigated confidence-accuracy correlation across subjects, which was determined by finding the mean accuracy and mean confidence for a single subject from all trials completed by the subject, doing so for all subjects, and correlating the mean values of all the subjects together to extract a single confidence-accuracy correlation per experimental subset (see “Methods”).

Using their respective across-subject correlations, this study determined for many experiments whether their correlation strengths differed between each other by experimental characteristics, such as number of tasks and conditions, experimental task type, and number of difficulty levels. This was done in an effort to investigate how these characteristics affected the

strength of the correlation obtained. Confidence-accuracy correlation values were extracted using Python data querying for 143 cognitive neuroscience experiments from the Confidence Database on Open Science Framework and statistically analyzed by experimental characteristics (including those mentioned above). Statistical analysis was used to detect any significant differences between all 143 correlations by a single experimental characteristic, such as task type (e.g. assessing whether correlations from memory studies significantly differ from those from perception studies). All correlations were reanalyzed by each characteristic, until all experimental characteristics had been tested. This study aimed to provide more information about how other design characteristics, besides task type and confidence recording method, impact the strength of the confidence-accuracy relationship. Additionally, by looking at how experimental design characteristics may influence resulting correlations, this study aimed to address the wider implication that experimental results may be reached and altered by modifying the experimental setup.

After conducting this experiment towards these aims, this study found a significant, positive mean correlation from all subsets, following the general trend in confidence-accuracy correlation. Correlations between experiments of different categories, specifically perception and memory, were also found to be significantly different, with mixed-type experiments having the highest correlations. There was a significant effect of confidence range on confidence-accuracy correlation, but no significant effect of feedback, number of subjects, minimum trials per subject, maximum trials per subject, number of tasks times conditions, or number of difficulty levels. These findings suggest that some experimental design choices (such as choosing a certain type of experimental task) may be more favorable for a strong confidence-accuracy relationship than others.

METHODS

Dataset and trial selection

To support this project, datasets from 145 cognitive decision-making experiments were taken from the Confidence Database found on Open Science Framework (Rahnev et al., 2020). The Confidence Database is a collection of experimental datasets presenting the various data gathered from a number of neuroscience experiments from university labs around the world. The experiments whose data were included in the database shared similar experimental setups. In most cases, they presented subjects with tasks that required them to respond by making a choice from many possible choices (one or more of which was the "correct" choice) in a decision-making process. In other cases, subjects had to make a best estimate to an actual value, which would serve as the target or "correct" value; an example would be pressing a spacebar for a certain amount of time to match a previously-shown duration as closely as possible (from the dataset "data_Akdogan_2017_Exp1"). All experiments required subjects to determine and rate how confident they were that they made the "correct" decision or came close to the target value. Each dataset included the decisions/estimates and confidence ratings that each subject made in every trial. From this setup, the mean accuracy (or error, in the case of estimation tasks) of each subject's decisions over the course of the experiment (with the exception of one dataset, mentioned next) as well as the mean confidence could be determined in every experiment. The mean accuracy/error and mean confidence values for all subjects were later correlated to find Pearson correlation values for each experiment.

In the end, 143 of these experimental datasets were included in this study. The dataset "data_Dildine_unpub" was excluded due to having an experimental design from which the mean accuracy from subjects could not be determined, since the experimental task did not involve

having a “correct” answer or target value to which subjects’ responses could be compared. Datasets with fewer than four subjects were excluded, as having very few subjects may result in outlier correlation values. Thus, the dataset “data_Zylberberg_2016” was excluded, which had only three subjects and a Pearson correlation of -0.99976. Practice trials were excluded from this study. All subjects in each experiment were included in determining the Pearson correlation(s) for each dataset. Subjects provided informed consent, and experiments were approved by their associated Institutional Review Boards.

Subset identification and subject separation

Each experiment had one or more subsets (subsets = conditions * tasks). For experiments with only one subset or if all subjects completed all subsets, one coefficient was found for each of these experiments. For experiments with more than one subset, with subjects only completing one or a select few of the subsets, a single coefficient value was found for each subset, resulting in multiple coefficient values for a single experiment. The format of subsets could be Condition A with Task A, Condition A with Task B, or Condition B with Task A.

In order to determine each experiment’s subsets and whether or not subjects should be separated by subset, experiments were manually and individually inspected, using the information from the Excel file “Database_Information,” the readme file descriptions for each experiment, and the experimental data files themselves, all of which can be found on the Confidence Database. Using the details from the “Manipulation” column from the file “Database_Information” (describing each experiment’s manipulations) as well as looking into each experiment’s data file for the column(s) indicating condition separation, the exact identity of each subset was determined based on the name entered on the data file. In order to determine

whether subjects completed all subsets or less, manual inspection and running Python (version 3.7.0) code to identify which subsets subjects completed were both used.

Data extraction and computation

Using Python (version 3.7.0) and imported packages (xlsxwriter, glob, pandas, os, numpy), the mean decision accuracy (or error) and mean confidence were found for every subject in each experiment. The mean accuracy/error and confidence values across all subjects were correlated to find one or more Pearson coefficient(s) for each experiment (depending on whether subjects were separated by subsets, explained previously) in order to quantify the strength of the relationship between accuracy and confidence in each experiment with its unique experimental design. These experimental coefficients were written and stored into Excel files for further analysis.

In order to complete this process, all experimental datasets (stored as csv files) were placed into a single folder. The contents of this folder were loaded into the program using the os listdir() function, and the code parsed through each individual dataset file in the folder to find the necessary values for completing all computations. To find the mean accuracy for each subject, for non-estimation scenarios, the total accuracy was found by summing the accuracy values from all trials completed by the subject, then dividing this number by the number of trials completed by the subject. For each trial in which the subject was correct, accuracy was denoted as 1. Incorrect trials had accuracy values of 0. For the experiments that utilized estimation in their tasks, mean error values (rather than mean accuracy) were found for each trial by calculating the absolute value difference from the subject's estimation with the target value. These error values were summed for all trials completed by the subject and then divided by the number of trials

completed. To find the mean confidence for each subject, all confidence values were summed to find total confidence, which was then divided by number of trials completed. After mean values were found for all subjects in the experiment, mean accuracy (or error) values were placed into one or more numpy arrays, and mean confidence values were placed into one or more numpy arrays. In the case that the experiment only had one subset or subjects completed all subsets, one array was used to store all subjects' mean accuracy/error values, and another array was used to store all subjects' mean confidence values. Each subject's accuracy and confidence values were in the same order in their respective arrays, and the values in the mean accuracy/error array were correlated to the values in the mean confidence array using the numpy corrcoeff() function to find a single Pearson correlation coefficient. In the case that the experiment had more than one subset and subjects were to be separated by subset, one mean accuracy/error array and one mean confidence array were designated for each subset (e.g. six total arrays, three for accuracy/error and three for confidence, for an experiment with three subsets). Subjects' values were placed into the appropriate arrays based on which subset they had completed (e.g. subject one's accuracy/error and confidence values placed in the accuracy/error and confidence arrays for subset X, while subject two's values were placed in arrays for subset Y). After all values have been placed into the appropriate arrays, the mean accuracy/error and confidence values for each subset were correlated to find a single Pearson coefficient per subset, resulting in multiple coefficients for the experiment. An extra step was taken if the experiment utilized estimation; after the Pearson coefficient was calculated, the sign was flipped so that positive coefficients became negative and vice versa.

A blank Excel file was opened using the xlsxwriter functions workbook() and add_worksheet(). After the coefficient(s) for each experiment were calculated, they were written

into the Excel file, with the values in each row corresponding to each experiment and spanning multiple columns in the row if there were multiple coefficients. At the end of this process of computing coefficients, the Excel file was filled with the coefficient values for all experiments.

Statistical analysis

For statistical analysis, tests were run in Excel using the Analysis ToolPak (AT) and the externally-retrieved Real Statistics Resource Pack (RSRP) add-ins. Using the data from the created Excel file, the mean correlation was computed from all coefficients found and compared to a mean of zero using a one-sample t-test (using RSRP).

Furthermore, a secondary analysis was conducted to determine whether experimental coefficients were significantly associated with different values of each experimental characteristic. For example, correlating all experimental coefficients with their respective number of choices on a confidence scale (e.g. 2 for a 2-point scale, 4 for a 4-point scale) would aid in determining whether having a higher or lower coefficient is associated with using a confidence scale with a wider or narrower range of choices. In this analysis, the experimental coefficients were correlated with the experimental values of the particular characteristic of interest to find another Pearson r and also a p-value indicating significance of the correlation. This correlation process was done for all experimental characteristics, except experiment type and feedback type: number of subjects, minimum trials per subject, maximum trials per subject, number of tasks times conditions, confidence scale (discrete integer values or on continuum), and number of difficulty levels (discrete number of difficulty levels or varying difficulty on continuum). Six Pearson coefficients and six p-values were determined, with one r-value and one p-value for each characteristic.

In order to complete this process, Python (version 3.7.0) and imported packages (xlsxwriter, glob, pandas, os, numpy, scipy.stats) were again used. All subset values from all experiments were placed into one numpy array while the values of interest corresponding to each subset value were placed into another array. For example, if the experimental characteristic of interest was number of subjects, all subset values were placed into one array, and the other array held the number of subjects for each subset in the corresponding order. The values from the two arrays were correlated to find a Pearson coefficient as well as a p-value indicating statistical significance using the scipy.stats function pearsonr(). This process was repeated for all experimental characteristics of interest: number of subjects, minimum trials per subject, maximum trials per subject, number of tasks times conditions, confidence scale, and number of difficulty levels. A blank Excel file was opened using the xlsxwriter functions workbook() and add_worksheet(). After each Pearson coefficient and p-value were calculated, they were written into the Excel file, with the values in each row corresponding to the particular experimental characteristic and spanning two columns, one column for each value.

For the number of difficulty levels, experiments with continuous difficulty were excluded from secondary analysis (21 experiments). Continuous difficulty refers to the conditions in which difficulty varies across a continuous range (such as visibility of a stimulus ranging from very visible to barely visible, including all visibilities in between) rather than distinct levels of difficulty set by the experimenters (such as when the task visibility is set to only a few different, distinct values). For confidence scale, experiments with continuous confidence or a varied use of discrete and continuous confidence scales were excluded from secondary analysis (33 experiments). Continuous confidence refers to using scales (such as sliding 0-100 scale that subjects move a marker across to indicate their confidence) that have set maximum and

minimum values and allow for confidence ratings to be any values in between, including decimals. Discrete confidence scales (2-point, 3-point, 4-point, etc.) utilize distinct integer values (0, 1, 2, 3, etc.) to denote level of confidence.

For experiment type, experimental coefficients were grouped by the experimental categories found from the Confidence Database, specifically into cognitive, perception, motor, mixed (combination of two or more of the other categories), and memory experimental categories. After separation into groups, a single-factor ANOVA was run, followed by a Tukey's HSD test (using RSRP). For feedback type, experimental coefficients were grouped into feedback (trial-by-trial feedback present) and no feedback categories (no feedback or feedback only given after blocks), and an independent sample t-test for unequal variances was run (using AT), after determining whether variances were equal or not (using two-sample F-test on AT).

Confidence Database

The database from which experimental data was taken can be found here: https://osf.io/s46pr/

Real Statistics Resource Pack

The Excel statistical analysis add-in can be found here: https://www.real-statistics.com/free-download/real-statistics-resource-pack/

RESULTS

General trend in correlation

One aim of this study was to determine whether a significant relationship between confidence and accuracy existed in subject responses by looking at correlation coefficients from all experiments. After computing all of the experimental coefficients, in order to determine if a significant relationship existed, coefficient values were run in a one-sample t-test comparing the mean value from all coefficients with the mean of zero. Overall, a significant correlation was found from the experimental data, as the mean Pearson correlation from all subsets was significantly different from zero (mean = 0.209; $t(215) = 10.017$, $p = 1.257\text{E-}19$, Cohen's $d = 0.682$). Not only was the mean correlation significant, but it was also slightly positive (0.209). This result indicates that a significant, positive relationship exists between confidence and accuracy values, which follows the logical trend that confidence and accuracy increase and decrease together when proper metacognitive ability is present.

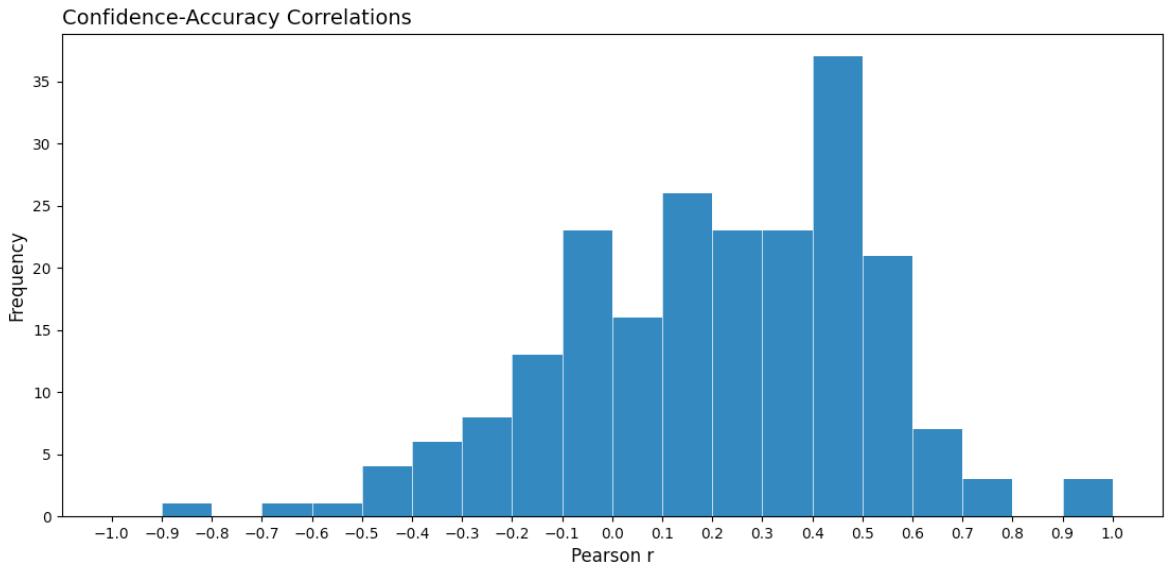

Figure 1. Histogram of all experimental r-values (mean of 0.209, significantly different from 0).

Experiment type effect

Another aim of this study was to determine whether the type of experimental task used had a significant effect on the correlative relationship between confidence and accuracy, similar to the significant effects of task type on confidence calibration found in previous research. After separating experimental correlations into the pre-set categories (mixed, memory, perception, cognitive, motor), a single-factor ANOVA was run, followed by a Tukey's HSD test. For experiment type, a significant difference was found between different types of experiments ($F(4, 211) = 4.131$, $p = 0.003$), particularly between perception ($M = 0.152$, $SD = 0.319$) and memory ($M = 0.309$, $SD = 0.268$). The 'mixed' experimental category had the highest confidence-accuracy correlation value of all types ($M_{mixed} = 0.326$, $M_{memory} = 0.309$, $M_{perception} = 0.152$, $M_{cognitive} = 0.218$, $M_{motor} = -0.109$). This result suggests that the type of task administered (thus determining the type of experiment or category the experiment belongs to) significantly affects the confidence-accuracy correlations found, with some experimental tasks associated with higher correlations than others.

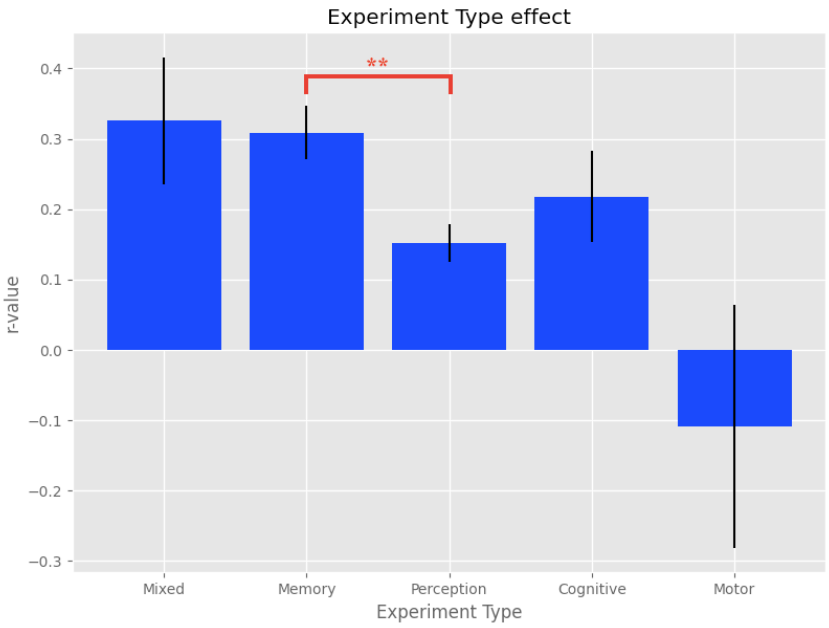

Figure 2. A significant difference was found between the r-values of different types of experiments (p = 0.003), particularly between perception and memory (p = 0.008). The 'mixed' experimental category had the highest mean r-value of all types.

Feedback effect

Feedback was another experimental characteristic that this study sought to investigate, specifically whether trial-by-trial feedback had a significant effect on confidence-accuracy correlation. In order to test the effect of feedback on correlation values, experimental coefficients were grouped into feedback (trial-by-trial feedback present) and no feedback categories (no feedback or feedback only given after blocks), and an independent sample t-test for unequal variances was run (after determining unequal variances with F-test). For feedback type, no significant difference was found between feedback and no feedback groups ($t(12) = 0.932$, $p = 0.370$). This finding runs against the intuitive notion that feedback should aid in helping subjects reach better calibration (in this case, a higher correlation between confidence and accuracy), as this result indicates that feedback exerted no significant effect on confidence-accuracy correlation.

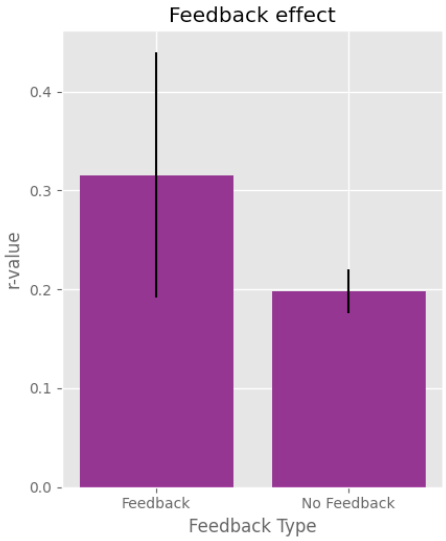

Figure 3. No significant difference was found between the r-values of feedback and no feedback groups ($p = 0.370$).

Effect of remaining characteristics

To complete a comprehensive analysis of experimental characteristics, this study correlated correlation coefficients with values from the remaining experimental characteristics of interest, such as the number of subjects, minimum trials per subject, maximum trials per subject, number of tasks times conditions, number of difficulty levels, and confidence scale range. For the remainder of the experimental characteristics, no significant correlation was found when experimental coefficients were correlated with number of subjects ($r = 0.025$, $p = 0.714$), minimum trials per subject ($r = -0.077$, $p = 0.258$), maximum trials per subject ($r = -0.095$, $p = 0.162$), number of tasks times conditions ($r = 0.007$, $p = 0.916$), or number of difficulty levels ($r = -0.051$, $p = 0.489$). However, a significant correlation was found for confidence scale ($r = -0.168$, $p = 0.028$). This finding indicates that the number of subjects, minimum or maximum trials per subject, number of tasks times conditions (i.e. subsets), and number of difficulty levels each did not have a significant effect on confidence-accuracy correlation. In contrast, confidence scale range had a significant effect on correlation values, with smaller ranges associated with higher confidence-accuracy correlation.

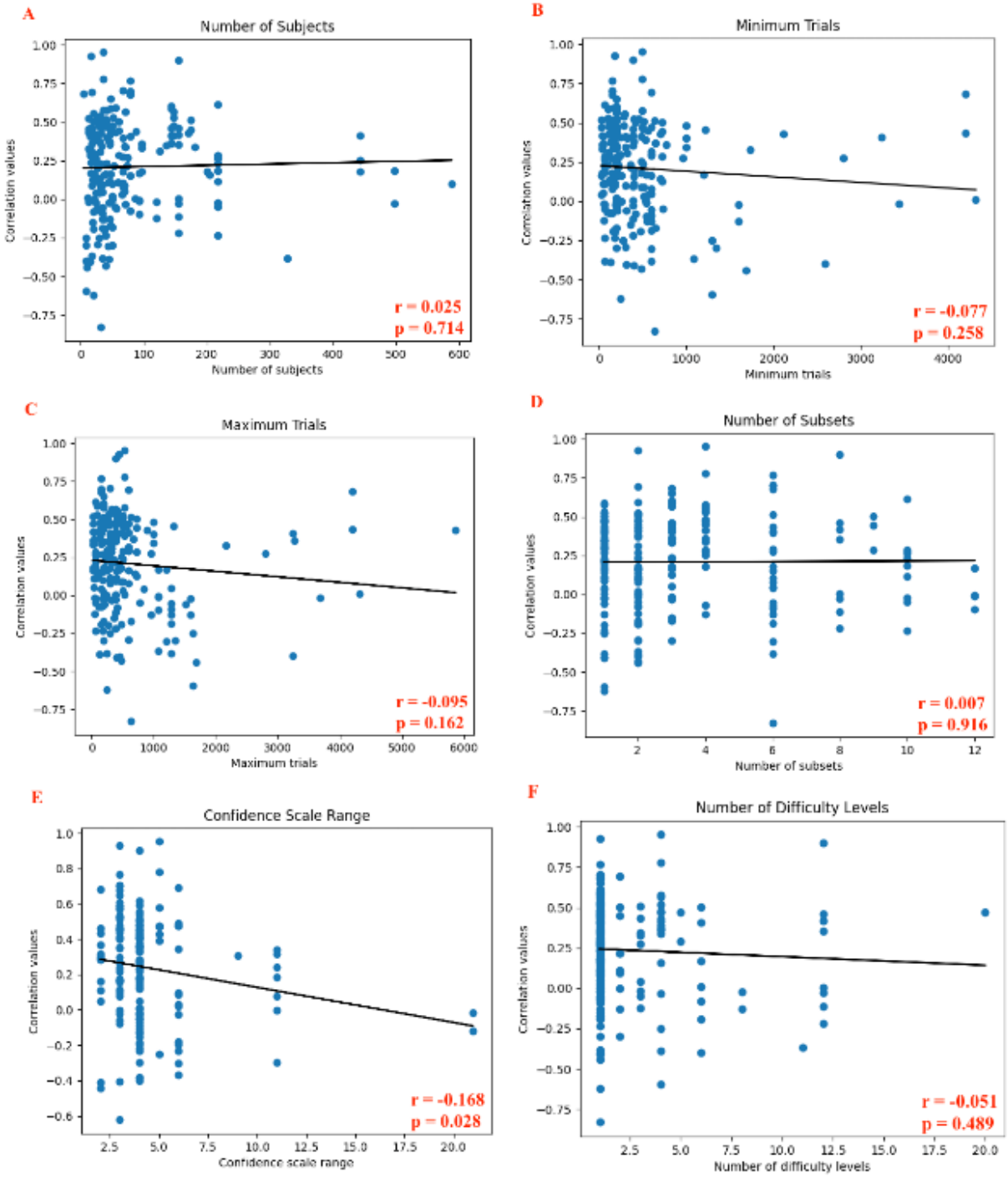

Figure 4. A significant effect was found only for confidence scale range on experimental correlations (E). No significant correlation was found when experimental coefficients were correlated with number of subjects (A), minimum trials per subject (B), maximum trials per subject (C), number of tasks times conditions (D), or number of difficulty levels (F).

DISCUSSION

The aim of this study was to determine how confidence-accuracy correlations significantly differed between various neuroscience experiments by experimental design characteristics and how design choices for each characteristic may have influenced the strength of the correlations. A higher correlation between confidence and accuracy indicates a stronger ability to judge the accuracy of one's decisions, as confidence and accuracy are well-calibrated. Good confidence-accuracy calibration can be useful in correcting the internal biases that contribute to erroneous judgements in many real-world scenarios, such as medical diagnosis and eye-witness testimonies (Yang & Thompson, 2010; Tenney, Spellman, & MacCoun, 2008). This study investigates the experimental conditions in which confidence and accuracy are best-calibrated.

To begin, a significant, positive mean correlation from all subsets was found, following the general trend in confidence-accuracy correlation, with accuracy and confidence values increasing and decreasing together, indicating the presence of metacognitive ability in experimental subjects. Differences in confidence calibration have been found between decision-making experiments of different tasks: sensory discrimination, general knowledge, and memory-recall (Bjorkman, Juslin, & Winman, 1993; Luna & Martín-Luengo, 2012). Similarly, the results of this study indicate that correlations between experiments of different categories (indicating type of task), specifically perception and memory, are significantly different, with mixed-type experiments having the highest correlations. This suggests that the type of task given to subjects influences their metacognitive ability, with subjects showing higher metacognitive ability for tasks of some types than others.

While the effects of feedback on metacognition are still relatively unclear, previous research suggests that feedback can improve calibration depending on the task given, such as in globally hard tasks, where there are more hard trials than easy (Petrusic & Baranski, 1997). Feedback may have an effect on metacognitive behavior by changing subjects' decision-making strategies after they receive it (Vollmeyer & Rheinberg, 2005). For example, a subject is biased to answer one choice over another in a two-choice decision-making task without feedback. However, he or she may decide to choose that choice with less frequency if feedback indicates that the frequency of wrong answers increases with that choice. The subject may also choose to lower the confidence rating when choosing that choice. Thus, metacognitive ability and confidence calibration may be enhanced. However, the results of this study indicate no significant difference in confidence-accuracy correlation between feedback and no feedback groups.

While previous studies have focused on the differences in calibration between different types of confidence reporting methods, this study tested whether a significant correlation between confidence-accuracy correlation and confidence rating range was present. Previous research has shown that differences in calibration were present between the data collected using the perceptual awareness scale, confidence rating, and post-decision wager techniques (Sandberg, Timmermans, Overgaard, & Cleeremans, 2010). Research on eye-witness testimonies, specifically testing face recognition in suspect lineups, show that a superior confidence-accuracy calibration was found when subjects used a half-range confidence scale (50-100%) as opposed to a full-range (0-100%) (Weber & Brewer, 2003). In line with previous research, a significant correlation between confidence-accuracy correlation and confidence rating range was found, and smaller scales were associated with higher correlation. In particular, for

confidence scales with discrete integer ratings to choose from, scales with smaller ranges were significantly different from scales with larger ranges in terms of effect on correlation, with correlation decreasing as range increased. While a significant correlation was found for confidence ratings, no significant correlations were found between confidence-accuracy correlations and other experimental characteristic values, such as number of subjects, minimum trials per subject, maximum trials per subject, number of tasks times conditions, and number of difficulty levels.

With the large number of experimental datasets included in this study, there was a wide variety of experimental setups and tasks and thus the possibility for many confounding variables. We aimed to extract all of the relevant information needed to conduct this study, such as mean accuracy, mean confidence, and confidence-accuracy correlation, using a methodology that minimizes the impact of experimental differences so that values from the experiments could be comparable to each other and grouped together for statistical analysis. However, we had no control over how experimenters inputted their data and the various choices that they made for how they wanted to record the data given to the database. Experimenters may have used different methodologies for formatting their data for presentation; this suggests that the information that we used from the datasets may have already been filtered by the individual experimenters' choices. Differences in data input methods may act as a confounding variable, contributing to the differences in correlation values extracted between experiments. In our methodology, for an experiment in which subjects only completed one or a select few of the subsets (subsets = conditions * tasks), a single coefficient value was found for each subset, resulting in multiple coefficient values for this experiment. However, we had no control over the differences that discriminated one task from another task or one condition from another condition. In one

experiment, one task may be very similar to the other task(s), with only a minor difference, while in another experiment, tasks may be significantly different from each other. These possible differences in the judgements and decisions made by individual researchers concerning the meanings of terms such as condition, manipulation, and task, as well as what each condition, manipulation, and task consisted of, present another possibility for confounding variables. These two examples of input methods and term meanings are only two areas out of many that confounding variables can present themselves to influence the results of our study. In future studies, we could use experiments that were more similar to each other, such as experiments performed by the same researchers, in order to minimize the number of confounding variables. These experiments could be set up using the same methodologies for determining what separates one condition or task from another, so that the results from our future studies are more likely due to effects from the various experimental characteristics (number of subjects, number of subsets, number of difficulty levels, etc.), and not differences from experimenters' choices on what constitutes a separate condition, task, or difficulty level.

Our own decisions in methodology may have impacted our study results. For experiments in which subjects did not complete all subsets, we decided to extract one correlation for each subset, resulting in multiple correlation values for each of these experiments. Experiments with continuous difficulty were excluded from the secondary analysis of the relationship between number of difficulty levels and confidence-accuracy correlation. Experiments with continuous confidence ranges were excluded from the secondary analysis of the relationship between confidence ranges and confidence-accuracy correlation. Experimental results could have differed if we extracted one correlation per subset for experiments in which subject completed all subsets. Results could also have differed if we developed a method to quantify continuous difficulty

levels and continuous confidence scales in a way that would make them comparable to discrete difficulties and confidence ranges and included them in analysis with discrete difficulty or confidence.

In future studies, we could investigate how results would change if we extracted one correlation for each subset in all experiments, regardless of whether subjects completed all subsets or not, and used these correlations for analysis. We could also include continuous difficulty and continuous confidence in future analyses, such as investigating whether there is a significant difference between confidence-accuracy correlations using a discrete confidence scale and a continuous scale.

The results of this study suggest that high confidence-accuracy correlation can be induced by setting up the experimental design with select characteristics associated with high correlation. By developing an experiment with the right combination of design features, high association between confidence and accuracy may be induced in an experimental setting. This implication can be further researched by setting up an experiment testing how different combinations result in different correlations. More broadly, effective combinations could be translated and applied to real-world settings in which high confidence-accuracy calibration is needed.

CONCLUSION

This study aimed to determine the experimental design conditions under which confidence-accuracy correlations are the strongest. To this end, this study investigated confidence-accuracy correlation across subjects for 143 cognitive neuroscience experiments from the Confidence Database on Open Science Framework. Using their respective across-subject correlations, this study determined for these experiments whether their correlation strengths differed between each other by their unique experimental design characteristics. A significant, positive mean correlation was found from all subsets, following the general trend in confidence-accuracy correlation. It also found that correlations between experiments of different categories, specifically perception and memory, are significantly different, with mixed-type experiments having the highest correlations. There was a significant effect of confidence range on confidence-accuracy correlation, but no significant effect of feedback, number of subjects, minimum trials per subject, maximum trials per subject, number of tasks times conditions, and number of difficulty levels.

This study was able to reveal significant differences between experiments of different categories based on task type and confidence scale range as well as the presence of general metacognitive ability in experimental subjects. Future studies are needed to further investigate the effects of the design characteristics for which this study could not find a significant difference. By finding the right combinations of design characteristics for good metacognition, these combinations could be translated and applied to real-world settings in which high confidence-accuracy calibration is needed.

REFERENCES

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Percept Psychophys, 61*(7), 1369-1383. doi:10.3758/bf03206187

Bjorkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: the underconfidence phenomenon. *Percept Psychophys, 54*(1), 75-81. doi:10.3758/bf03206939

Fleming, S. M., & Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: implications for measures of awareness. *Conscious Cogn, 19*(1), 352-363. doi:10.1016/j.concog.2009.11.002

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?: The calibration of probability judgments. *Organizational Behavior and Human Performance, 20*(2), 159–183. doi:10.1016/0030-5073(77)90001-0

Luna, K., & Martín-Luengo, B. (2012). Confidence–Accuracy Calibration with General Knowledge and Eyewitness Memory Cued Recall Questions. *Appl. Cognit. Psychol., 26*(2): 289-295. doi:10.1002/acp.1822

Petrusic, W. M., & Baranski, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *The American Journal of Psychology, 110*(4), 543-572. doi:http://dx.doi.org/10.2307/1423410

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., … Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring

consciousness: Is one measure better than the other?. *Conscious Cogn, 19*(4), 1069-1078. doi: 10.1016/j.concog.2009.12.013

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

Tenney, E., Spellman, B., & MacCoun, R. (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. doi:10.1016/j.jesp.2008.04.006

Vollmeyer, R., & Rheinberg, F. (2005). A surprising effect of feedback on learning. *Learning and Instruction*, *15*(6), 589–602. https://doi.org/10.1016/j.learninstruc.2005.08.001

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *J Appl Psychol, 88*(3), 490-499. doi:10.1037/0021-9010.88.3.490

Yang, H., & Thompson, C. (2010). Nurses' risk assessment judgements: a confidence calibration study. *J Adv Nurs, 66*(12), 2751-2760. doi:10.1111/j.1365-2648.2010.05437.x

Zarnoth, P., & Sniezek, J. A. (1997). The Social Influence of Confidence in Group Decision Making. *J Exp Soc Psychol, 33*(4), 345-366. doi:10.1006/jesp.1997.1326