Elicitations of Confidence: A Tale of Two Methods

By

Sandy Campbell

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Don A. Moore, Chair
Professor Leif Nelson
Professor Stefano DellaVigna

Spring 2024

Elicitations of Confidence: A Tale of Two Methods

ABSTRACT

Elicitations of Confidence: A Tale of Two Methods

by

Sandy Campbell

Doctor of Philosophy in Business Administration

University of California, Berkeley

Professor Don A. Moore, Chair

People often report confidence in two stages: first, they select an answer; second, they report how confident they are that that answer is correct. But research suggests an alternative: one-stage elicitations ask for a belief distribution over the entire set of answers. This dissertation compares these two elicitation methods. In so doing, I shed light on to the psychology of confidence and provide guidance for survey design.

I find across seven studies that one-stage elicitation formats lead to higher average peak confidence than two-stage elicitation formats and that difficulty moderates this relationship, such that the effect is stronger when the question posed is difficult. Familiarity with elicitation format does not attenuate the effect. The effect remains stable across different stimuli, and when participant decisions are incentive compatible.

In examining why this effect occurs, I find that it is driven by two-stage responses that violate rationality. Across three of the four pre-registered studies, I find that the two-stage responses that indicate less than 50% confidence in the selected answer are driving the main result. When these responses are filtered out, the difference in peak confidence between elicitation formats disappears, as does the moderating effect of question difficulty.

Our confidence in our ability to select the right option has enormous consequences. When we are sure, we can commit without hesitation. When we are unsure, then it might be wise to hedge our bets, to delay, or to gather more information. By gaining insights into how people form and communicate confidence in their judgments and identifying effective confidence elicitation techniques, decision-makers can make more informed and effective decisions, leading to better outcomes for organizations and individuals alike.

TABLE OF CONTENTS

ABSTRACT 1

CHAPTER 1 | THEORETICAL BACKGROUND AND LITERATURE REVIEW 1

One-stage Confidence Elicitation 1
Two-stage Confidence Elicitation 2
Extant Literature on Confidence Elicitations 2
Competing Hypotheses 3

CHAPTER 2 | EXPLORATORY STUDIES 6

Confidence Elicitation Manipulation 6
Stimuli 7

CHAPTER 3 | PRE-REGISTERED STUDIES 9

Study 4: Does the main effect of elicitation format replicate in a sample of undergraduates? 9
Method 9
Participants 9
Design 9
Procedure and Materials 10
Results 10
1) Pre-registered analysis: ANOVA model 10
2) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition 12
3) Other Notes 14
Discussion 15
Study 5: Is there an interaction between elicitation format and question difficulty? Does deliberation moderate the relationship between elicitation format and peak confidence? 16
Method 16
Participants 16
Design 17
Procedure and Materials 17
Results 17
1) Pre-registered analysis: ANOVA model 17
2) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition 19
3) Pre-registered exploratory analysis: Exploring deliberation as a mediator 22
4) Exploratory analysis: Coin flip 24
5) Other Notes 26
Discussion 26
Study 6: Do participants prefer one elicitation over the other, and if so, does this change peak confidence and accuracy? 28
Method 28
Participants 28
Design 28
Procedure and Materials 28

Results 28
1) Descriptives..28
2) Pre-registered analysis: ANOVA models 32
3) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition35
4) Other Notes 37
Discussion...38
STUDY 7: DOES THE EFFECT REPLICATE WHEN PARTICIPANT CHOICES ARE INCENTIVE COMPATIBLE, AND WHEN DIFFERENT STIMULI ARE USED?.. 39
Method...39
Participants..39
Design..39
Procedure and Materials..39
Results 40
1) Pre-registered analysis: ANOVA Model 40
2) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition41
3) Other Notes 43
Discussion...43
CHAPTER 4 | GENERAL DISCUSSION AND FUTURE DIRECTIONS 44
Why studying confidence elicitations is important 44
Contributions to the literature..44
Future Directions 46
Concluding Remarks 47
REFERENCES 48
SUPPLEMENTARY MATERIALS 53
EXPLORATORY STUDIES... 53
STUDY 1: DOES A ONE-STAGE ELICITATION FORMAT LEAD TO HIGHER AVERAGE PEAK CONFIDENCE THAN A TWO-STAGE FORMAT IN A SAMPLE OF UNDERGRADUATES? 53
Method...53
Participants..53
Design..53
Procedure and Materials..53
Results 54
1) ANOVA model with the task factor 54
2) Main effect of question difficulty 55
3) Main effect of condition..58
4) Interaction between condition and difficulty 59
5) Deviations from rationality in the two-stage condition 60
6) Other Notes 64
Discussion...64
STUDY 2: DOES THE MAIN EFFECT OF CONDITION REPLICATE IN A SAMPLE OF CLOUD RESEARCH APPROVED AMAZON MECHANICAL TURK PARTICIPANTS? 65
Method...65
Participants..65
Design..65
Procedure and Materials..65
Results 66
1) ANOVA model 66

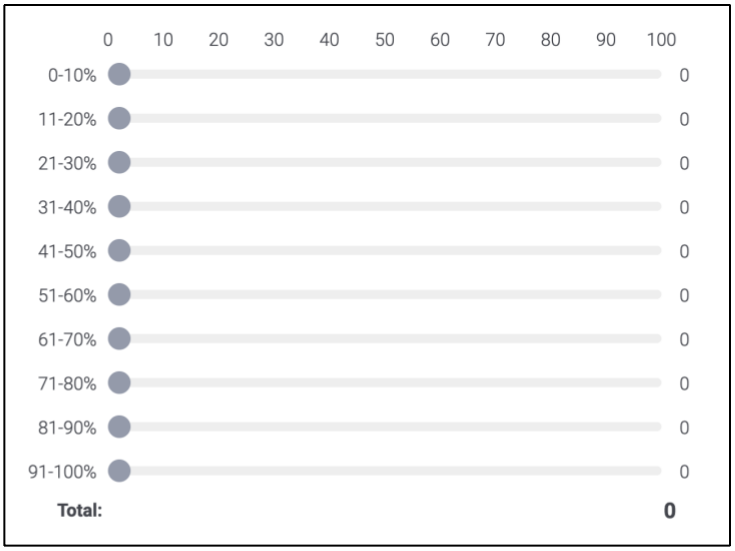
2) Main effect of question difficulty 66
3) Main effect of condition....68
4) Interaction between condition and difficulty 68
5) Deviations from rationality in the two-stage condition 69
6) Other Notes 71
Discussion....72
STUDY 3: (WHY) DO PARTICIPANTS VIOLATE RATIONALITY IN THE TWO-STAGE CONDITION?.... 73
Method....73
Participants....73
Design....73
Procedure and Materials....73
Results 74
1) ANOVA model....74
2) Main effect of question difficulty 74
3) Main effect of condition....76
4) Interaction between condition and difficulty 77
5) Deviations from rationality in the two-stage condition 77
6) Other Notes 82
Discussion....83
SUPPLEMENTAL ANALYSES FOR PRE-REGISTERED STUDIES.... 84
Study 4: Main effect of difficulty....84
Study 5: Main effect of difficulty....85
Study 6: Main effect of difficulty....86
Study 7: Main effect of difficulty....87

CHAPTER 1 | Theoretical Background and Literature Review

Of the many elicitations of confidence, one key distinction stands out. Two-stage confidence elicitations ask people to choose the answer they think is most likely correct, and then how confident they are in that chosen answer. The two-alternative forced choice paradigm, so commonly employed in the decades-old literature on confidence calibration, frequently uses this sort of two-stage elicitation (Hoffrage, 2004; Liberman, 2004). By contrast, one-stage confidence elicitations ask people to state their belief distribution over the set of possible answers. One-stage elicitations have grown in popularity, in part thanks to claims that they reduce overconfidence (Goldstein & Rothschild, 2014; Haran et al., 2010). In what follows, I will describe both more broadly, along with other types of confidence elicitations, and discuss how one-stage and two-stage elicitations fit into the literature.

One-stage Confidence Elicitation

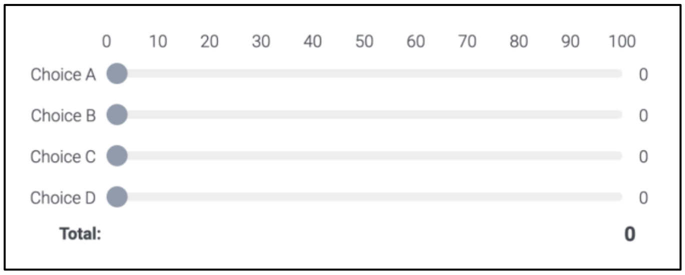
One-stage confidence elicitations ask people to state their belief distribution over all possible outcomes. Research suggests it can reduce people's tendency to fixate on, and become excessively certain in a favored hypothesis (Haran et al., 2010; Moore, 2020). Figure 1 depicts an example of this one-stage elicitation method, based on materials developed by Haran et al (2010). People see an entire range of possible outcomes (e.g., 0% to 100%), partitioned into mutually exclusive bins, and indicate the probability that each bin includes the correct answer. The associated question is naturally one that invites numerical responses (e.g., What percent of UC Berkeley students are business majors?).



0-10%	0
11-20%	0
21-30%	0
31-40%	0
41-50%	0
51-60%	0
61-70%	0
71-80%	0
81-90%	0
91-100%	0
Total:	**0**

Figure 1. Example of the One-Stage Belief Elicitation Interface for Numerical Questions

The one-stage elicitation method can also accommodate non-numerical options. In this case, people see a set of exhaustive options, and indicate the probability that each option is correct. Figure 2 depicts an example of this type of question. You can imagine taking a test that asks how sure you are in each of the answer options, from Choice A to D.



0 10 20 30 40 50 60 70 80 90 100

Choice A 0

Choice B 0

Choice C 0

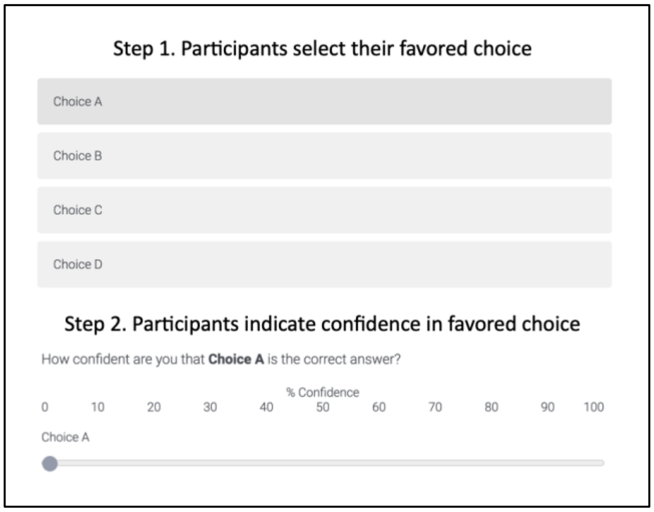
Choice D 0

Total: **0**

Figure 2. Example of the One-Stage Belief Elicitation for Non-Numerical Questions

Two-stage Confidence Elicitation

Two-stage confidence elicitations ask people to choose the answer they think is most likely correct, and then indicate how confident they are in that chosen answer. Figure 3 depicts an example of this two-stage elicitation method with non-numerical answers. People are presented with a set of exhaustive answers and are asked to indicate which they think is the correct answer. After they make their selection, respondents indicate their confidence.



Step 1. Participants select their favored choice

Choice A

Choice B

Choice C

Choice D

Step 2. Participants indicate confidence in favored choice

How confident are you that **Choice A** is the correct answer?

% Confidence

0 10 20 30 40 50 60 70 80 90 100

Choice A

Figure 3. Example of the Two-Stage Belief Elicitation for Non-Numerical Questions

Extant Literature on Confidence Elicitations

Confidence elicitation refers to the process of asking people to report their degree of confidence or uncertainty about a given event or decision. One- and two-stage elicitations are two commonly used methods in the literature for eliciting confidence. The other predominantly used measure is confidence intervals.

One-stage elicitations are more commonly known as histogram measures or SPIES (Subjective Probability Interval Estimates). They involve asking an individual to distribute probabilities over the set of mutually exclusive and exhaustive outcomes. For instance, one might ask an economic forecaster to assign probabilities to different possibilities for GDP growth in the next quarter. One-stage elicitations have been widely used in various fields, including forecasting (Casey, 2021; Murphy & Winkler, 1987) and risk analysis (Clemen & Winkler, 1999).

Two-stage elicitations are more commonly known as item-confidence measures. They involve asking an individual first to make a decision about a specific event or task and then to rate their confidence in that decision. This method is widely used in psychological research and has been applied in domains such as decision making (Kahneman & Tversky, 1972).

In addition to one- and two-stage elicitations, confidence intervals are another commonly used method for eliciting confidence. Confidence intervals are a range of values around a point estimate that reflects the degree of uncertainty in that estimate. For example, a pollster might predict a candidate vote share with a 95% confidence interval of plus or minus 3 percentage points. Confidence intervals have been widely in statistics (Cox, 2006), social sciences (Cumming, 2013), and economics (Giordani & Söderlind, 2003, 2006), among other fields. One potential advantage of confidence intervals is their simplicity and ease of interpretation, making them a popular choice for communicating uncertainty to decision-makers.

Despite the popularity of different elicitation methods, there has been ongoing debate regarding their ability to capture individuals' "real" beliefs accurately. Perhaps the elicitation that has received the most criticism is confidence intervals, with research questioning its credibility (Teigen & JØrgensen, 2005). Evidence suggests that people make a number of systematic errors when specifying confidence intervals (Soll & Klayman, 2004; Teigen & JØrgensen, 2005). These errors are severe enough that it is worth questioning the degree to which people are able to faithfully report percentiles from a subjective probability distribution, as confidence intervals require (Hoffrage, 2004; Moore, Tenney, et al., 2015).

An important thread of this ongoing debate asks whether these different elicitation methods support better calibration – that is, aligning confidence with accuracy. For example, Hu and Simmons (2023) examine whether a one-stage elicitation truly reduces overconfidence. My dissertation focuses largely on whether one elicitation format may lead to higher average peak confidence than another, holding questions constant. It also explores whether the two-stage method captures real beliefs accurately.

Competing Hypotheses

The extant literature suggests that the one-stage confidence elicitation method might lead to lower average peak confidence than the two-stage method. There are good reasons why one-stage confidence elicitations should effectively reduce overconfidence. One-stage elicitations, by construction, build in the most general and effective debiasing technique: they force respondents to "consider the opposite" (Larrick, 2004; Lord et al., 1984). That is, by requiring respondents to report the probability of not only their favored response, but of others as well, it forces them to consider the possibility that they might be wrong. Similarly, Moore (2020, 2023) writes that asking people to complete a belief distribution "forces them to broaden their thinking and consider the possibility that their best guess is wrong," and that this may be what helps reduce overprecision. There is evidence from the forecasting literature that one-stage elicitations do indeed reduce overprecision; this literature notes that one way to help forecasters think through uncertainty is by getting them to think probabilistically (Bruine De Bruin et al., 2000; Chang et al., 2016), and that asking forecasters to report probability distributions leads to better-calibrated forecasts (Haran et al., 2010).

Moore's (2020, 2023) rationale is consistent with the "considering the alternative" literature. Given how significant and pervasive overconfidence is (Kahneman, 2011), prior literature has attempted to reduce this bias by prompting people to consider the outcomes that they do not favor. Hu and Simmons (2023) provide a brief review of this literature, noting that researchers have attempted to get people to consider the alterative through direct instructions (Hoch, 1985; Koriat et al., 1980; Lord et al., 1984; Walters et al., 2016), or by altering the elicitation format (Soll & Klayman, 2004). The one-stage belief distribution elicitation fits more with the latter, and is a more natural way to prompt individuals to consider alternative answers.

An additional reason to expect one-stage elicitations to reduce overprecision comes from research on the partition dependence (Fox & Clemen, 2005). Conversational norms may lead people to assume that all answer options are viable, and deserve at least some credence. Especially when they are unsure, people tend to assign some credence to all the answer options before them. Unknown or unknowable possibilities lead people to assign equal probabilities to each of the given options. This may, for instance, help account for the prevalence of 50/50 probability estimates. For example, Fischhoff and Bruine de Bruin (1999) asked their undergraduate respondents about the probability of experiencing a bombing at their university, a break-in at home, contracting cancer, or being infected with AIDS. The prevalence of 50% reports vastly exceeded the actual probability of these rare events.

Hu and Simmons (2023) note that the one-stage belief distributions have gained popularity in recent years, with researchers in management and psychology increasingly using the method to attempt to capture true participant beliefs (Dietvorst & Bharti, 2020; Goldstein & Rothschild, 2014; Hofman et al., 2020; Moore, Carter, et al., 2015; Morewedge et al., 2016; Reinholtz et al., 2021; Soll et al., 2023).

Informed by three exploratory studies, I designed a series of pre-registered studies to examine competing hypotheses. The first hypothesis is that the one-stage confidence elicitation method does indeed lead to lower overprecision than the two-stage method, for the stated reasons above.

The competing hypothesis predicts the opposite – that the two-stage method leads to lower confidence than the one-stage method. How might two-stage elicitations reduce confidence? Two-stage elicitations leave open the possibility that participants report confidence less than 50% in their favored answer choice, even when there are only two options.

This is unlikely for easy questions, where I expect respondents to a one-stage elicitation would place close to 0% confidence in the wrong answer (A) and close to 100% confidence in the obviously right answer (B). In the two-stage condition, similarly, most participants will choose B, then put close to 100% confidence in B.

However, consider an impossibly difficult question. Those in the one-stage elicitation can place close to 50% confidence on each of the answer options, indicating they have no idea which is correct. However, those in the two-stage condition have another outlet to express their utter lack of confidence: Even though it might violate rationality, some participants could report less than 50% confidence in their selected answer. The presence of even a few such reports might bring average confidence down lower in the two-stage condition than it could be in the one-stage condition.

In three exploratory studies, I ask whether there are any interesting differences in confidence under the two elicitation methods and identify results that inform a series of pre-registered studies. The materials, data, and code for all seven studies can be found on OSF.

CHAPTER 2 | Exploratory Studies

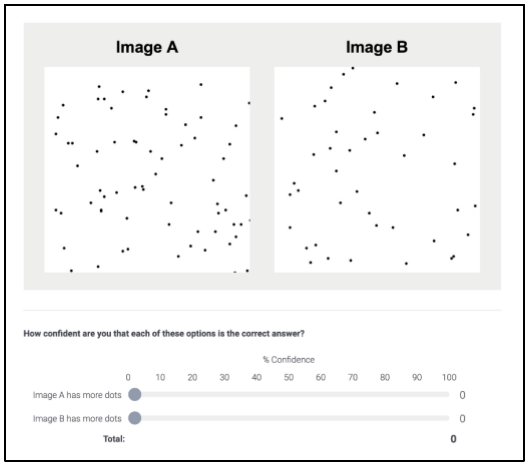
Study 1 tested the basic question of whether the two confidence elicitation formats led to a difference in average peak confidence and helped with stimuli selection. Study 2 focused on the stimuli for which difficulty level was better differentiated and attempted to replicate the results with Cloud Research approved Amazon Mechnical Turk participants, rather than undergraduates. Study 3 extended the results from Study 2 and examined the reasoning behind violations of rationality in the two-stage condition.

Across the three exploratory studies, I found initial evidence of a main effect of elicitation format, and an interaction between elicitation format and question difficulty. I also found evidence suggesting that responses in the two-stage condition were driving the observed effects. However, the results from study to study were somewhat inconsistent (e.g., whether there was a main effect, an interaction effect, or both), as was the size of the effect I found. Given the lack of pre-registration and the underpowered nature of the samples, I relegate the exploratory studies to the Supplementary Materials.

Below, I describe the manipulation and stimuli used in the exploratory studies, as well as the pre-registered studies. I abbreviate the description for the following studies, highlighting only the noteworthy differences.

Confidence Elicitation Manipulation

My experiments typically manipulated how confidence was elicited. In the one-stage confidence elicitation condition, participants distributed their confidence over two answer choices, with one slider scale per answer choice. The survey forced confidence to sum to 100%. Figure 4 presents an example of an easy question from the survey.



How confident are you that each of these options is the correct answer?

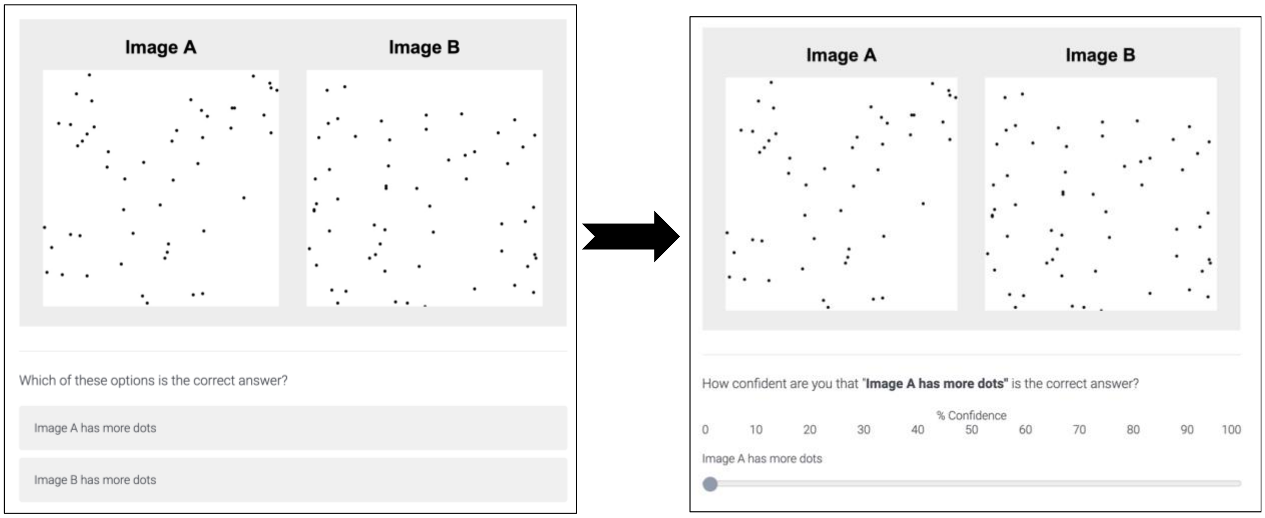
Image A has more dots 0

Image B has more dots 0

Total: 0

Figure 4. Example of the One-Stage Belief Elicitation for Easy Dots Question

In the two-stage confidence elicitation condition, participants first chose which option they believed was the correct answer, and then indicated how confident (0-100% slider scale) they were that their selected answer was indeed correct. **Error! Reference source not found.** below d epicts an example from the survey of a question of medium difficulty under two-stage elicitation. The primary dependent variable was the confidence participants placed on their favored answer choice: I call this peak confidence in discussing any analyses and results.



Which of these options is the correct answer?

Image A has more dots

Image B has more dots

How confident are you that "**Image A has more dots**" is the correct answer?

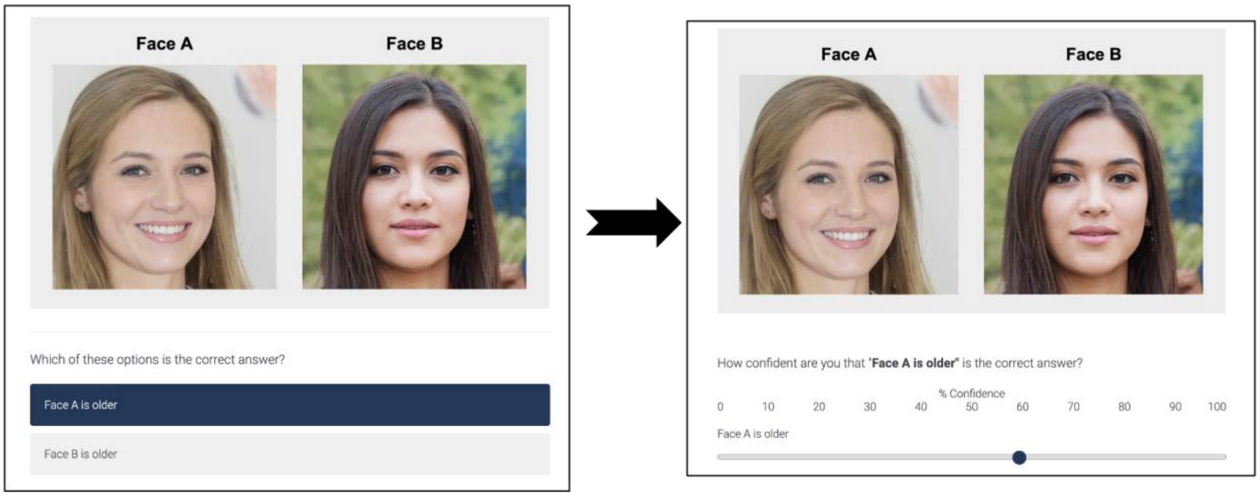
Image A has more dots

Figure 5. Example of the Two-Stage Belief Elicitation for Medium Dots Question

Stimuli

I used two stimuli tasks across the entire package of studies, focusing primarily on the dot task. Both tasks consisted of three questions each (easy, medium, and hard). The dot task showed two images of white squares with black dots inside, and asked participants to indicate which image had more dots. I generated dots uniformly on a 200 by 200 grid. I generated images with some number of dots between 20 and 99, inclusive. For the **hard task**, the images differed by 2 dots (80 vs. 82), for the **medium task**, the images differed by 8 dots (50 vs. 58), and for the **easy task**, the images differed by 24 dots (67 vs. 43).

The face task showed participants two images of AI generated faces, and asked them to indicate which face they thought was older. I generated faces from the FakeFace API. For the hard task, ages differed by 2 years (20 vs. 22), for the medium task, ages differed by 5 years (25 vs. 30), and for the easy task, ages differed by 15 years (25 vs. 40). **Error! Reference source not found.** b elow depicts an example from the survey of a question of medium difficulty under two-stage elicitation.



Which of these options is the correct answer?

Face A is older

Face B is older

How confident are you that "**Face A is older**" is the correct answer?

Figure 6. Example of the Two-Stage Belief Elicitation for Medium Ages Question

CHAPTER 3 | Pre-registered Studies

The three exploratory studies raised a number of interesting questions. They prompted me to run a series of pre-registered studies to further investigate the question of whether confidence elicitation formats can differentially impact average peak confidence. In Study 4, I tested whether the main effect of condition found in the undergraduate sample in Study 1 would replicate. In Study 5, I built upon the results of Study 4 by examining whether a one-stage elicitation leads to higher average peak confidence than a two-stage elicitation for hard versus easy questions, and whether the amount participants deliberated made a difference. Study 6 examined participant preferences for one elicitation method over another, and whether preferences impacted confidence and accuracy. Lastly, Study 7 created an environment in which participant choices were incentive compatible and tested whether the effect replicated with different stimuli. The pre-registration, materials, data, and code for all four studies in this package can be found on OSF.

Study 4: Does the main effect of elicitation format replicate in a sample of undergraduates?

This study examines whether the main effect of condition found in the undergraduate sample in exploratory Study 1 replicates.

Method

Participants

The final sample includes 247 UC Berkeley undergraduate participants (123 Female, 122 Male, 1 Other; 149 Asian, 50 White/Caucasian, 21 Hispanic, 15 Other, 11 African American). Participants were undergraduates who were taking the core business class, undergraduate 105, in Spring semester 2023. As a part of their course requirements, all students in the course were required to complete a pre-survey, which included questions from researchers. This study was pre-registered, and excluded participants whose recorded Qualtrics progress was less than 95%. 124 participants were in the one-stage condition, and 123 participants were in the two-stage condition.

Design

The experiment had a two-cell between-subjects design that manipulated how confidence was elicited. In the one-stage confidence elicitation condition, participants distributed their confidence over two answer choices, with one slider scale per answer choice. In the two-stage confidence elicitation condition, participants first chose which option they believed was the correct answer, and then indicated how confident (0-100% slider scale) they were that their selected answer was indeed correct.

The dot task showed two images of white squares with black dots inside, and asked participants to indicate which image had more dots. Participants in both conditions completed three questions each (easy, medium, and hard).

I did not force confidence to sum to 100 in the one-stage condition, and instead told participants I would normalize to 100. If a participant left both sliders at 0 for a given question, I assigned 50% probability to the two answer choices.

Procedure and Materials

The procedure and materials followed that of the exploratory studies, using the dot image pairs. Participants in both conditions completed one task (the dot task) with three questions (easy, medium, and hard), such that each participant completed three questions in total.

The survey randomly assigned participants to the two between-subjects conditions and randomized the order of questions.

Results

1) Pre-registered analysis: ANOVA model

I planned to run a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on peak confidence. Based on the results from exploratory Study 1, I predicted a main effect of condition, such that average peak confidence in the one-stage condition would be higher than average peak confidence in the two-stage condition.

I find that there is indeed a significant main effect of condition, $F(1, 245) = 57.38, p < .001$. There is of course a significant main effect of difficulty on peak confidence, $F(2, 490) = 59.46, p < .001$. I do also observe an interaction between condition and difficulty, $F(2, 490) = 3.81, p = .023$. I did not pre-register finding this interaction, as it did not appear in the exploratory study with the undergraduate sample. Figure 7 below depicts peak confidence as a function of condition and difficulty, with diamonds indicating average accuracy.

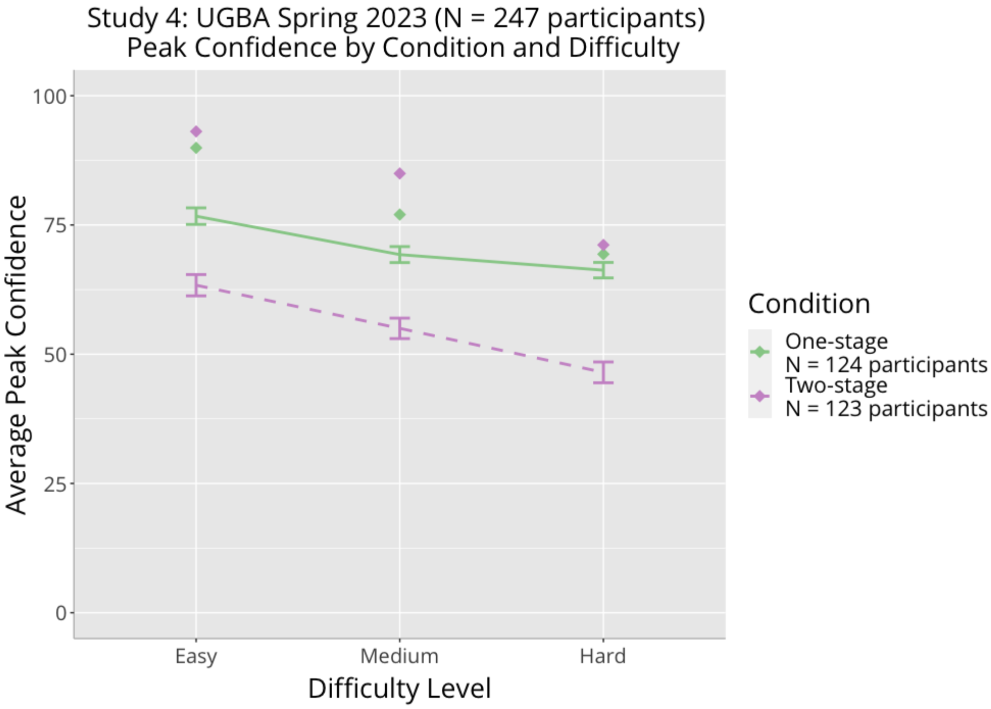


Figure 7. (Study 4) Average peak confidence by condition and question difficulty level. Diamonds indicate average accuracy.

Collapsing across question formats, I find that average peak confidence is 70.06% (*SD* = 21.50) Collapsing across difficulty levels, I find that average peak confidence is 70.74% (*SD* = 17.76) in the one-stage condition, and 54.95% (*SD* = 23.35) in the two-stage condition. Average accuracy does not differ between conditions (one-stage *M* = 78.76%, two stage *M* = 83.06%), *t*(738.02) = -1.64, *p* = .101. Figure 8 depicts average peak confidence in each condition, with average accuracy represented by the red diamonds.

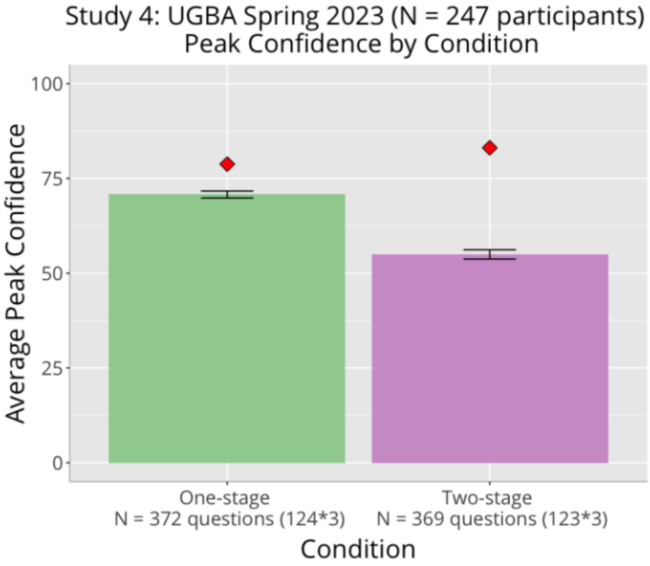


Figure 8. (Study 4) Average Peak Confidence by Condition. Red diamonds indicate average accuracy.

Figure 9 depicts peak confidence by condition and question difficulty, with red diamonds indication average accuracy, and transparent diamonds indicating average peak confidence.

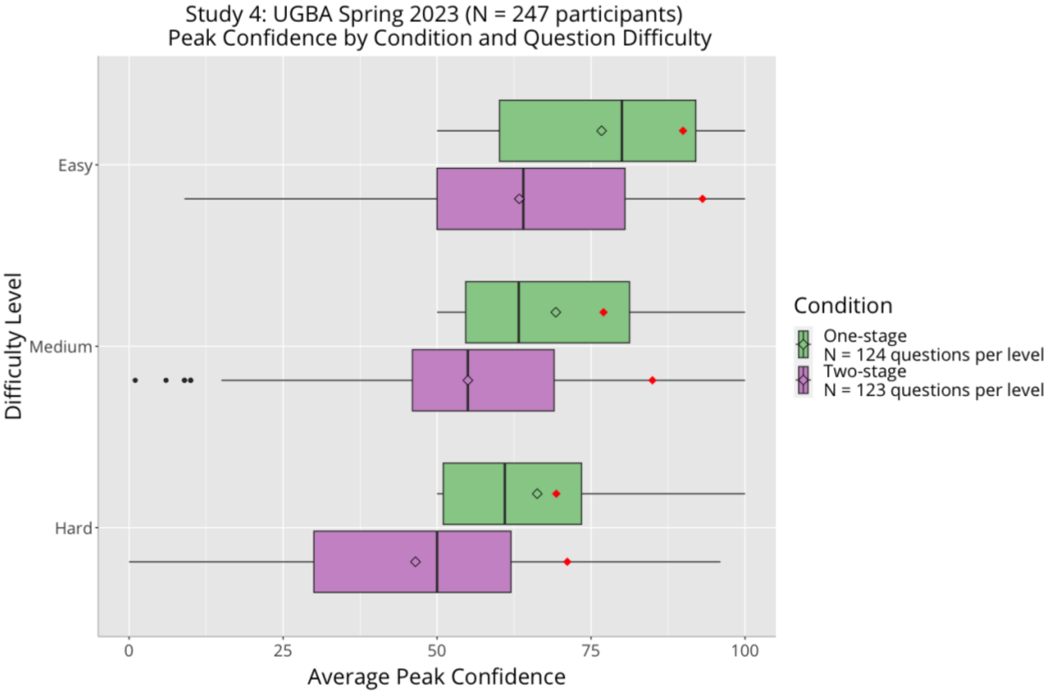


Figure 9. (Study 4) Average peak confidence by condition and by question difficulty level. Red diamonds indicate accuracy, transparent diamonds indicate mean.

2) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition

I examine whether participants violate rationality in the two-stage condition. I find that across all tasks and difficulty levels, 114 out of 369 responses (30.89%) had less than 50 percent confidence in the answer selected in the first stage. Figure 10 depicts average peak confidence for each difficulty level, broken down by those who answered less than 50% confidence and those who answered more.

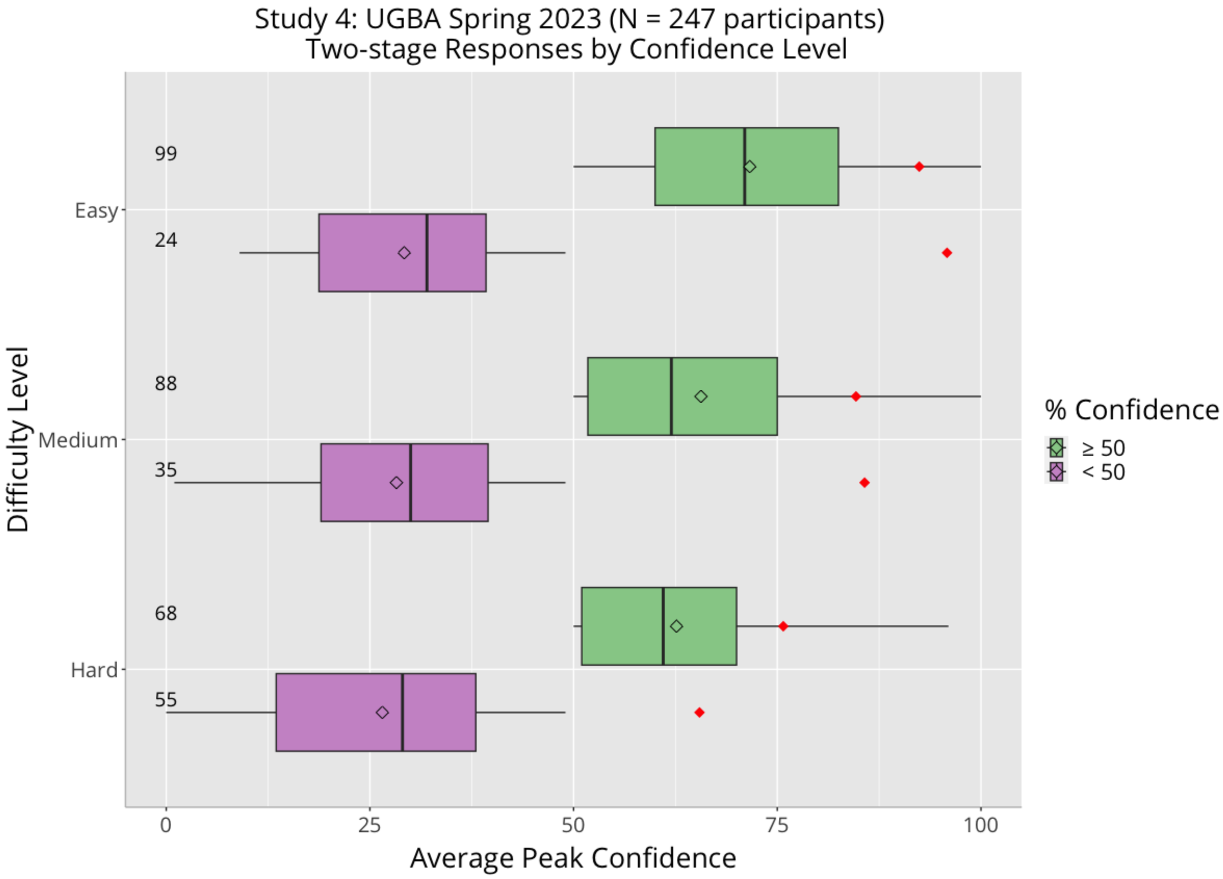


Figure 10. (Study 4) Average peak confidence by question difficulty and confidence level. Red diamonds indicate accuracy, transparent diamonds indicate mean. Number of responses per confidence level listed on the lefthand side.

The percentage of responses indicating less than 50% confidence differs by difficulty level, with 44.72% of responses indicating less than 50% confidence for hard questions in the two-stage condition, 28.46% for medium, and 19.51% for easy (see Table 1 below).

	Question Difficulty Level		
	Easy	Medium	Hard
< 50% confidence	24 responses	35 responses	55 responses
≥ 50% confidence	99 responses	88 responses	68 responses

Table 1. (Study 4) Number of Responses by Task and Confidence Level

Exploring the data further, I filter the dataset and remove two-stage responses that indicate less than 50% confidence in the selected answer and compare peak confidence by difficulty level against the one-stage responses. I planned to run a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on peak confidence with the restricted dataset. The analysis yields a significant main effect of condition $F(1, 229) = 4.37$, $p = .038$. There is of course also a significant main effect of difficulty on peak confidence, $F(2, 390) =$

39.18, *p* < .001. There is no interaction between condition and difficulty, *F*(2, 390) = 0.152, *p* = .859.

Breaking the results down further, I find that direction of the means by condition stay the same. However, the difference in peak confidence between conditions is only significant for the easy questions. Figure 11 depicts these results.

1. For the easy questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 76.71, *SD* = 17.82) and the two-stage condition (*M* = 71.63, *SD* = 16.00), *t*(217.94) = 2.24, *p* = .026.
2. For the medium questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 69.27, *SD* = 17.22) and the two-stage condition (*M* = 65.64, *SD* = 14.21), *t*(205.21) = 1.68, *p* = .095.
3. For the hard questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 66.25, *SD* = 16.69) and the two-stage condition (*M* = 62.63, *SD* = 11.98), *t*(176.61) = 1.74, *p* = .084.

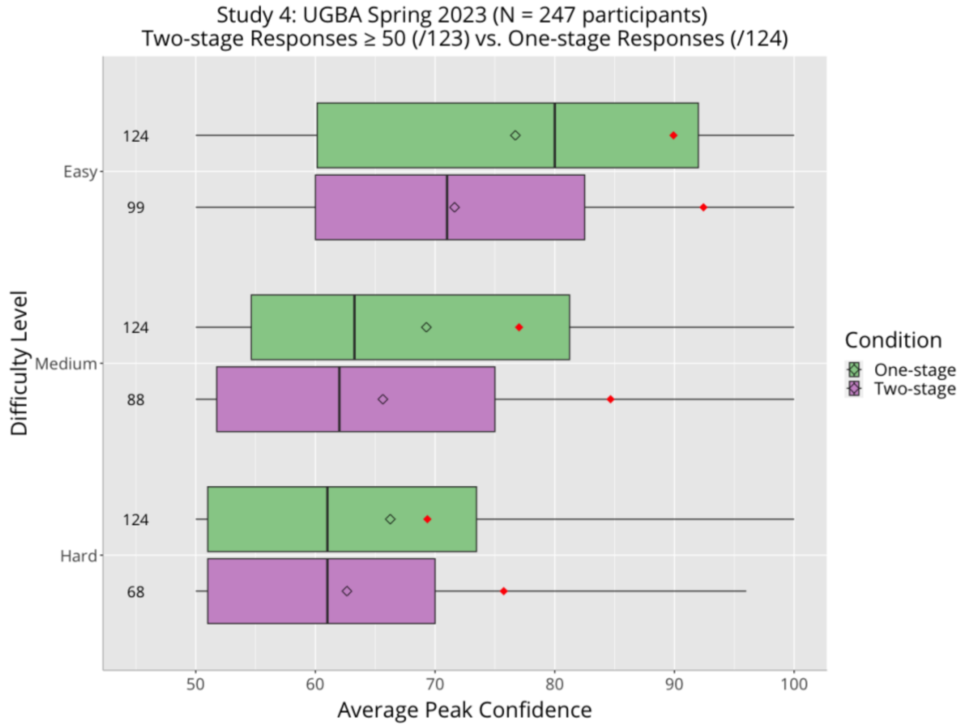


Figure 11. (Study 4) Boxplot showing average peak confidence by question difficulty level and condition, using filtered dataset. Red diamonds indicate accuracy, transparent diamonds indicate mean.

3) Other Notes

I asked participants in both conditions two open-ended questions at the end of the survey. The first asked what they thought the survey was asking, and the second was on whether they found anything in the survey weird or confusing.

In both conditions, participants generally reported the survey was asking about confidence, perception, images, bias, etc. The majority of participants did not find anything about the survey weird or confusing, though there were some open-ended responses by participants in the two-stage condition that noted it was difficult to know what to do if they thought both images had the same number of dots. There were also a few one-off responses in both conditions confused about the relevance of the questions with respect to the rest of the survey, which touched on questions like inequality.

Discussion

In summary, the pre-registered analysis provide support for my hypothesis that there would be a main effect of elicitation format on average peak confidence. This pattern holds directionally when two-stage responses that indicate less than 50% confidence in the selected answer are removed from the dataset. I also find the interaction effect, such that the effect of condition is stronger for the hard questions than for the easy questions.

Study 5: Is there an interaction between elicitation format and question difficulty? Does deliberation moderate the relationship between elicitation format and peak confidence?

I designed Study 5 to build upon the results of Study 4 by examining whether a one-stage elicitation leads to higher average peak confidence than a two-stage elicitation for hard versus easy questions, and whether the amount participants deliberate can make a difference. In other words, I investigated whether there was an interaction effect, and hypothesized that the effect may come about as a result of lower deliberation. As discussed in exploratory Study 1, there are reasons to believe undergraduates may simply be racing through the survey as compared to the best Amazon Mechanical Turk workers (Douglas et al., 2023; Zhang et al., 2024). I saw this as a big difference between exploratory Studies 1 and 2. Undergraduate students might have deliberated less given that the questions were part of a longer pre-survey, whereas Cloud Research approved MTurkers who have a high approval rating might care and deliberate longer. Low deliberation might lead to a main effect of condition, whereas high deliberation might lead to an interaction effect, where the difference in peak confidence between conditions is significant only for difficult questions. I note that given the difficulties inherent in studying deliberation, the pre-registration focused only on the high deliberation sample.

Method

Participants

I calculated sample size based on the effect size found in an exploratory study with the same design, but a smaller sample size. The power analysis yielded a necessary sample size of 322 per group. I planned to recruit 1400 participants: 700 MTurkers in the "low deliberation" group and 700 in the "high deliberation" group.

I ran two HITs concurrently on CloudResearch. For the low deliberation sample, I filtered for MTurkers who had a 0-80% approval rating and had completed 100 or more HITs.[1] For the high deliberation sample, I filtered for MTurkers who had a 95%+ approval rating, had done 100 HITs or more, and who were CloudResearch approved. Those who took previous versions of the study were excluded. I checked post data collection that the samples were mutually exclusive, and they indeed were.

Notably, I had to stop recruitment for the low deliberation sample due to insufficient sample pool size – there were not enough MTurkers who fit the requirements of a 0-80% approval rating, with 100 or more HITs. In hindsight, this makes sense because of selection bias – MTurkers who are not paying much attention are more likely to have a lower approval rating and are less likely to persist and complete 100 HITs. Thus, I stopped data collection after about a month of data collection (June 19, 2023 – July 31, 2023).

The final high-deliberation sample includes 700 Cloud Research approved Amazon Mechanical Turk participants (332 female, 356 male, 7 gender neutral; mean age = 42.60; 526 White, 78

[1] Note that I made two amendments to the pre-registration during data collection.

Black or African American, 46 Asian, 33 Hispanic or Latino, 7 American Indian or Alaska Native, and 8 Other). I filtered out anyone whose recorded Qualtrics progress was less than 95% upon receiving the data. 346 participants were in the one-stage condition, and 354 participants were in the two-stage condition.

The final low deliberation sample includes 162 Amazon Mechanical Turk participants (113 Female, 47 Male, 2 Gender neutral; mean age = 37.90; 115 White, 23 Black or African American, 7 Asian, 9 Hispanic or Latino, 2 Other, 1 Prefer not to answer). I filtered out anyone whose recorded Qualtrics progress was less than 95% upon receiving the data. 82 participants were in the one-stage condition, and 80 participants were in the two-stage condition.

Design

As in prior studies, the experiment used a two-cell between-subjects design that manipulated how confidence was elicited. In this study, I added a coin flip question to the end for exploratory purposes. Participants were told that a fair coin would be flipped, and asked which side they thought the coin would land on. They were then asked to indicate how confident they were in their answer – the question was posed in the one-stage format or two-stage format, consistent with their condition.

Procedure and Materials

The procedure and materials followed that of the exploratory studies, using the dot task. Participants in both conditions completed the dot task with three questions (easy, medium, and hard), such that each participant completed three questions in total.

The survey randomly assigned participants to the two between-subjects conditions and randomized the order of questions.

Results

1) Pre-registered analysis: ANOVA model

For the high deliberation sample, I planned a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on peak confidence. I predicted an interaction between condition and difficulty, such that average peak confidence in the one-stage condition would be higher than average peak confidence in the two-stage condition, but only for the difficult questions.

I find a significant main effect of condition, $F(1, 698) = 42.09$, $p < .001$. There is of course a significant main effect of difficulty on peak confidence, $F(2, 1396) = 257.46$, $p < .001$. I do also observe the predicted interaction between condition and difficulty, but it is just barely significant, $F(2, 1396) = 2.93$, $p = .054$. Figure 12 below depicts peak confidence as a function of condition and difficulty, with diamonds indicating average accuracy.

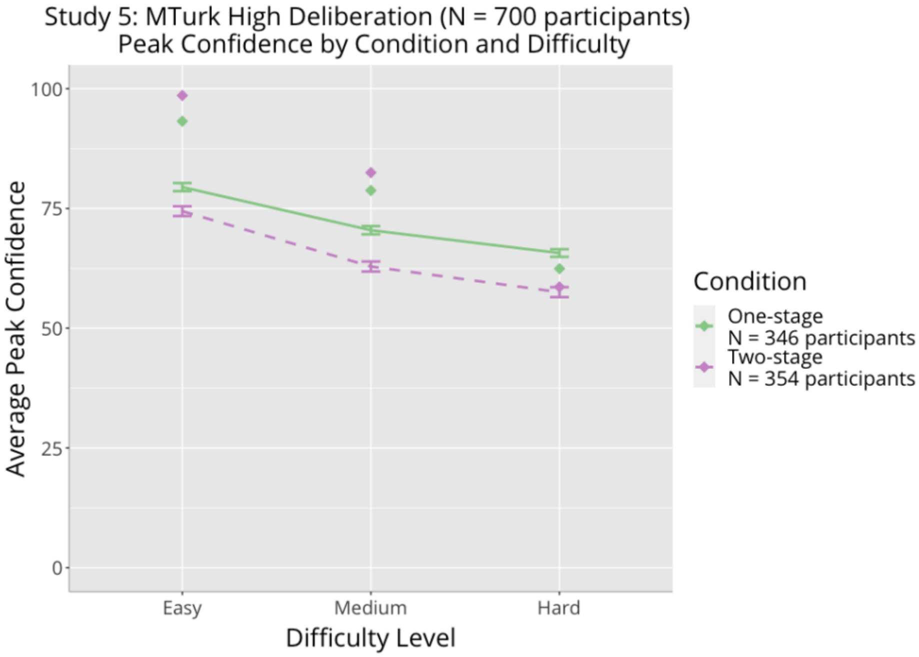


Figure 12.(Study 5) Average peak confidence by condition and question difficulty level. Diamonds indicate average accuracy.

Collapsing across difficulty levels, I find that average peak confidence is 71.87% (*SD* = 16.52) in the one-stage condition, and 64.94% (*SD* = 20.69) in the two-stage condition. Average accuracy does not differ between conditions (one-stage *M* = 78.13, two-stage *M* = 79.89), *t*(2095) = -1.06, *p* = .290. Figure 13 depicts average peak confidence in each condition, with average accuracy represented by the red diamonds.

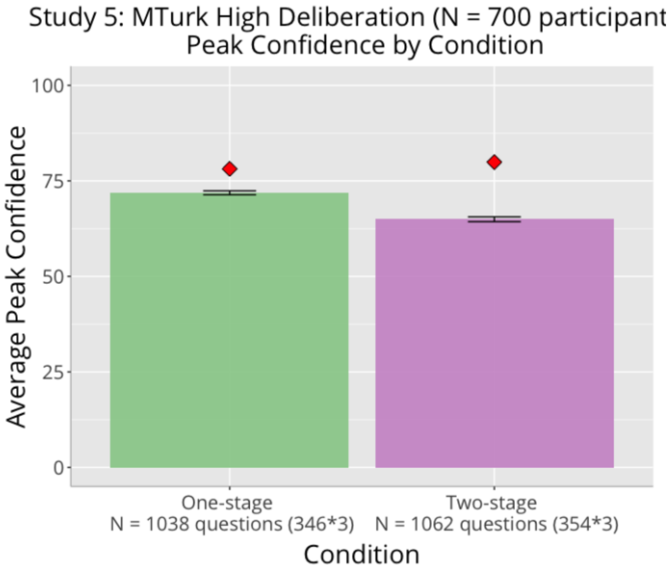


Figure 13. (Study 5) Average Peak Confidence by Condition. Red diamonds indicate average accuracy.

Figure 14 depicts peak confidence by condition and question difficulty, with red diamonds indication average accuracy, and transparent diamonds indicating average peak confidence.

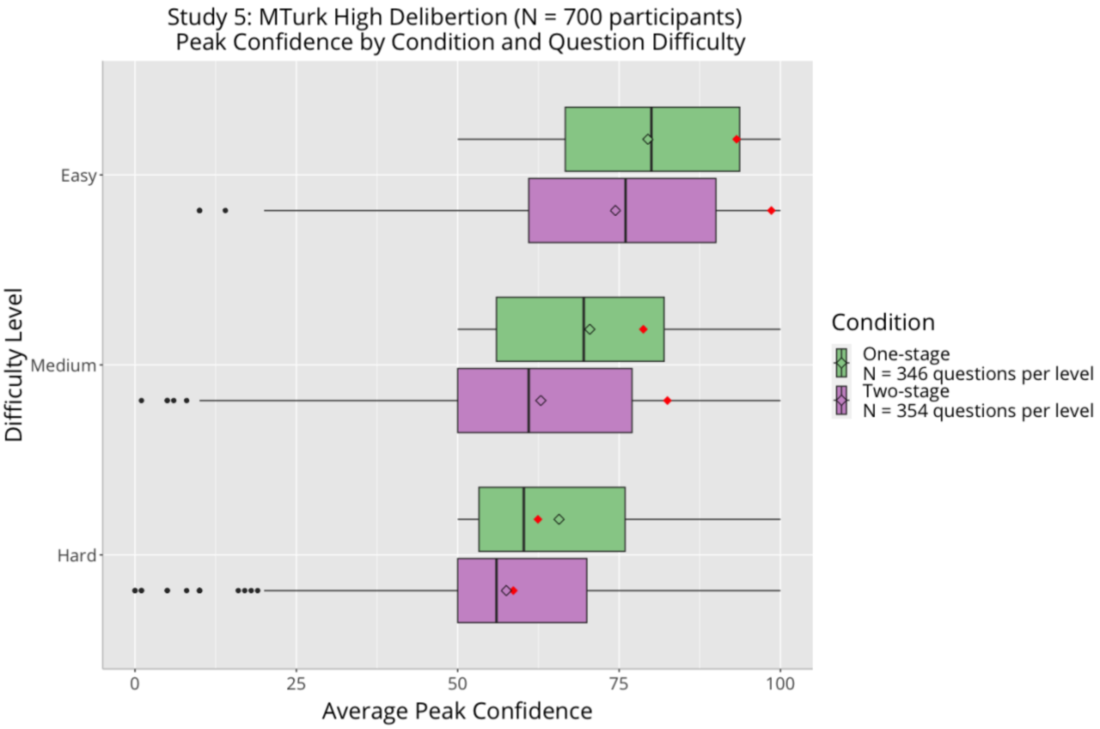


Figure 14. (Study 5) Average peak confidence by condition and by question difficulty level. Red diamonds indicate accuracy, transparent diamonds indicate mean.

2) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition

I examine whether participants violate rationality in the two-stage condition. I find that across all tasks and difficulty levels, 154 out of 1062 responses (14.50%) had less than 50 percent confidence in the answer selected in the first stage. Figure 15 depicts average peak confidence for each difficulty level, broken down by those who answered less than 50% confidence and those who answered more.

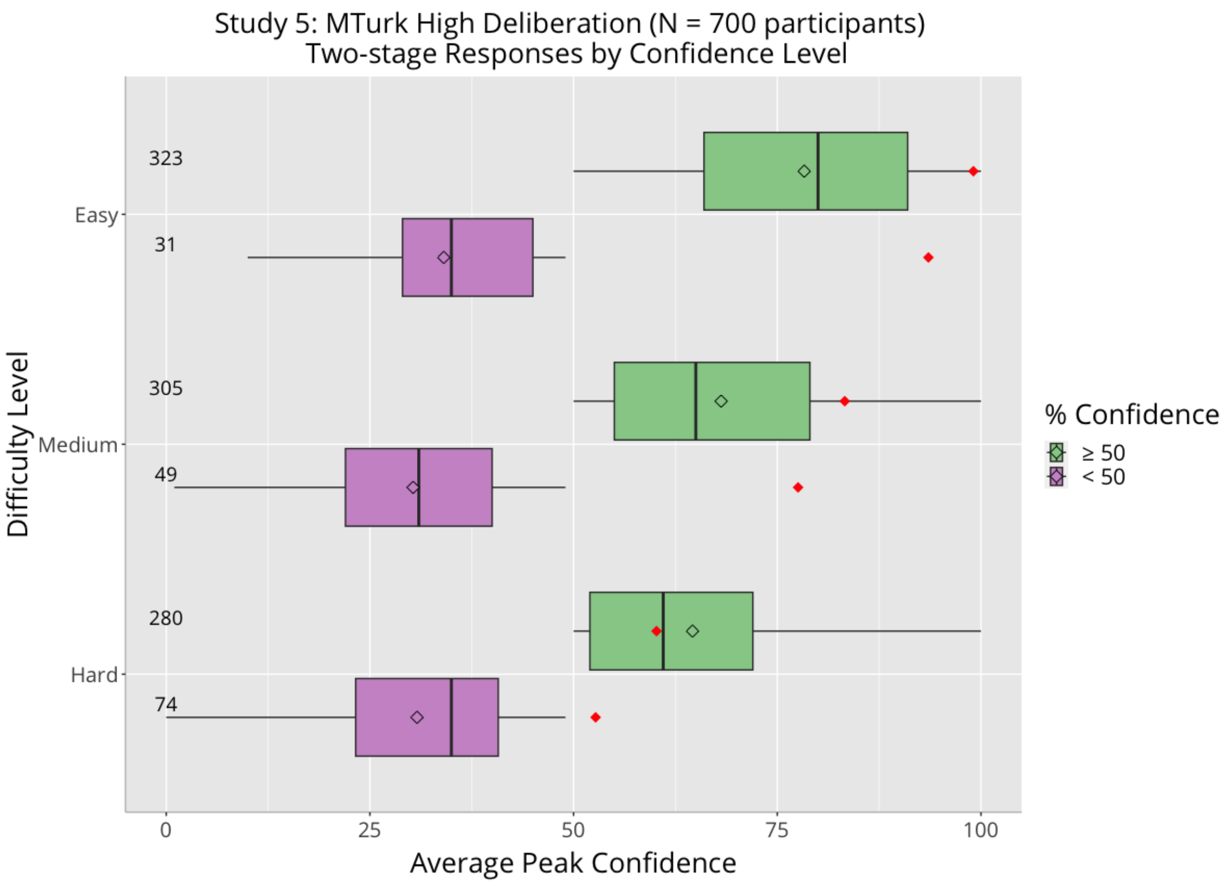


Figure 15. (Study 5) Average peak confidence by question difficulty and confidence level. Red diamonds indicate accuracy, transparent diamonds indicate mean. Number of responses per confidence level listed on the lefthand side.

The percentage of responses indicating less than 50% confidence differs by difficulty level, with 20.90% of responses indicating less than 50% confidence for hard questions in the two-stage condition, 13.84% for medium, and 8.76% for easy (see Table 2 below).

	Question Difficulty Level		
	Easy	Medium	Hard
< 50% confidence	31 responses	49 responses	74 responses
≥ 50% confidence	323 responses	305 responses	280 responses

Table 2. (Study 5) Number of Responses by Task and Confidence Level

Exploring the data further, I filter the dataset and remove two-stage responses that indicate less than 50% confidence in the selected answer and compare peak confidence by difficulty level against the one-stage responses. I planned to run a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on peak confidence with the restricted dataset. The analysis yields no significant main effect of condition, *F*(1, 682) = 1.83, *p* = .176, and no interaction between condition and difficulty, *F*(2, 390) = 0.15, *p* = .859. This suggests that the responses in the two-stage condition under fifty percent are driving the results seen in the main analysis depicted in Figure 12.

Breaking results down further, I find that the observed pattern stays the same directionally, and is only significant for the medium difficulty question, but barely so. Figure 16 depicts these results.

1. For the easy questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 79.46, *SD* = 15.75) and the two-stage condition (*M* = 78.30, *SD* = 14.71), *t*(667) = 0.98, *p* = .326.
2. For the medium questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 70.46, *SD* = 15.96) and the two-stage condition (*M* = 68.11, *SD* = 14.89), *t*(646.9) = 1.95, *p* = .052.
3. For the hard questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 65.70, *SD* = 14.81) and the two-stage condition (*M* = 64.59, *SD* = 13.91), *t*(610.26) = 0.96, *p* = .338.

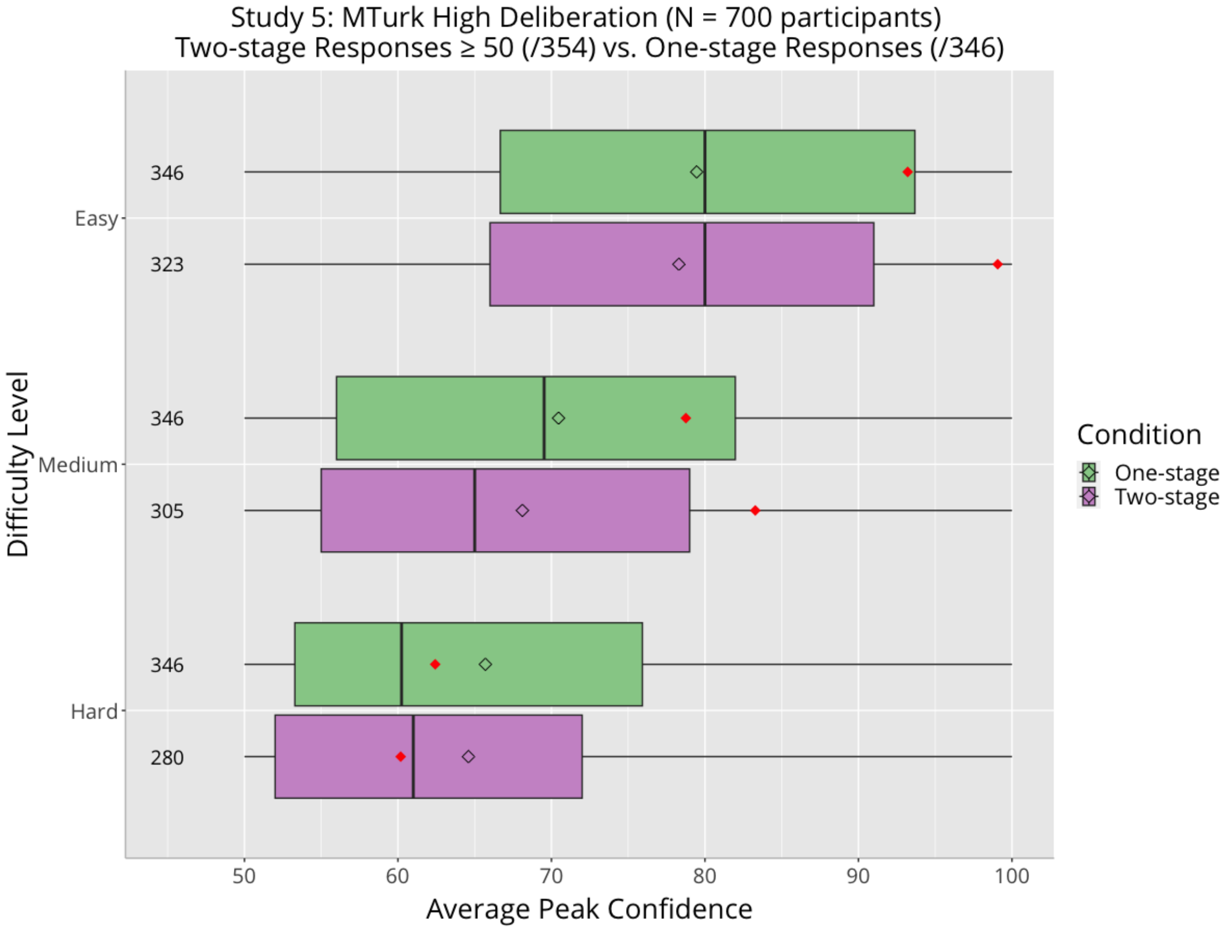


Figure 16. (Study 5) Boxplot showing average peak confidence by question difficulty level and condition, using filtered dataset. Red diamonds indicate accuracy, transparent diamonds indicate mean.

3) Pre-registered exploratory analysis: Exploring deliberation as a mediator

To explore the effect of deliberation, I pre-registered a plan to run an exploratory 2 (condition: one-stage, two-stage) x 2 (deliberation: low, high) x 3 (difficulty: easy, medium, hard) mixed ANOVA. My theory would predict a three-way interaction between condition (one-stage vs two-stage), deliberation (low vs high), and difficulty (easy, medium, hard). For those low in deliberation, the prediction was that regardless of difficulty level, there would be a significant difference between one-stage and two-stage, such that participants in the one-stage condition would report higher average peak confidence than those in the two-stage condition. For those high in deliberation, the prediction was that difficulty level would not matter – specifically that there would be a significant difference in peak confidence between one-stage and two-stage conditions only for the difficult questions, such that participants in the one-stage condition would report higher peak confidence than in the two stage condition only for the difficult questions. I noted in my pre-registration that the study may be underpowered to detect this three-way interaction, which is why the critical analysis focused on the high deliberation sample only.

Due to the issues with recruitment, the final sample size is indeed too low. The following analyses are therefore entirely exploratory. Before running the full model, I first ran a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on the low deliberation dataset only. There is no significant main effect of condition, $F(1, 160) = 0.43$, $p = .513$. There is a significant main effect of difficulty on peak confidence, $F(2, 320) = 26.61$, $p < .001$, which does provide some assurance that participants were not simply answering randomly. There is no interaction between condition and difficulty, $F(2, 320) = 0.07$, $p = .993$. Figure 17 below depicts peak confidence as a function of condition and difficulty, with diamonds indicating average accuracy. If anything, this result runs contrary to what I may have expected.

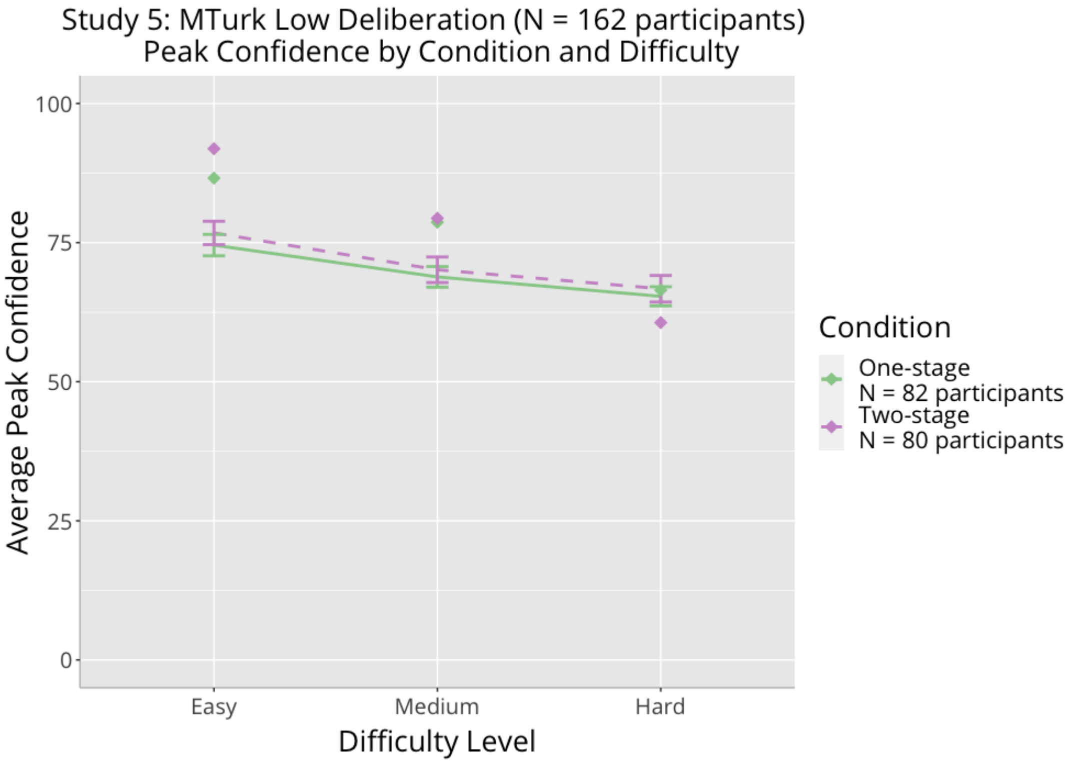


Figure 17. (Study 5 Low) Average peak confidence by condition and question difficulty level. Diamonds indicate average accuracy.

I also run the 2 (condition: one-stage, two-stage) x 2 (deliberation: low, high) x 3 (difficulty: easy, medium, hard) mixed ANOVA. I break the results down by main effects and interactions below. Figure 18 below depicts peak confidence as a function of condition, difficulty, and deliberation with colors indicating condition, and shapes indicating deliberation level.

Main Effects

- There is a significant main effect of condition on peak confidence, *F*(1, 858) = 29.51, *p* < .001. This suggests that peak confidence significantly differs between the one-stage and two-stage conditions.
- The main effect of deliberation on peak confidence is not significant, *F*(1, 858) = 2.39, *p* = .123, indicating that peak confidence does not significantly differ between low and high deliberation samples.
- There is a significant main effect of difficulty on peak confidence, *F*(2, 1716) = 280.55, *p* < .001. This means that the difficulty level significantly impacts peak confidence.

Interaction Effects

- Condition × Deliberation: There is a significant interaction between condition and deliberation, *F*(1, 858) = 11.53, p < .001, suggesting that the effect of condition on peak confidence varies depending on the level of deliberation.
- Condition × Difficulty: The interaction between condition and difficulty is not statistically significant, *F*(2, 1716) = 2.86, *p* = .057, implying that the effect of condition on peak confidence does not significantly change across different difficulty levels.

- Deliberation × Difficulty: There is a significant interaction between deliberation and difficulty, *F*(2, 1716) = 7.03, *p* < .001, indicating that the impact of deliberation on peak confidence varies across different levels of difficulty.
- Condition × Deliberation × Difficulty: The three-way interaction is not significant, *F*(2, 1716) = 0.29, *p* = .748, suggesting that the combined effect of condition, deliberation, and difficulty on peak confidence does not differ significantly across the levels tested.

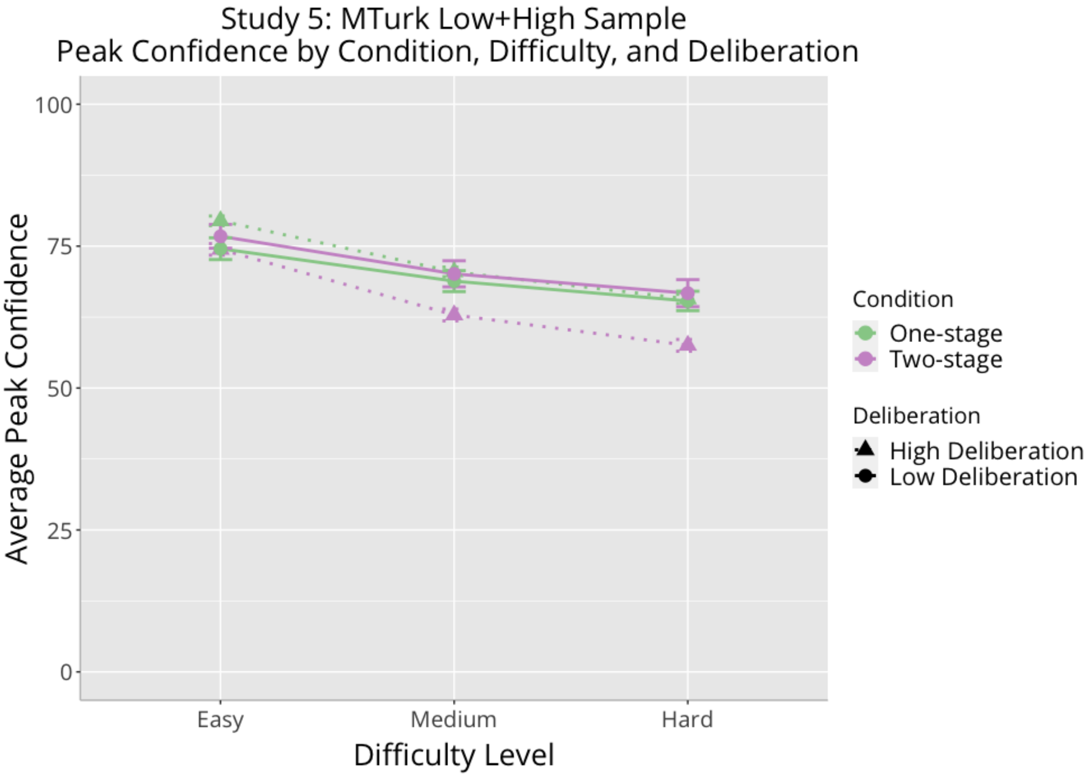


Figure 18.(Study 5 Low and High) Average peak confidence by condition and question difficulty level. Colors indicate condition, and shapes indicate deliberation level.

Overall, I find that condition and question difficulty significantly affect peak confidence, but level of deliberation does not. The interaction between condition and deliberation indicates that the influence of the experimental condition on confidence is affected by how much individuals deliberate. The relationship between deliberation and difficulty also changes across different levels, highlighting that the way deliberation impacts confidence depends on the task's difficulty. In sum, while both condition and question difficulty independently influence peak confidence, the effect of deliberation on confidence is more complex and depends on both the condition and the difficulty of the tasks. I note these results are entirely exploratory and may not replicate.

4) Exploratory analysis: Coin flip

Given the issues with reliability of the low deliberation sample, I look at the coin flip question using only the high deliberation sample. Because the data is not normally distributed, I use a Wilcoxon rank sum test to compare peak confidence between the two conditions. The results do not indicate a significant difference in peak confidence between the two conditions (one-stage *M*

= 52.24, two-stage *M* = 54.80), W = 58457, p = 0175. Figure 19 graphs the counts of each answer by condition.

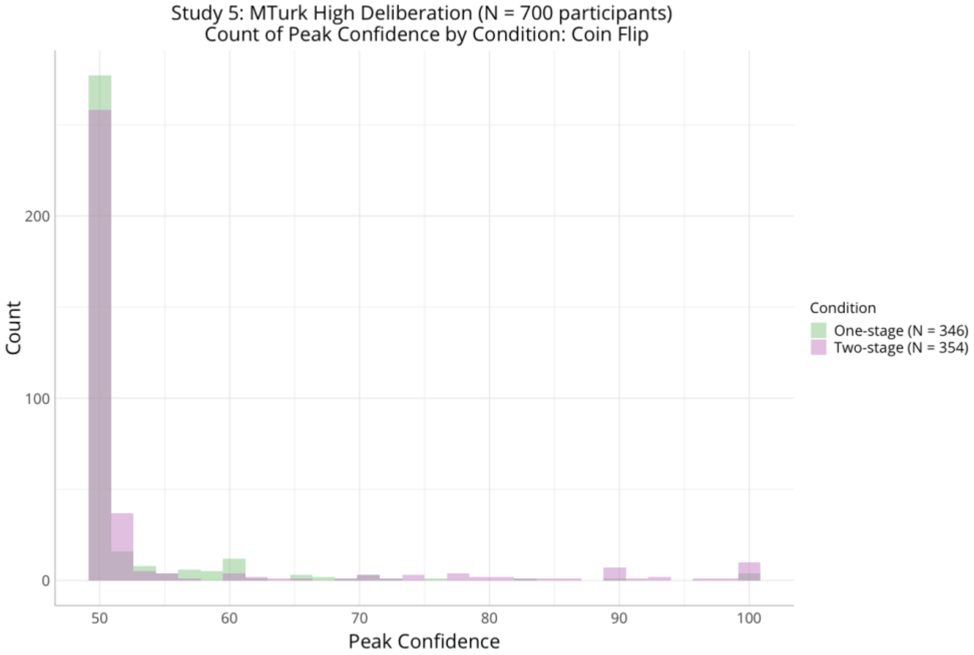


Figure 19. (Study 5 Coin) Count of peak confidence by condition.

Note again that peak confidence is the highest confidence reported in the one-stage condition, but the implied highest confidence in the two-stage condition (e.g., if a participant chose Tails and said 25%, peak confidence would reflect 75% in the data). Figure 20 plots the counts with the reported confidence in the two-stage condition, rather than implied peak confidence.

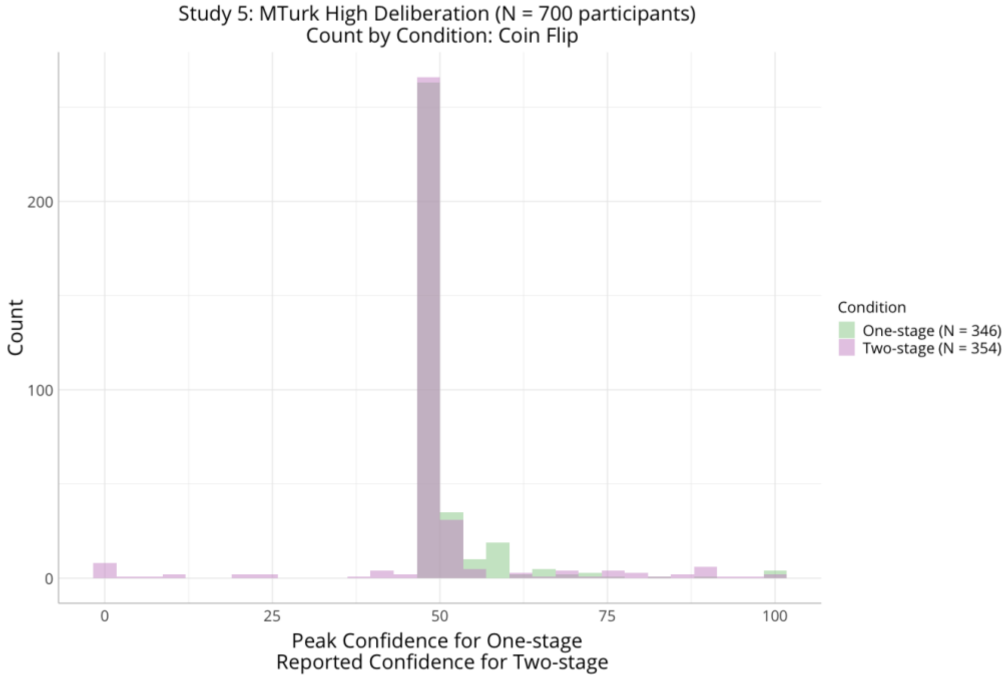


Figure 20. (Study 5 Coin) Count of peak confidence for the one-stage condition and reported confidence for the two-stage condition.

5) Other Notes

I asked all participants two open-ended questions at the end of the survey. The first was on what they thought the survey was asking, and the second was on whether they found anything in the survey weird or confusing.

In both the high and low deliberation samples, all participants generally reported the survey was asking about dots, perception, probability, confidence, image, perceptions, and so on. The vast majority of responses indicated nothing about the survey was weird or confusing.

I also asked three attention check questions, hoping to use them as another measure of deliberation. However, the number of people who did not pass the attention check questions proved quite low in both samples, so the questions could not provide another operationalization of deliberation as I had hopes.

Discussion

Study 5 builds upon the results of Study 4 by examining whether a one-stage elicitation leads to higher average peak confidence than a two-stage elicitation for hard versus easy questions, and whether the amount participants deliberate makes a difference. The exploratory studies hinted that undergraduates may simply be racing through the survey as compared to the best Amazon Mechanical Turk workers. There is research that suggests undergraduate data quality is of lesser quality than MTurk, and that setting Cloud Research and MTurk requirements also significantly improves data quality (Douglas et al., 2023; Zhang et al., 2024).

The low deliberation sample proved more difficult to obtain than originally anticipated. Although I explored the results of the full model including the deliberation variable, I note the exploratory nature of the findings, and do not put too much stock in results. However, Study 5 provides the first fully powered test of the main question: Is there a difference in peak confidence by elicitation format, and does question difficulty impact this relationship? I find there is indeed a main effect of condition, such that average peak confidence is higher in the one-stage condition than in the two-stage condition. I also observe an interaction between condition and difficulty, such that the difference is larger for harder questions.

Study 6: Do participants prefer one elicitation over the other, and if so, does this change peak confidence and accuracy?

Study 6 examines familiarity as a potential mediator. I surveyed which confidence elicitation – the one-stage or the two-stage – participants preferred, if either, and tested how they performed under their preferred elicitation.

Method

Participants

The final sample includes 600 Cloud Research approved Amazon Mechanical Turk participants (278 Female, 309 Male, 5 Gender neutral; mean age = 45.39; 441 White, 58 Black or African American, 47 Asian, 36 Hispanic or Latino, 2 American Indian or Alaska Native, 1 Native Hawaiian or Other Pacific Islander and 1 Other). I filtered out anyone whose recorded Qualtrics progress was less than 95% upon receiving the data. 137 participants were in the one-stage-pref condition, 50 participants were in the one-stage-nopref condition, 366 participants were in the two-stage-pref condition, and 49 participants were in the two-stage-nopref condition.

Design

The experiment has a four-cell between-subjects design. Participants first saw an explanation for what the one-stage and two-stage confidence elicitation formats were and then ran through two examples of each. Then, they were asked which elicitation they preferred, if either. The explanation order was randomized. In the prefer-one-stage confidence elicitation condition, participants distributed their confidence over two answer choices, with one slider scale per answer choice. In the prefer-two-stage confidence elicitation condition, participants first chose which option they believed was most likely to be the correct answer, and then indicated how confident (0-100% slider scale) they were their selected answer was indeed correct. In the no-preference confidence elicitation condition, participants were randomized into seeing questions in either the one-stage elicitation, or the two-stage elicitation.

Procedure and Materials

I used the same dot task as in prior studies. Participants in both conditions saw three pairs of dot images (easy, medium, and hard) in random order.

Results

1) Descriptives

Over half the participants indicated a preference for the two-stage elicitation format (N = 366 participants). 137 participants indicated a preference for the one-stage elicitation format, and 99 participants indicated they had no preference between the two formats.

366 participants were placed in the two-stage-pref condition and 137 were placed in the one-stage-pref condition. Of the 99 participants who indicated having no preference, 50 participants were randomized in the one-stage-nopref condition, and 49 participants were randomized in the two-stage-nopref condition.

Participants who indicated having a preference were then asked on the following page why they preferred being asked questions in the format they selected. Participants were asked two additional questions on norms in daily life vs. on MTurk. When asked “Which confidence elicitation format do you encounter more in daily life?” 384 participants (63.89%) selected the two-stage format, and 217 (36.11%) selected the one-stage format. When asked “Which confidence elicitation format do you encounter more on MTurk?” the numbers did not differ much: 387 participants (64.29%) selected the two-stage format, and 215 (35.71%) selected the one-stage format.

Those who preferred the two-stage format gave a variety of reasons as to why – I attempt to summarize the most frequent reasons below and provide select quotes from participants. Because the vast majority of participants preferred the two-stage condition, there are many more quotes than appear here. It would be interesting to do a deeper analysis of the open-ended responses in the future. What stood out to me was that participants found the two-stage format simpler – my original thought before running the study was that participants, especially on MTurk, might choose the one-stage format simply because it is one less step. A number of participants noted how the two-stage format seems more natural, and how it might actually reflect true preferences better.

It is easier to understand (e.g., more intuitive) and simpler to do. The one-stage elicitation format seems redundant.

- I don't want to break down how confident I am for every item.
- I think it is simpler to understand and requires less work.
- It eliminates one slider bar I have to set.
- I am asked to make the choice first then give my confidence in my answer. It lets me think about it first and then list how strong I feel instead of having to do two decisions at once.
- From an end user's perspective, it's easier to deal with 1 slider instead of 2.
- It is a more direct format; I think it is more clear and easy to understand
- It seems more intuitive that way.
- It seems like it is structured in a more logical way
- I think it is easier to use. I like just answering the question and then giving my confidence answer.
- It's simpler and a more efficient way to communicate confidence of my answer choice.
- The format is easier and less confusing and I have to read less.

It better captures participants’ preferences

- I get to pick which i think is correct which already seems more confident
- I feel like the two stages allow me to really think about my answer and possibly refine it. There is just an extra step that encourages more thought about the question.

- It gives me a black and white option to choose and then I can decide how confident I am with the option. It gives me more control and less uncertainty around what I choose.
- Just makes more logical sense to me; I'd rather have to put my thought into which answer is correct rather than having to decide which I'm more confident in, the latter of which is slightly more work.
- I am going to feel strongly that one answer is correct so I would like to choose my answer right away.
- I think it gives a better idea overall of what I think is true.
- I prefer being asked int this format because its somewhat more direct so I feel more confident in my answers. The other one-stage confidence elicitation format seems more broad and vague so I prefer two-stage where I feel more firm with my answers.
- It seems more direct and consistent. It requires less thought outside of initial approximation.
- I think it presents a answering format more consistent with the thought process.
- If I already have a preference for one answer choice over the other, I don't really want to be asked about the answer choice that I don't think is correct. I'd rather be asked further about the answer choice I picked.
- Because it gives me a limitation in the process, feels more structured, more of a psychological thing really.
- It seems more natural to my thinking process. I choose which answer I feel is correct - and then give a percentage estimate of my confidence. The other process - it is simple math - if one is (say) 80%, the other is 20% by definition. Why make me go thru that? When you already know the answer if I give a percentage (say 80%) to one of them...

It improves the quality and precision of the answer

- I like having to choose the answer directly. And, then determining how confident I am. I feel like there is less room for error.
- I think its more to the point, asking someone to distribute percentages across a set of data might lead to more overthinking rather than a real \"guess\".
- It better separates the decision from the assessment of confidence, making it easier to produce a good answer
- The fact that you chose a particular answer already skews confidence in favor of that that option & the next stage fine tunes it further.

It makes me more confident in my selection

- I think it's better to have an affirmative answer at the beginning and then test confidence. You don't waffle on the decision because it's already made.
- Not quite sure. But I think it helps me be more certain with my answer.
- I prefer two-stage confidence elicitation because i would only have to answer one-time and never re-evaluate my first answer thus building confidence.

It makes me second-guess my selection

- Because it gives me a little more time to think carefully about my first choice.
- It gives you a chance to seconds guess it in a sense.

The one-stage format seems more complicated

- Seeing the scale next to each option doesn't allow me to focus on comparing and contrasting the two options as much.
- Being given all the options is a bit overwhelming.

There also were participants who very clearly understood the two-stage format:

- It seems the most straight-forward, obviously if I think my answer is 80% correct then I think the other is 20% correct. It seems redundant to have to enter all that.
- I feel like which I prefer semi depends on the questions being asked and how many options there are, but especially if there are two options I think this says more about what I'm thinking and it's also mechanically better because sliders are a pain to deal with.
- It seems the same as one-state confidence elicitation except I am giving confidence points to both answers. In two-stage I feel like I am giving confidence points to both answers by selecting the one - the other option would be the difference. Since I view these as the same, the two-state is easier.
- One reason is less math, ha ha. It is just less complicated, I of course don't really know the exact percentages of my feelings and if there are only two different options then it is just a waste of time to have to choose percentages for both when the answer I think is wrong would just be 100 minus the percentage I chose for the answer I thought was right.

It’s the norm

- Because I don't know, I'll guess that's the approach most people have taken with me, which is why I prefer it. Habit.
- I'm not sure, it was the easiest to understand and was in the format that is normally used in school and on tests.

Those who preferred the one-stage format also gave a variety of reasons as to why – I attempted to summarize the most frequent reasons below, and provide select quotes from participants. What stood out to me was that the reasons overlapped with the ones that participants who preferred the two-stage format gave. The two reasons that I saw as being more different concerned how the one-stage format might better capture uncertainty, and feels like it provides more detailed information.

The one-stage format is easier to understand

- i think its easier to understand
- It is easier to think through
- I feel like I only have to answer one item instead of two

The one-stage format is faster

- To get the questions answered quicker
- Less questions to answer
- It is quicker to give the same answer
- eh, its shorter and im lazy
- It is quicker, more efficient and still gets the same result.

I like the one-stage format better

- I like being able to put a percentage of confidence for each answer.
- Even though the probability percentage of the unchosen answer is implied in the two-stage confidence elicitation by the degree of confidence not stated in the second stage, I feel better seeing both probabilities stated numerically.
- I like how compact the one-stage confidence elicitation format it. It asks the question in a way in which you present your level of confidence in both responses. The two-stage confidence elicitation just feels very redundant to me, where as the one-stage feels more streamline in answering questions while still getting the same quality of response.

It better captures participant preferences

- The format helps me organize my thinking and likely provides a more accurate reflection of my uncertainty. If I have to select one of two options first, I will probably overestimate my confidence in that choice because the act of choosing itself has an effect.
- It provides options to answer for all possible answers and to show how much more I prefer one answer over another
- I am able to more clearly show which item I am most confident in being correct, but also clearly shows the alternative option and the odds I believe that will occur.
- I think it makes more sense in a visual way to say whether or not the answer is one way or the other with the amount of confidence at the same time. It eliminates an extra step.

It better captures uncertainty

- Because if I'm somewhat split on the answer it's easier to convey that, especially if there are more than two choices and I'm split amongst all of them.
- I would not be so sure about my answers. Given the option to give a percentage between the options makes me feel like the answer I chose is more likely to be correct while acknowledging I could be wrong and the other answer correct.

It's better than the two-stage format in the level of detail, and that helps in decision making

- It feels more detailed versus the two-stage confidence elicitation format.
- I like that it's more nuanced. Most answers that are opinion-based are not black and white, yes/no answers.
- I like having options for answers, it helps make a more informed decision.
- Because, often we meet situations where there can be multiple answers possible in our life. If the questions has a definite answer, I still prefer to consider the possibility of the other case.
- You have the opportunity to give a likelihood for all the answer options as to just one. By establishing a probability for all answers I'm more likely to arrive at the correct one.

2) Pre-registered analysis: ANOVA models

To test if peak confidence is significantly different among the four groups, I planned to run a one-way analysis of variance (ANOVA). The four groups in this study are:

1. One-stage-pref: Those who prefer one-stage format
2. One-stage-nopref: Those with no preference who are randomized into one-stage format
3. Two-stage-pref: Those who prefer two-stage format
4. Two-stage-nopref: Those with no preference who are randomized into two-stage format

Following the ANOVA, I planned to use post-hoc tests to determine which groups significantly differed (Scheffé's Test).

Model 1
In running the analyses, I realized my pre-registered analyses were not sufficiently clear – I should have specified the exact models I intended to run. Based on what was written, I ran two models. The first looked at whether peak confidence significantly differed among the four groups. A one-way repeated measures ANOVA was conducted to examine the impact of condition on peak confidence, with participants treated as a subject factor to account for repeated measures. The analysis revealed a significant effect of condition on peak confidence, *F*(3, 598) = 3.58, *p* = .014. This suggests that peak confidence significantly differs across the four conditions.

Following the ANOVA, I planned to use post-hoc tests to determine which groups significantly differed (Scheffé's Test). Note that Scheffé's Test uses the simplified model without the PID repeated measures factor. Scheffé's test was used to control for Type I error across multiple comparisons. The mean square error was 304.77 with 1802 degrees of freedom, with a computed F-value of 2.61. The Scheffé's test, applied at a 0.05 level of significance, indicated differences in peak confidence levels among the groups.

The means for peak confidence were as follows: "one-stage-nopref" at 67.48 (*SD* = 15.82, N = 150), "one-stage-pref" at 68.13 (*SD* = 15.29, N = 411), "two-stage-nopref" at 62.12 (*SD* = 20.55, N = 147), and "two-stage-pref" at 64.58 (*SD* = 17.97, N = 1098).

According to Scheffé's groupings, "one-stage-pref" (*M* = 68.13) and "one-stage-nopref" (*M* = 67.48) did not significantly differ from each other as indicated by their shared group labels (a and ab), suggesting similar levels of peak confidence in these conditions. However, both "two-stage-pref" (*M* = 64.58) and "two-stage-nopref" (*M* = 62.12) were significantly different from "one-stage-pref" but not from each other, as denoted by their group label (b). These results suggest that the one-stage conditions tend to have higher peak confidence compared to the two-stage conditions. Figure 21 below depicts peak confidence as a function of condition, with diamonds indicating average accuracy.

I also conducted Scheffé's test to look at whether accuracy differs across conditions. The mean square error was 1421.163 with 1802 degrees of freedom, and a computed F-value of 2.61, suggesting variability in hit rates across conditions. Despite this variability, the Scheffé's test, conducted at an alpha level of 0.05, did not distinguish separate groups among the conditions, indicating that the differences in accuracy were not statistically significant at that threshold.

Mean accuracy was as follows: "one-stage-nopref" at 76.33% (*SD* = 37.85, N = 150), "one-stage-pref" at 76.40% (*SD* = 39.39, N = 411), "two-stage-nopref" at 80.61% (*SD* = 37.22, N = 147), and "two-stage-pref" at 81.19% (*SD* = 37.09, N = 1098). Although the hit rates for the two-stage conditions trended higher, Scheffé's group analysis classified all conditions into the same group (a), indicating no significant difference in hit rates among the conditions tested.

The consistent grouping across all conditions suggests that, despite the observed variances and the marginal differences in mean hit rates, the experimental conditions did not significantly affect accuracy.

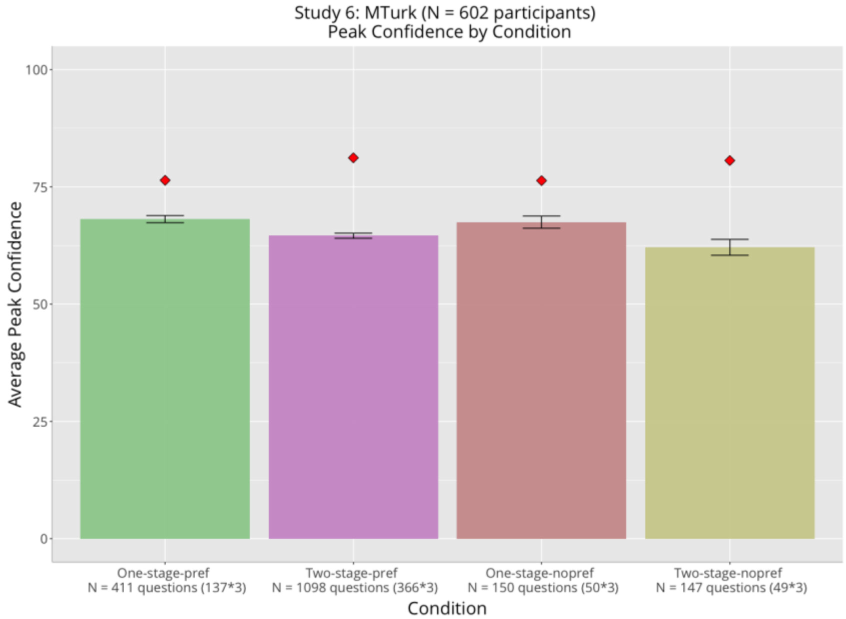


Figure 21. (Study 6) Average peak confidence by condition. Red diamonds indicate average accuracy.

Model 2

The second model tested the hypothesis that I specified in my pre-registration: I predicted an interaction between condition and difficulty, such that average peak confidence in the one-stage condition will be higher than average peak confidence in the two-stage condition, but only for the difficult questions.

The model yields a significant main effect of condition, *F*(3, 598) = 3.58, *p* = .014. There is of course a significant main effect of difficulty on peak confidence, *F*(2, 1196) = 173.33, *p* < .001. I also observe the expected interaction between condition and difficulty, *F*(6, 1196) = 2.29, *p* = .033. Figure 22 below depicts peak confidence as a function of condition and difficulty, with diamonds indicating average accuracy.

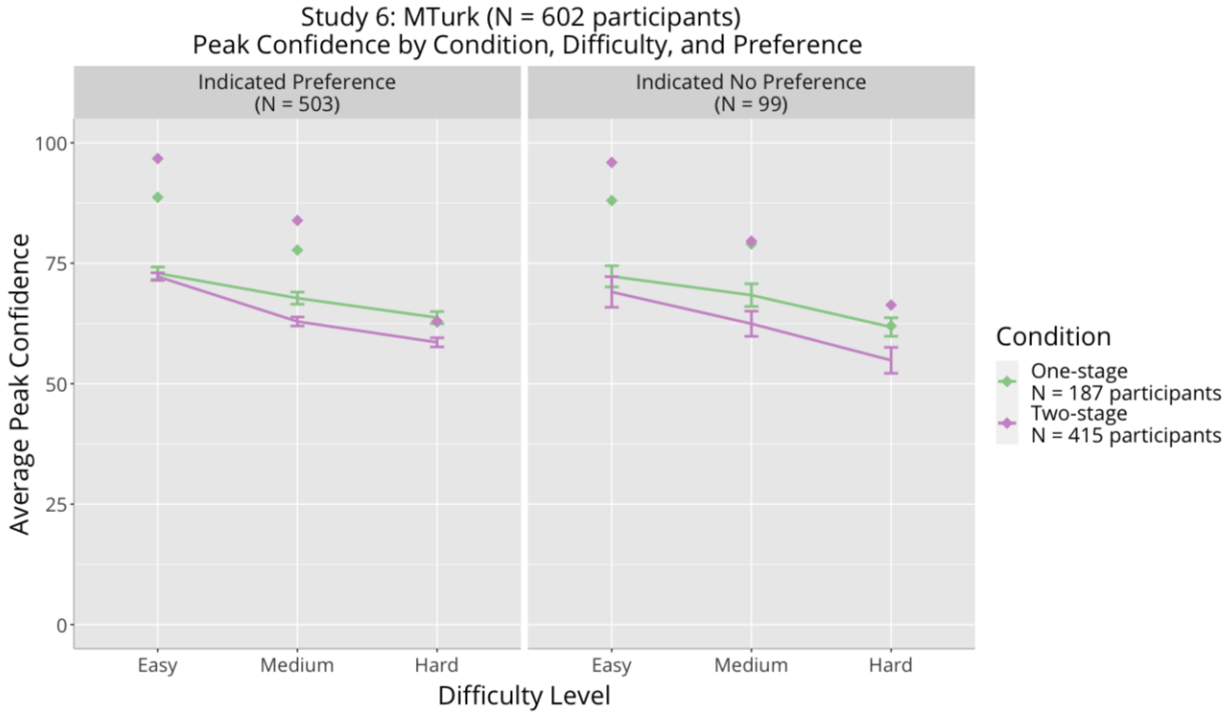


Figure 22. (Study 6) Average peak confidence by condition, question difficulty level, and whether or not participants indicated a preference between the two elicitations. Diamonds indicate average accuracy.

3) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition

I examine whether participants violate rationality in the two-stage-pref condition. I choose to focus on this rather than the two-stage-nopref condition given the sample size is much higher. I find that across all tasks and difficulty levels that 135 out of 1098 responses (12.30%) had less than 50 percent confidence in the answer selected in the first stage. Figure 23 depicts average peak confidence for each difficulty level, broken down by those who answered less than 50% confidence and those who answered more.

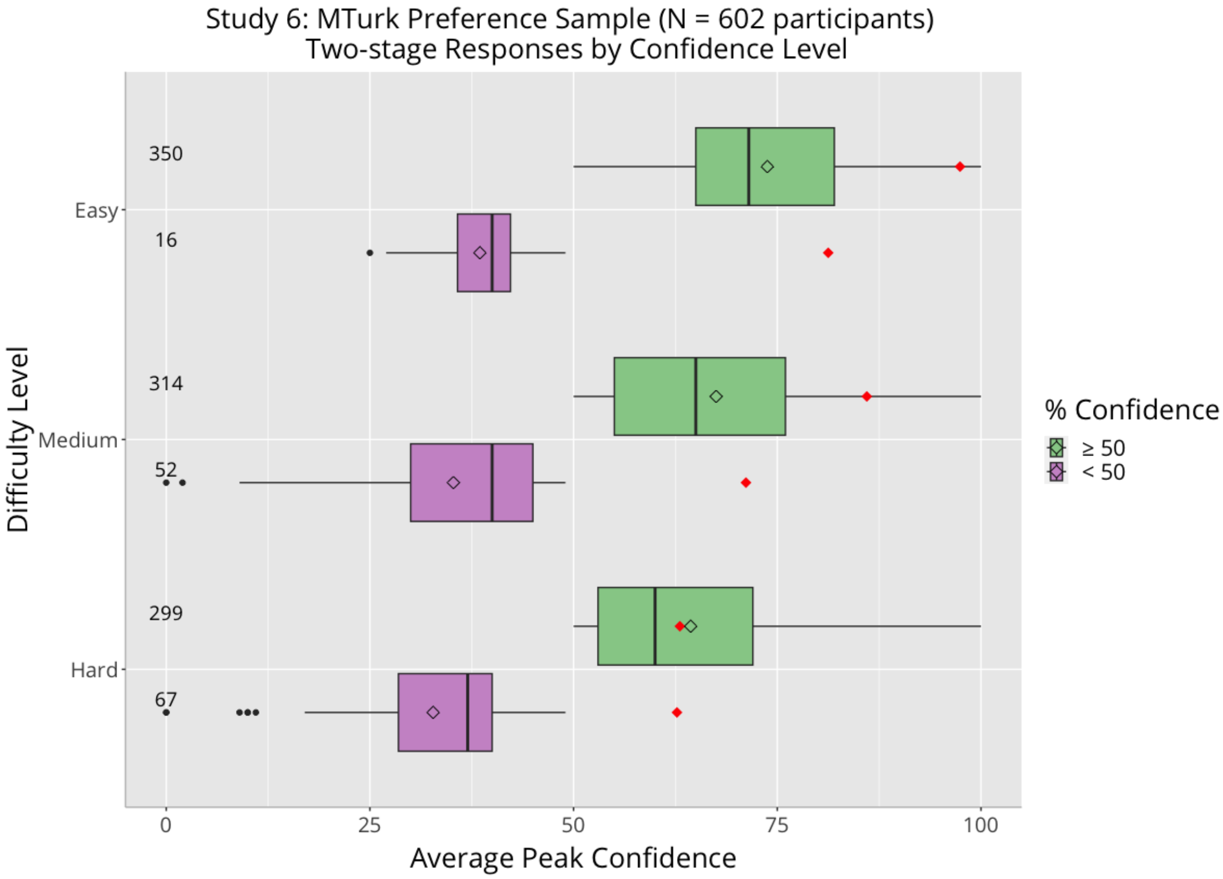


Figure 23. (Study 6 Preference) Average peak confidence by question difficulty and confidence level. Red diamonds indicate accuracy, transparent diamonds indicate mean. Number of responses per confidence level listed on the lefthand side.

The percentage of responses indicating less than 50% confidence differs by difficulty level, with 18.31% of responses indicating less than 50% confidence for hard questions in the two-stage condition, 14.21% for medium, and 4.37% for easy (see Table 3 below).

	Question Difficulty Level		
	Easy	Medium	Hard
< 50% confidence	16 responses	52 responses	67 responses
≥ 50% confidence	350 responses	315 responses	299 responses

Table 3. (Study 6 Preference) Number of Responses by Task and Confidence Level

Exploring the data further, I filter the dataset and remove two-stage responses that indicate less than 50% confidence in the selected answer and compare peak confidence by difficulty level against the one-stage responses. I planned to run a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on peak confidence with the restricted dataset. The analysis yields no significant main effect of condition, $F(1, 491) = 0.35$, $p = .556$, and no interaction between condition and difficulty, $F(2, 875) = 0.77$, $p = .465$. This suggests that the responses in the two-stage condition under fifty percent are driving the results seen in the main analysis depicted in Figure 22. Figure 24 depicts average peak confidence by question difficulty and condition, filtering the data for two-stage responses that more than or equal to 50.

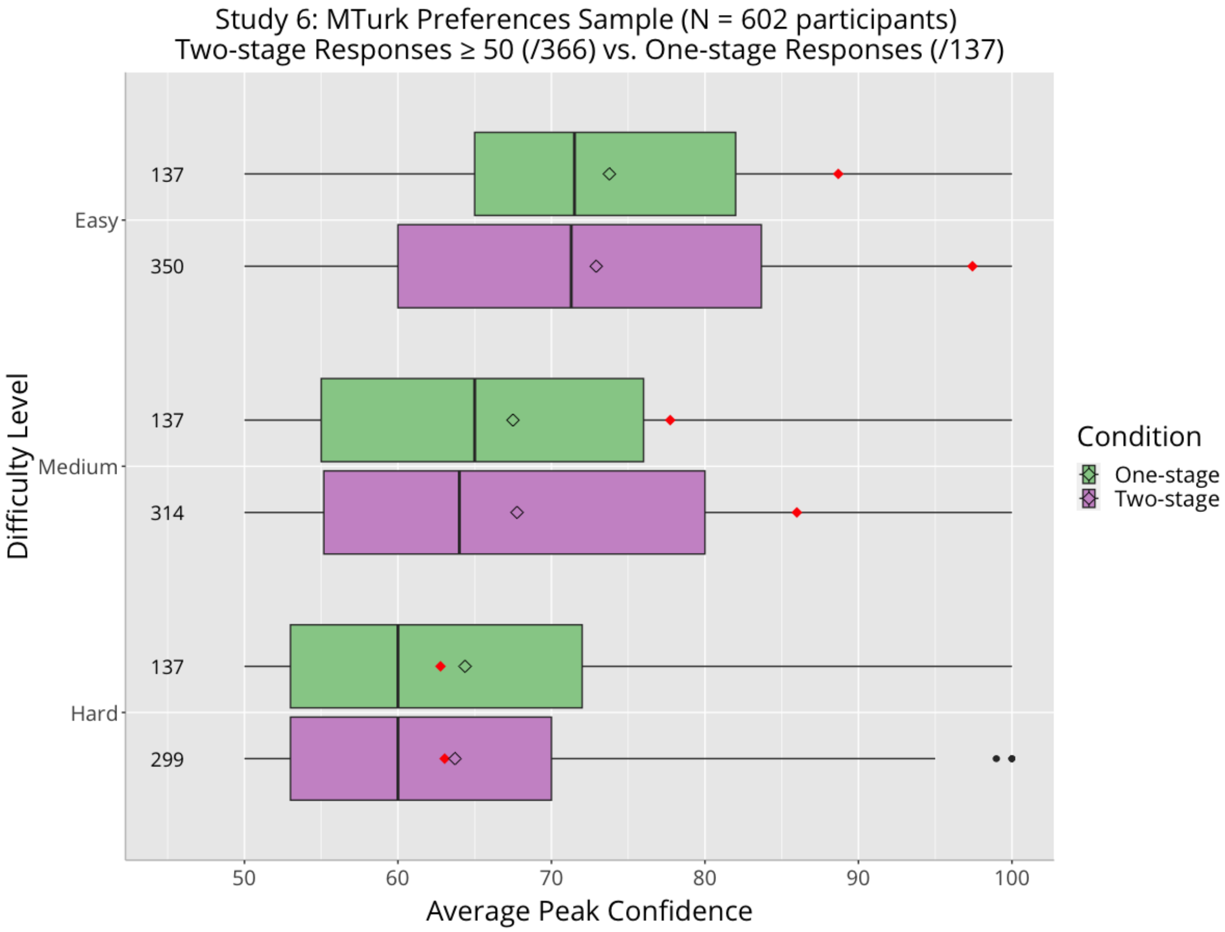


Figure 24. (Study 6 Preference) Boxplot showing average peak confidence by question difficulty level and condition, using filtered dataset. Red diamonds indicate accuracy, transparent diamonds indicate mean.

4) Other Notes

I asked all participants two open-ended questions at the end of the survey. The first was on what they thought the survey was asking, and the second was on whether they found anything in the survey weird or confusing.

In both conditions, all participants generally reported the survey asked about preferences on confidence elicitations, indicating the survey was indeed straightforward. The vast majority of responses indicated finding nothing about the survey weird or confusing. A few participants found the attention check questions a bit strangely worded, which might explain why a number of participants failed them – I will make a note on this for future studies.

Discussion

Study 6 examines which type of confidence elicitation – the one-stage or the two-stage – participants report being most familiar with, and whether familiarity with the elicitation used may lead to different results. Interestingly and perhaps unsurprisingly, participants report being most familiar with the two-stage elicitation format, encountering it more on MTurk and in daily life, and preferring to answer questions under that elicitation.

Why might preferences for a given elicitation matter? Presumably in selecting the two-stage elicitation format, participants are implicitly signaling that they understand it – this might lead to less violations of rationality, which would bring average peak confidence up. My results show that this is not the case – even when participants are able to select which elicitation they prefer, there is still a significant difference in average peak confidence in the two conditions, and a significant interaction effect of condition and question difficulty.

Those who indicated no preference in either elicitation provided a nice ground for a small replication – even with the limited sample size, I find that the effect from prior studies replicate. When participants are randomized into one elicitation format, I observe that condition effect and condition x difficulty interaction.

I note that I do not find a difference in average peak confidence when I compare those who selected a given elicitation format, and those who were randomized into it. In other words, the results suggest that the one-stage conditions tend to have higher peak confidence compared to the two-stage conditions, but that average peak confidence does not differ between the one-stage-pref and one-stage-nopref, and between the two-stage-pref and two-stage-nopref. Notably, accuracy did not differ across any of the conditions – participants were not more accurate when say, they were able to choose their elicitation format.

Study 7: Does the effect replicate when participant choices are incentive compatible, and when different stimuli are used?

Throughout the six studies, I used the same dot stimuli to enhance the comparability of findings. It allowed the various studies to build upon one another, increased the comparability of results, and provided more assurance that the effect I was seeing from study to study was real and robust. Study 7 aimed to see whether the effect would replicate with different stimuli. Critically, the study also employs incentive compatible payoffs, which raised the stakes on accurate responding.

Method

Participants

The final high-deliberation sample includes 968 Cloud Research approved Amazon Mechanical Turk participants (472 Female, 484 Male, 5 Gender neutral; mean age = 51.28; 751 White, 70 Black or African American, 84 Asian, 36 Hispanic or Latino, 4 American Indian or Alaska Native, 2 Native Hawaiian or Other Pacific Islander and 15 Other). I filtered out anyone whose recorded Qualtrics progress was less than 95% upon receiving the data. As planned, I also filtered out anyone who did not pass the two comprehension checks given right after the instructions – this amounted to 37 participants in practice. In the end, 489 participants were in the one-stage condition, and 479 participants were in the two-stage condition.

Design

The experiment had a two-cell between-subjects (one-stage condition vs. two-stage condition) design that manipulated how confidence was elicited. Decisions were incentive compatible. The survey randomly assigned participants to the two between-subjects conditions and presented tasks in randomized order within each condition.

Procedure and Materials

Participants in both conditions read that the questions in the survey would be scored using the quadratic scoring rule, and that they could maximize their score by honestly and accurately reporting their confidence. Their score earned them lottery tickets into a lottery for a $50 bonus. Participants who wanted to see more details on the QSR had the option of clicking a link that explained the payoff equation. QSR scores ranged 0 to 2. I divided by 2 and multiplied by 100, so that for each question, payoffs ranged from 0 to 100 lottery tickets.

The format of the conditions followed prior studies. Instead of the dots task, I used the faces task piloted in the first exploratory study but dropped the medium difficulty question. The faces task showed a pair of images of faces and asked participants to indicate which face was older. Participants in both conditions saw two pairs of images (easy and hard).

Results

1) Pre-registered analysis: ANOVA Model

I planned to run a 2 (condition) x 2 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on peak confidence. I predicted an interaction between condition and difficulty, such that average peak confidence in the one-stage condition would be higher than average peak confidence in the two-stage condition, but only for the difficult question.

There is a significant main effect of condition, $F(1, 966) = 28.99$, $p < .001$. There is of course a significant main effect of difficulty on peak confidence, $F(1, 966) = 1484.23$, $p < .001$. I also observe an interaction between condition and difficulty, $F(1, 966) = 14.02$, $p < .002$. Figure 25 below depicts peak confidence as a function of condition and difficulty, with diamonds indicating average accuracy.

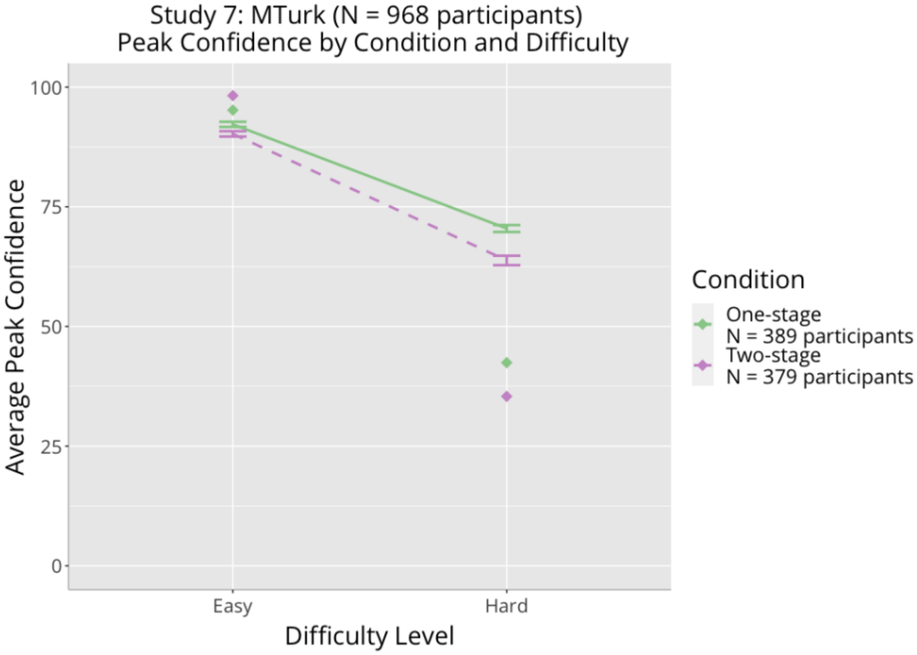


Figure 25. (Study 7) Average peak confidence by condition and question difficulty level. Diamonds indicate average accuracy.

Collapsing across difficulty levels, I find that average peak confidence is 81.35% (*SD* = 17.74) in the one-stage condition, and 77.01% (*SD* = 21.96) in the two-stage condition. Average accuracy does not differ between conditions (one-stage $M = 68.81$, two-stage $M = 66.81$), $t(1928) = 0.97$, $p = 0.333$). Figure 26 depicts average peak confidence in each condition, with average accuracy represented by the red diamonds.

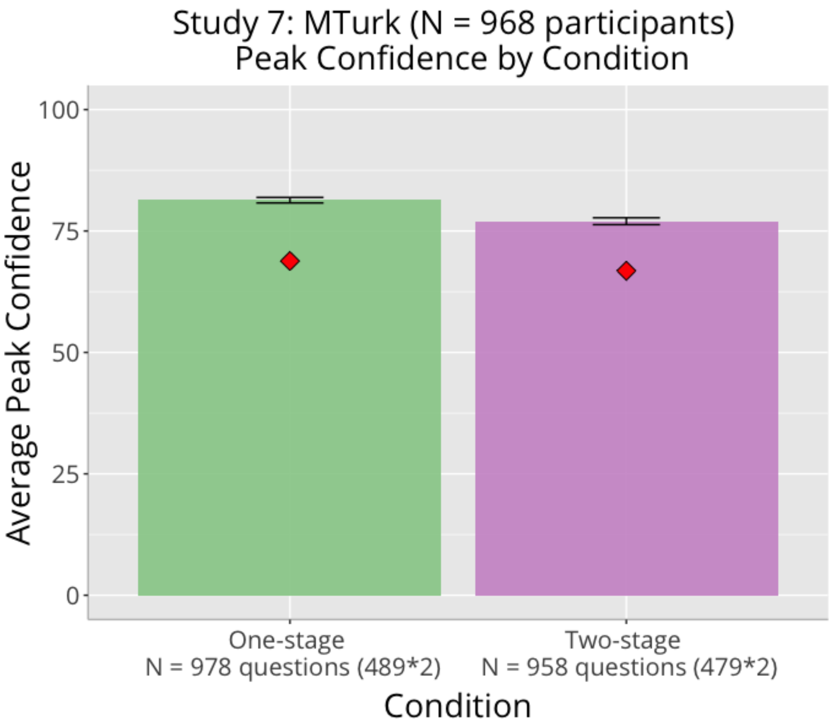


Figure 26. (Study 7) Average Peak Confidence by Condition. Red diamonds indicate average accuracy.

2) Pre-registered exploratory analysis: Deviations from rationality in the two-stage condition

I examine whether participants violate rationality in the two-stage condition. I find that across all tasks and difficulty levels, 76 out of 968 responses (7.85%) had less than 50 percent confidence in the answer selected in the first stage. Figure 27 depicts average peak confidence for each difficulty level, broken down by those who answered less than 50% confidence and those who answered more.

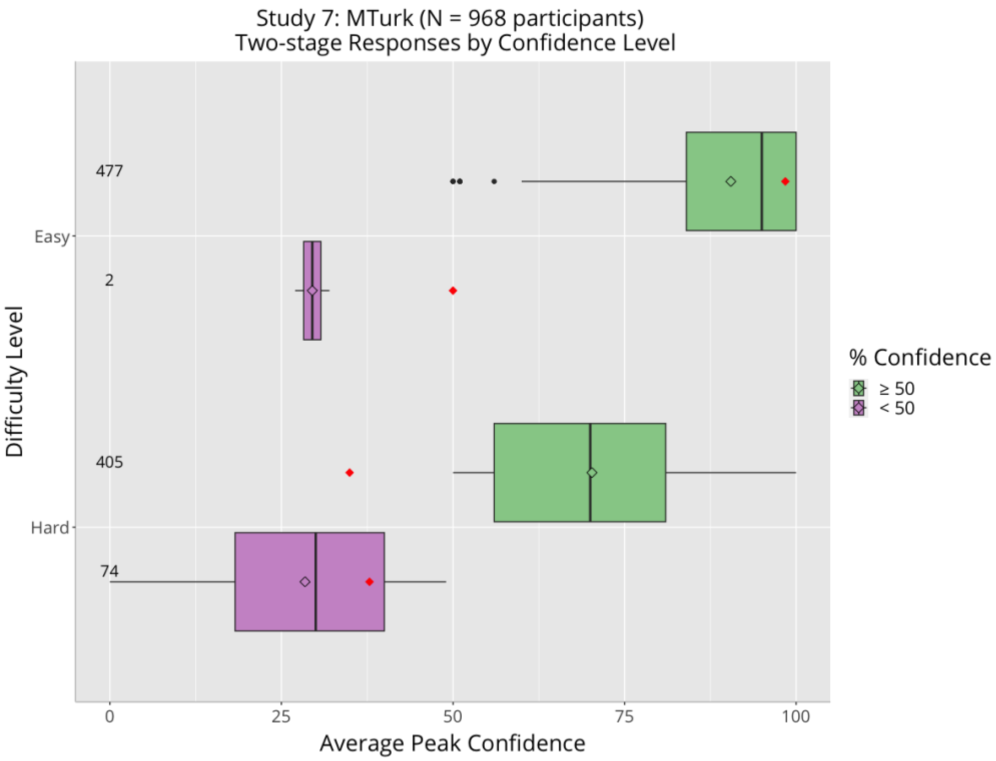


Figure 27. (Study 7) Average peak confidence by question difficulty and confidence level. Red diamonds indicate accuracy, transparent diamonds indicate mean. Number of responses per confidence level listed on the lefthand side.

The percentage of responses indicating less than 50% confidence differs by difficulty level, with 15.45% of responses indicating less than 50% confidence for hard questions in the two-stage condition and 0.42% for easy (see Table 4 below).

	Easy	Hard
< 50% confidence	2 responses	774responses
≥ 50% confidence	477 responses	405 responses

Table 4. (Study 7) Number of Responses by Task and Confidence Level

Exploring the data further, I filter the dataset and remove two-stage responses that indicate less than 50% confidence in the selected answer and compare peak confidence by difficulty level against the one-stage responses. I planned to run a 2 (condition) x 2 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty) on peak confidence with the restricted dataset. The analysis yields no significant main effect of condition, $F(1, 965) = 0.04$, $p = .835$), and no interaction between condition and difficulty, $F(1, 890) = 0.66$, $p = .415$. This suggests that the responses in the two-stage condition under fifty percent are driving the results seen in the main analysis depicted in Figure 25. Figure 28 depicts average peak confidence by question difficulty and condition, filtering the data for two-stage responses that more than or equal to 50.

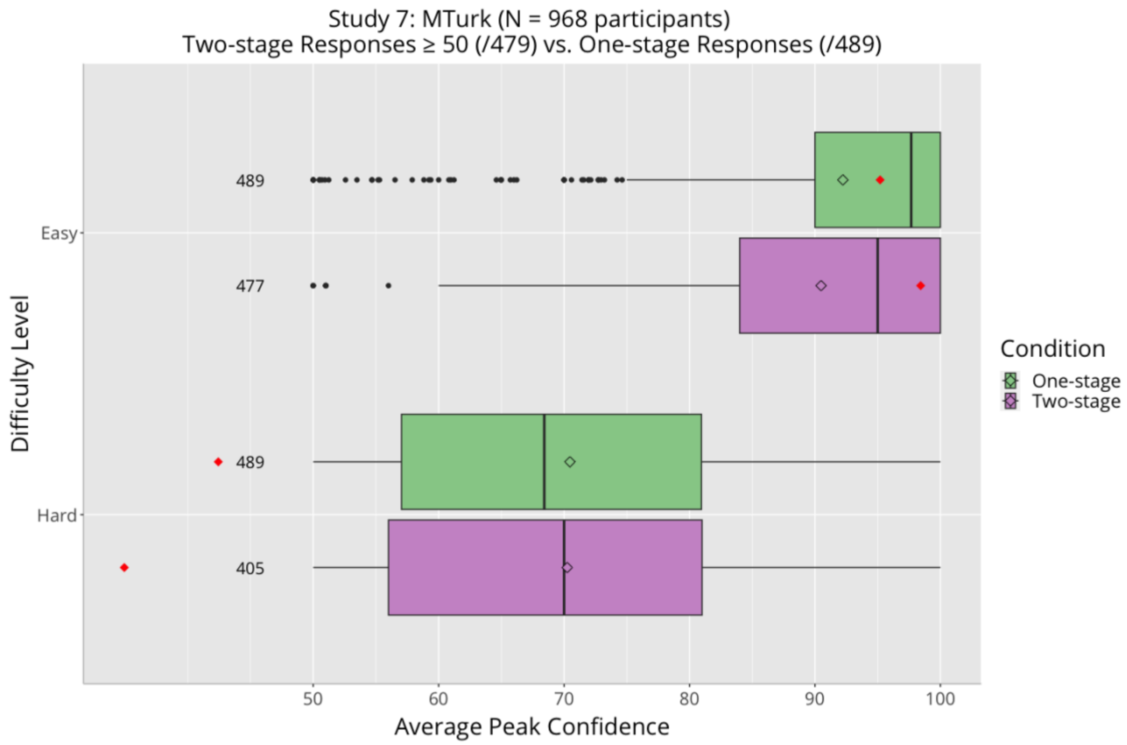


Figure 28. (Study 7) Boxplot showing average peak confidence by question difficulty level and condition, using filtered dataset. Red diamonds indicate accuracy, transparent diamonds indicate mean.

3) Other Notes

I asked all participants two open-ended questions at the end of the survey. The first was on what they thought the survey was asking, and the second was on whether they found anything in the survey weird or confusing.

In both conditions, all participants generally reported the survey was asking about age, confidence, people, perceptions, and so on. The vast majority of responses indicated finding nothing about the survey weird or confusing. A few participants did pick up on the fact that the faces were AI generated.

Discussion

I find that average peak confidence in the one-stage condition is higher than average peak confidence in the two-stage condition, even when using different stimuli, and even when participant choices are incentive compatible. The more salient differences in difficulty in this study made it easier to pick up on the interaction effect – average peak confidence is higher in the one-stage than in the two-stage condition, and this difference is exacerbated when questions are difficult. This implies that difficulty moderates the relationship between elicitation format and average peak confidence. Exploratory analyses suggest that the responses in the two-stage condition under fifty percent are driving the results.

CHAPTER 4 | General Discussion and Future Directions

Why studying confidence elicitations is important

Confidence elicitation is an important aspect of decision-making in various fields, including management and psychology. The assessment of an individual's level of confidence in their decisions or judgments is critical to understanding how decisions are made and can significantly impact the quality and outcomes of these decisions. This is particularly relevant in the fields of management and psychology, where decision-making plays a crucial role in organizational success and individual well-being.

Studying confidence elicitation is important for several reasons. First, it can help us gain insights into how people make decisions, particularly in complex and uncertain situations. By understanding how individuals form and communicate their confidence in their judgments, we can gain valuable insights into the decision-making process and identify potential biases or cognitive errors that may affect decision quality (Kahneman, 2011). Additionally, research on confidence elicitation can help us identify effective ways to communicate decision uncertainty, which is particularly important in situations where the consequences of decisions can have significant impact.

Second, studying confidence elicitation can help us improve decision-making outcomes. By using effective confidence elicitation techniques, we can obtain more accurate and reliable confidence estimates, which can lead to better decision-making outcomes (Lichtenstein et al., 1982). This is particularly relevant in high-stakes decision-making contexts, where the consequences of poor decision-making can be severe.

Third, understanding the methods used to elicit confidence is crucial for ensuring the accuracy and reliability of results. There are many different methods that can be used to elicit confidence, including rating scales, confidence intervals, and probability distributions. Each method has its strengths and weaknesses, and understanding what they are can guide selection of the most appropriate method for a given context (Kahneman et al., 1982). Moreover, understanding the impact of different methods on confidence estimates can help decision-makers choose the most effective technique for obtaining accurate and reliable confidence estimates.

Contributions to the literature

This dissertation focuses on the third reason, exploring how the way in which the question comes to us may affect how confident we feel and how we respond to our choices. It shows the power of question formats, and contributes more broadly to the literature looking at the psychology of survey responses (Tourangeau et al., 2000; Zaller & Feldman, 1992). Life forces choices between options every day. For instance, when people get offers of admission or employment, they must decide whether to take them. People who receive marriage proposals must choose whether to say “yes.” Medical patients must choose between treatments, such as medicines or surgeries. Voters must choose between candidates.

I studied two prominent methods of confidence elicitation, and explored whether they lead to different outcomes on the same questions, and examined why that might be the case. The extant literature suggests that the one-stage confidence elicitation method might lead to lower average peak confidence than the two-stage method for two primary reasons: (1) one-stage elicitations may force respondents to "consider the opposite" (Larrick, 2004; Lord et al., 1984), and (2) partition dependence suggests that unknown or unknowable possibilities lead people to assign equal probabilities to each of the given options (Fox & Clemen, 2005). Given the increasing popularity of the one-stage elicitation format in academic research (Hu & Simmons, 2023), it is important to systematically compare how these elicitation methods affect decision-making processes across different contexts. This comparative approach can elucidate the potential advantages and drawbacks of each method, providing valuable insights into their effectiveness and suitability for various research or practical applications.

In three initial exploratory studies, I examined whether there may be an interesting phenomenon worth studying when comparing the two confidence elicitations. I gathered enough evidence to motivate a series of pre-registered, well-powered studies that dug more deeply into whether elicitation formats can lead to a difference in average peak confidence where they should not, and why that might be. In Study 4, I found initial support for my hypothesis: there was a main effect of elicitation format on average peak confidence, and this effect was stronger for the hard questions than for the easy questions. Notably, Study 4 looked at a sample of undergraduate students.

Study 5 built upon the results of Study 4 by examining whether a one-stage elicitation leads to higher average peak confidence than a two-stage elicitation for hard versus easy questions, and whether the amount participants deliberate can make a difference. The exploratory studies hinted that undergraduates might simply be racing through the survey as compared to the best Amazon Mechanical Turk workers. Additionally, research suggests undergraduate data quality is of lesser quality than MTurk, and that setting Cloud Research and MTurk requirements also significantly improves data quality (Douglas et al., 2023; Zhang et al., 2024). Study 5 aimed to contribute to this literature and show response patterns under conditions that are more similar to day-to-day decision making. We often don't make judgements under conditions where we're incentivized to pay attention lest it affects our performance score and make many quick judgements in passing. Low deliberation might lead to a main effect of condition, whereas high deliberation might lead to an interaction effect, where the difference in peak confidence between conditions is significant only for difficult questions. Unfortunately, the low deliberation sample proved more difficult to attain than originally anticipated, and any results reported should be seen as exploratory. While Study 4 was pre-registered, it was underpowered. Study 5 provides the first fully powered test of the research question, and replicates the finding found in Study 4.

Study 6 has a practical bent; it examined which type of confidence elicitation – the one-stage or the two-stage – participants report being most familiar with, and whether familiarity with the elicitation used may lead to different results. To the best of my knowledge, there are no empirical studies investigating participant preferences in this area. Interestingly but perhaps unsurprisingly, participants report being most familiar with the two-stage elicitation format, encountering it more on MTurk and in daily life, and preferring to answer questions under that elicitation. Participants who select the two-stage format might be implicitly signaling that they understand it better – this might lead to less violations of rationality, which would bring average peak confidence up. My results show that this is not the case – even when participants are able to select which elicitation they prefer, I still see the main effect and the interaction. Notably, accuracy did not differ across any of the conditions – participants were more accurate when say, they were able to choose their elicitation format.

Study 7 nicely rounds out the package of studies, showing that elicitation format impacts peak confidence and that difficulty moderates this relationship, even when using different stimuli, and even when participant choices are incentive compatible.

In sum, I find across seven studies that one-stage elicitation formats lead to higher average peak confidence than two-stage elicitation formats and that difficulty moderates this relationship, such that the effect is stronger when the question posed is difficult. Familiarity with elicitation format does not attenuate the effect. The effect remains stable when using different stimuli, and when participant decisions are incentive compatible.

In examining why this effect occurs, I find that it is driven by two-stage responses that violate rationality. Across three of the four pre-registered studies, I find that the two-stage responses that indicate less than 50% confidence in the selected answer are driving the main result. When these responses are filtered out, the difference in peak confidence between elicitation formats disappears, as does the moderating effect of question difficulty.

Future Directions

The set of studies in this dissertation provide a great starting point for investigating how elicitation format might impact what decisions makers take away from surveys. It is interesting that there is not simply an interaction effect between elicitation format and question difficulty, but also a main effect, such that participants answering questions in the one-stage format report higher average peak confidence than those answering questions in the two-stage format. This extends prior work comparing different elicitation formats, such as item-confidence vs. SPIES (Moore, Carter, et al., 2015), and is critical to study because of the increasing use cases of the one-stage format in research (Hu & Simmons, 2023).

This work also contributes to extant research that looks at difficulty as a moderator. The hard-easy effect would suggest overconfidence on difficult tasks, and underconfidence on easy tasks (Erev et al., 1994). While I do observe this in a couple of the studies, one striking feature of the results is the prevalence of underconfidence, across all difficulty levels. This may just be due to the specific stimuli used across the studies – the hard easy-effect is clearly present when the face stimuli are used, and the difficulty levels are very well differentiated. A natural extension would be to see whether different stimuli do align with what the hard-easy effect would product, and to examine generally whether the findings hold when the option set is larger than two.

Future studies will focus on studying the effect in the context of consequential, meaningful stimuli. I intend to examine whether and how key organizational leaders might make very different decisions depending on the elicitation format used – when a one-stage elicitation format is used and average peak confidence is thereby higher, would decision makers be more likely to go with the favored option? Do survey designers take question difficulty into account when designing surveys that include difficult questions, and that use the two-stage elicitation format, which is reportedly more frequently used in daily life? The potential extensions are endless, and the area is ripe for exploration.

Furthermore, it is essential to investigate whether these findings hold consistent in real-world settings beyond controlled experimental environments, thereby enhancing the generalizability and applicability of the research outcomes. As such, future studies should consider incorporating a diverse range of scenarios and participant demographics to fully explore the impact of confidence elicitation methods on judgment and decision making.

Concluding Remarks

Our confidence in our ability to select the right option has enormous consequences. When we are sure, we can commit without hesitation. When we are unsure, then it might be wise to hedge our bets, to delay, or to gather more information.

By gaining insights into how people form and communicate confidence in their judgments, researchers can better understand the underlying mechanisms that influence decision making and identify effective confidence elicitation techniques. This can inform effective decision making, leading to better outcomes for organizations and individuals alike. Through this focus on confidence elicitation and methods, I hope to contribute in some small way to our understanding of judgement and decision making.

REFERENCES

Ariely, D., Loewenstein, G., & Prelec, D. (2005). Tom Sawyer and the myth of fundamental value. *Journal of Economic Behavior & Organization*, *60*(1), Article 1.

Bruine De Bruin, W., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty-fifty chance." *Organizational Behavior and Human Decision Processes*, *81*(1), 115–131.

Casey, E. (2021). Are professional forecasters overconfident? *International Journal of Forecasting*, *37*(2), 716–732.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*(2), Article 2.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge university press.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos One*, *18*(3), e0279720.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), Article 3.

Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty-fifty = 50%? *Journal of Behavioral Decision Making*, *12*(2), 149–163.

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, *51*(9), 1417.

Giordani, P., & Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, *47*(6), 1037–1059.

Giordani, P., & Söderlind, P. (2006). Is there evidence of pessimism and doubt in subjective distributions? Implications for the equity premium puzzle. *Journal of Economic Dynamics and Control*, *30*(6), 1027–1043.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1–14.

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*(7), Article 7.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), Article 4.

Hoffrage, U. (2004). Overconfidence. In R. F. Pohl (Ed.), *Cognitive illusions: Fallacies and biases in thinking, judgment, and memory* (pp. 235–254). Psychology Press.

Hofman, J. M., Goldstein, D. G., & Hullman, J. (2020). *How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results*. 1–12.

Hu, B., & Simmons, J. P. (2023). Does constructing a belief distribution truly reduce overconfidence? *Journal of Experimental Psychology: General*.

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), Article 3.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), Article 2.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*. Blackwell Publishers.

Liberman, V. (2004). Local and global judgments of confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 729–732.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art in 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–333). Cambridge University Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231.

Moore, D. A. (2020). *Perfectly confident*. Harper Collins.

Moore, D. A. (2023). Overprecision is a property of thinking systems. *Psychological Review*.

Moore, D. A., Carter, A. B., & Yang, H. H. (2015). Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organizational Behavior and Human Decision Processes*, *131*, 110–120.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making* (pp. 182–212). Wiley.

Morewedge, C. K., Tang, S., & Larrick, R. P. (2016). Betting Your Favorite to Win: Costly Reluctance to Hedge Desired Outcomes. *Management Science*, *64*(3), 997–1014. https://doi.org/10.1287/mnsc.2016.2656

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*(7), Article 7.

Reinholtz, N., Fernbach, P. M., & De Langhe, B. (2021). Do people understand the benefit of diversification? *Management Science*, *67*(12), 7322–7343.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299–314.

Soll, J. B., Palley, A., Klayman, J., & Moore, D. A. (2023). Overconfidence in probability distributions: People know they don’t know but they don’t know what to do about it. *Unpublished Manuscript*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3435793

Teigen, K. H., & JØrgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *19*(4), 455–475.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*.

Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2016). Known unknowns: A critical determinant of confidence and calibration. *Management Science*, *63*(12), Article 12.

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science*, 579–616.

Zhang, W., Wang, C., Mittal, A., Schroeder, J., & Moore, D. A. (2024). Comparison of Data Quality for Scientific Participant Pools. *Working Paper*.

SUPPLEMENTARY MATERIALS

Exploratory Studies

Study 1 tests the question of whether there is a difference in confidence under the two elicitation methods with two different stimuli and three difficulty levels. Study 2 focuses on the stimuli for which difficulty level is better differentiated and attempts to replicate the results with Cloud Research approved Amazon Mechanical Turk participants. Study 3 attempts to replicates results from Study 2 and examines the reasoning behind violations of rationality in the two-stage condition.

Study 1: Does a one-stage elicitation format lead to higher average peak confidence than a two-stage format in a sample of undergraduates?

Method

Participants

The final sample includes 191 UC Berkeley undergraduate participants (111 Women, 76 Men, 2 Other; 117 Asian, 30 White/Caucasian, 18 Hispanic, 18 Other, 6 African American, 1 Native American, 1 Pacific Islander). Participants were undergraduates who were taking the core business class, undergraduate 105, in Fall semester 2023. As a part of their course requirements, all students in the course were required to complete a pre-survey, which included questions from researchers. This study was exploratory, was not pre-registered, and made no exclusions. 93 participants were in the one-stage condition, and 98 participants were in the two-stage condition.

Design

The experiment had a two-cell between-subjects design that manipulated how confidence was elicited. In the one-stage confidence elicitation condition, participants distributed their confidence over two answer choices, with one slider scale per answer choice. The survey forced confidence to sum to 100.

In the two-stage confidence elicitation condition, participants first chose which option they believed was the correct answer, and then indicated how confident (0-100% slider scale) they were that their selected answer was indeed correct. The primary dependent variable was the confidence participants placed on their favored answer choice.

Procedure and Materials

Participants in both conditions completed two tasks (a dot task and a face task) with three questions each (easy, medium, and hard), such that each participant completed six questions in total. The dot task showed two images of white squares with black dots inside, and asked participants to indicate which image had more dots. I generated dots uniformly on a 200 by 200 grid. I generated images with some number of dots between 20 and 99, inclusive. For the **hard**

task, the images differed by 2 dots (80 vs. 82), for the **medium task**, the images differed by 8 dots (50 vs. 58), and for the **easy task**, the images differed by 24 dots (67 vs. 43).

The face task showed them two images of AI generated faces, and asked them to indicate which face they thought was older. I generated faces from the FakeFace API. For the hard task, ages differed by 2 years (20 vs. 22), for the medium task, ages differed by 5 years (25 vs. 30), and for the easy task, ages differed by 15 years (25 vs. 40).

The survey randomly assigned participants to the two between-subjects conditions, presented tasks in randomized order within each condition, and randomized the order of questions for each task.

Results

1) ANOVA model with the task factor

I first ran an ANOVA model on the full data, then broke the results down further, looking at main effects and interactions. All analyses are exploratory and were not pre-registered. I conducted a 2 (condition) x 2 (task) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty, task) on peak confidence. I report the results with the task factor, but note that I am not interested in differences between stimuli and have no predictions for how task might differentially impact results. The results of an ANOVA without the task factor are consistent.

There is a significant main effect of condition, $F(1, 189) = 78.03$, $p < .001$, and of difficulty, $F(2, 378) = 305.72$, $p < .001$, on peak confidence. There is also a significant interaction between condition and difficulty, $F(2, 378) = 3.28$, $p = .039$, suggesting that the effect of condition on peak confidence varies depending on the difficulty level.

The main effect of task is significant, $F(1, 189) = 13.89$, $p < .001$, indicating that participants' peak confidence is higher for one of the two tasks. There is no significant interaction between condition and task, $F(1, 189) = 0.37$, $p = .543$, indicating that the effect of condition does not vary depending on the task.

There is a significant interaction between difficulty and task, $F(2, 378) = 60.62$, $p < .001$, indicating that the effect of difficulty on peak confidence varies depending on the task. Finally, there is a significant three-way interaction between condition, difficulty, and task, $F(2, 378) = 3.49$, $p = .032$, suggesting that the effects of condition and difficulty on peak confidence vary depending on the task. However, this effect is relatively small.

Overall, these results suggest that both condition and difficulty level have significant effects on participants' peak confidence, while the effect of task is relatively smaller. Figure S 1 below depicts peak confidence as a function of condition, difficulty, and task. The following sections break the results of the ANOVA down.

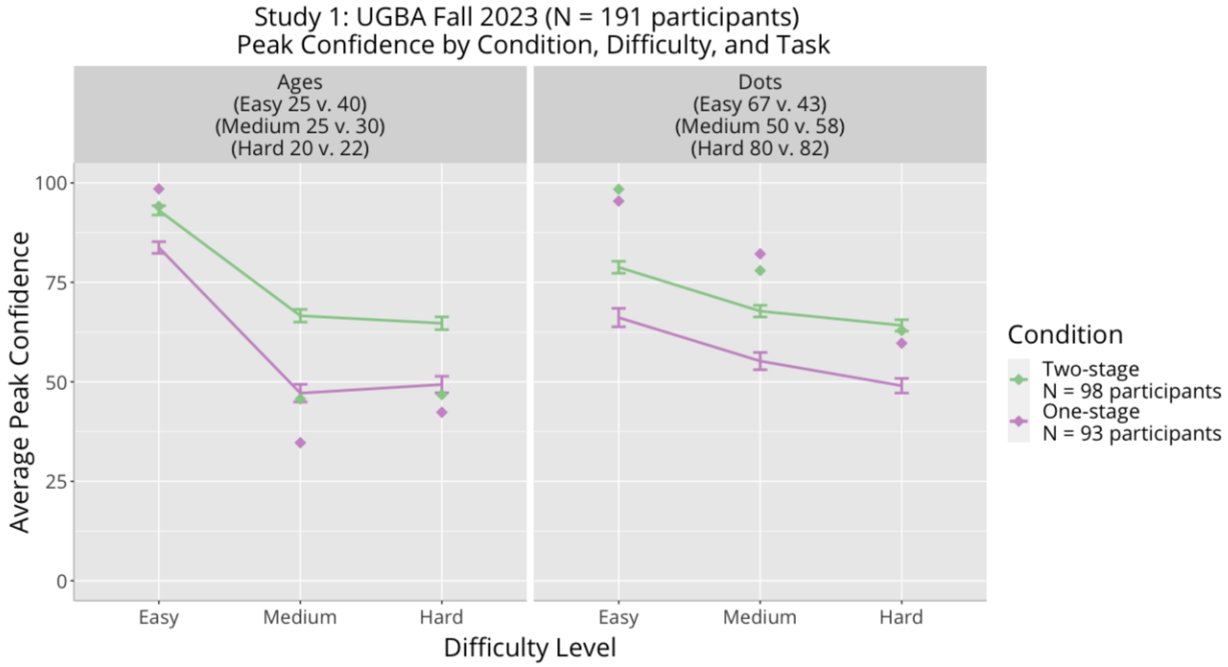


Figure S 1. (Study 1) Average peak confidence by condition, question difficulty level, and task. Diamonds indicate average accuracy.

2) Main effect of question difficulty

I examine whether there are any main effects of question difficulty. This is simply a check of whether questions were appropriately difficult based on level. I would expect to find that peak confidence is highest for the easy question, lower for the medium question, and lowest for the hard question. Collapsing across tasks and questions, I find that average peak confidence is 80.30% (*SD* = 19.09) for the easy question, 58.97% (*SD* = 20.44) for the medium question, and 56.61% (*SD* = 18.81) for the hard question. Figure S 2 depicts average peak confidence for the three difficulty levels, with average accuracy shown in red.

Independent samples *t*-tests reveal two of the comparisons to be significant (easy vs. medium, *t*(758.49) = 14.90, *p* < .001; easy vs. hard, *t*(761.83) = 17.27, *p* < .001) and one not (medium and hard, *t*(756.82) = 1.66, *p* = .097).

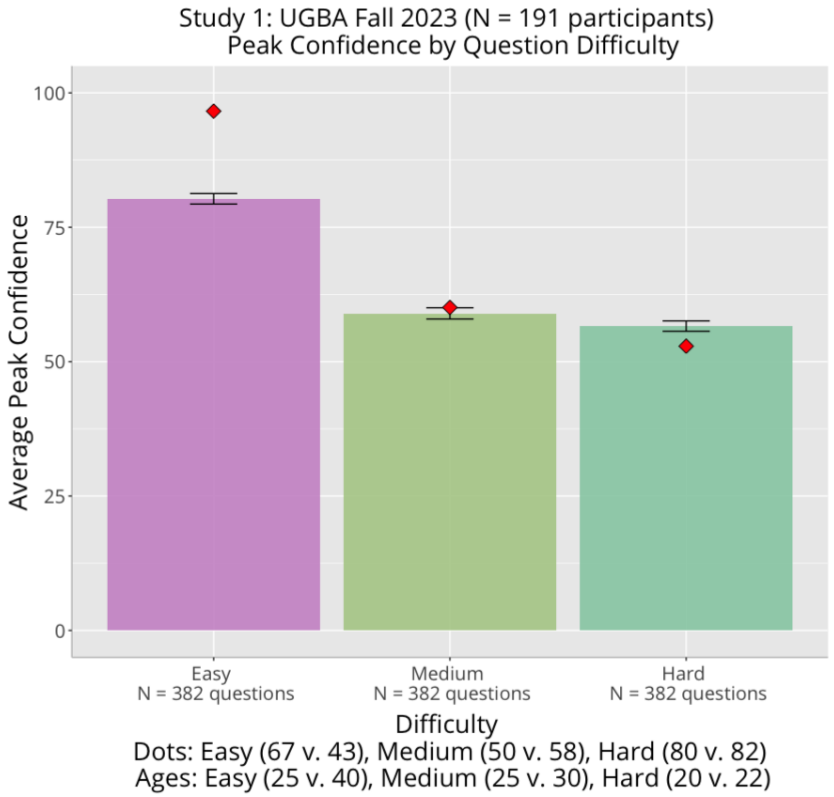


Figure S 2. (Study 1) Average peak confidence by question difficulty level. Red diamonds indicate average accuracy.

I break this down further by task below, and find that peak confidence for the medium and hard question for the ages task does not differ, *t*(377.72) = 2.59, *p* = .010, suggesting that perhaps the questions were too similar in difficulty level.

For the dots task, all three pairwise comparisons are significant, suggesting the three questions were well differentiated in terms of difficulty.

- Peak confidence differs significantly for the easy questions and the medium questions, *t*(379.02) = 5.42, *p* < .001, with higher peak confidence for the easy questions (*M* = 72.30, *SD* = 20.28) than for the medium questions (*M* = 61.32, *SD* = 19.27).
- Peak confidence differs significantly for the easy questions and the hard questions, *t*(373.88) = 8.14, *p* < .001, with higher peak confidence for the easy questions (*M* = 72.30, *SD* = 20.28) than for the hard questions (*M* = 56.40, *SD* = 17.83).
- Peak confidence differs significantly for the medium questions and the hard questions, *t*(377.72) = 2.59, *p* = .010, with higher peak confidence for the medium questions (*M* = 61.32, *SD* = 19.27) than for the hard questions (*M* = 56.40, *SD* = 17.83).

For the ages task, only two of the three pairwise comparisons are significant. Peak confidence for the medium and hard question does not differ, suggesting that perhaps the questions were too similar in difficulty level.

- Peak confidence differs significantly for the easy questions and the medium questions, *t*(325.63) = 17.22, *p* < .001, with higher peak confidence for the easy questions (*M* = 88.30, *SD* = 13.82) than for the medium questions (*M* = 56.62, *SD* = 21.34).

- Peak confidence differs significantly for the easy questions and the hard questions, $t(339.74) = 18.02$, $p < .001$, with higher peak confidence for the easy questions ($M = 88.30$, $SD = 13.82$) than for the hard questions ($M = 56.82$, $SD = 19.79$).
- Peak confidence does not differ significantly for the medium questions and the hard questions, $t(377.88) = -0.09$, $p = .927$; average peak confidence for the medium questions ($M = 56.62$, $SD = 21.34$) is similar to that of the hard questions ($M = 56.82$, $SD = 19.79$).

Again, I do not predict differences by task and am not interested in these differences, but report these analyses to see whether difficulty level differed by task and to guide stimuli selection for future studies. Figure S 3 depicts average peak confidence for the three difficulty levels, broken down by task.

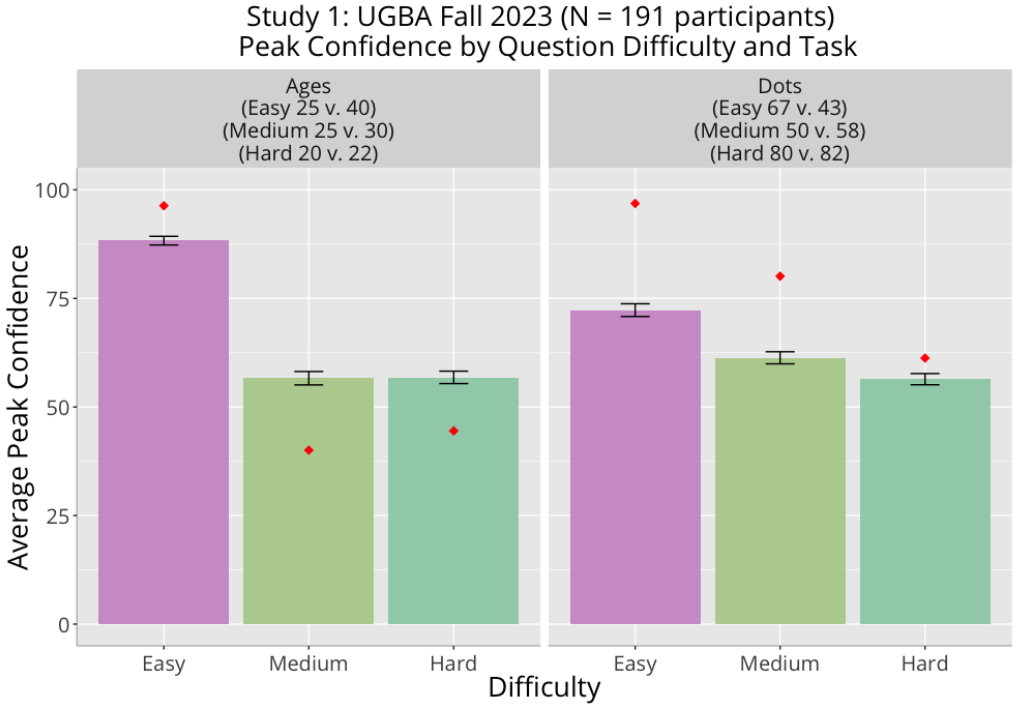


Figure S 3. (Study 1) Average peak confidence by question difficulty level and task. Red diamonds indicate average accuracy.

As an additional check on question difficulty, I note the average time taken to click the submit button for a given question. For the one-stage questions, the timer measured the time the participants spent on the page allocating confidence across the two answer options. For the two-stage questions, the timer measured the time participants spent on the page where they chose between the two answer options. Table S 1 below provides the summary statistics. Note that data was windsorized at the 5th and 95th percentiles.

Condition	Task	Difficulty	N	Mean	SD	Min	Max
One-stage	Dots	Easy	93	17.87	11.72	5.92	48.20
One-stage	Dots	Medium	93	18.21	11.56	6.36	48.10
One-stage	Dots	Hard	93	21.18	12.96	7.58	52.83

Two-stage	Dots	Easy	98	8.09	3.56	5.64	19.76
Two-stage	Dots	Medium	98	9.68	4.97	5.76	23.97
Two-stage	Dots	Hard	98	11.58	6.74	5.70	30.36
One-stage	Ages	Easy	93	11.91	6.24	5.83	28.39
One-stage	Ages	Medium	93	15.03	7.97	6.15	36.10
One-stage	Ages	Hard	93	19.61	15.82	6.87	71.85
Two-stage	Ages	Easy	98	7.62	2.94	5.67	16.62
Two-stage	Ages	Medium	98	12.22	7.65	5.78	32.77
Two-stage	Ages	Hard	98	11.95	6.34	5.99	30.22

Table S 1. (Study 1) Summary Statistics on Time Per Question in seconds.

3) Main effect of condition

Collapsing across tasks and questions, I find that average peak confidence in the one-stage condition (*M* = 72.52%, *SD* = 17.53) is significantly different from average peak confidence in the two-stage condition (*M* = 58.43%, *SD* = 23.89), *t*(1076.9) = 11.42, *p* < .001. Average accuracy does not differ between conditions (one-stage *M* = 70.97%, two stage *M* = 68.79%), *t*(1143.4) = 453.39, *p* = .388. Figure S 4 depicts average peak confidence in each condition, with average accuracy represented by the red diamonds.

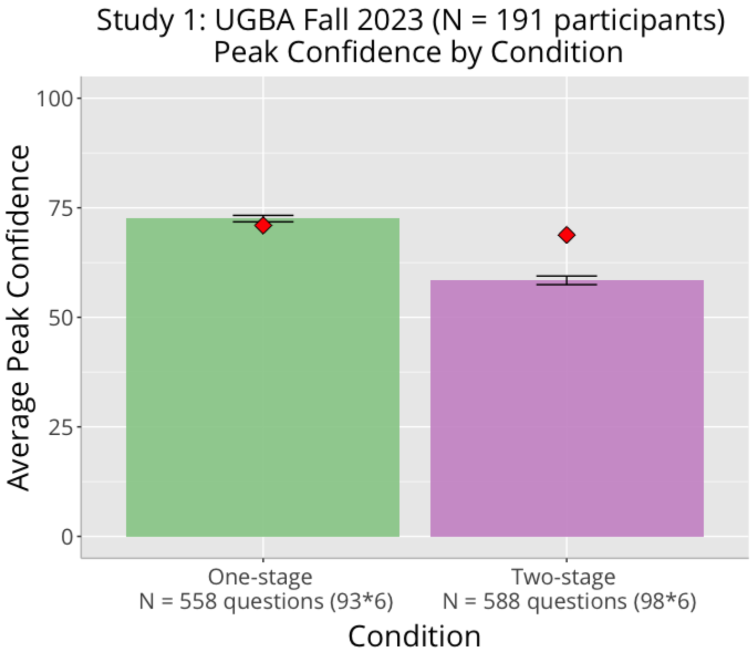


Figure S 4. (Study 1) Average peak confidence by condition. Red diamonds indicate average accuracy.

The finding that average peak confidence differs significantly by condition holds true when the result is broken down by task. For both the dots task, *t*(524.77) = 8.50, *p* < .001, and the ages task, *t*(543.28) = 7.83, *p* < .001, average peak confidence is higher in the one-stage condition than the in two-stage condition.

4) Interaction between condition and difficulty

I examine whether peak confidence differs by condition for each of the difficulty levels, with data collapsed across tasks. I find that for all three difficulty levels, there is a significant difference between average peak confidence in the one-stage condition and the two-stage condition: Easy one-stage (*M* = 85.94, *SD* = 14.81) vs. two-stage (*M* = 74.94, *SD* = 21.09), *t*(350.57) = 5.92, *p* < .001; Medium one-stage (*M* = 67.18, *SD* = 14.75) vs. two-stage (*M* = 51.18, *SD* = 22.01), *t*(342.44) = 8.38, *p* < .001; Hard one-stage (*M* = 64.45, *SD* = 14.60) vs. two-stage (*M* = 49.17, *SD* = 19.37), *t*(361.65) = 8.73, *p* < .001. Figure S 5 depicts these results, with red diamonds indicating average accuracy.

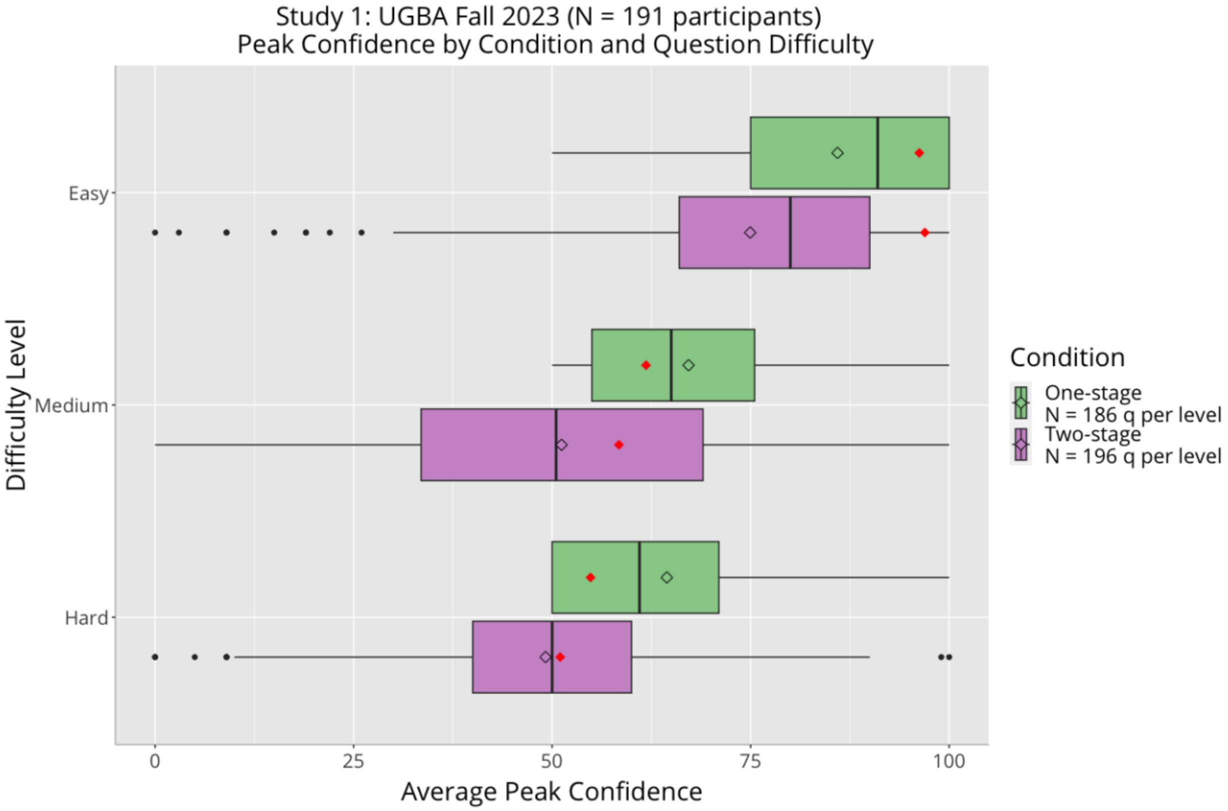


Figure S 5. (Study 1) Average peak confidence by condition and by question difficulty Level. Red diamonds indicate accuracy, transparent diamonds indicate mean.

Given that the difficulty levels were more successfully differentiated for the dots task than the ages task, I break down the effect condition has on peak confidence by difficulty and task, and find that the results hold.

The results for the ages task are as follows:

- For the easy questions in the ages task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 93.10, *SD* = 11.33) and the two-stage condition (*M* = 83.74, *SD* = 14.48), *t*(182.48) = 4.99, *p* < .001.
- For the medium questions in the ages task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 66.60, *SD* = 15.41) and the two-stage condition (*M* = 47.15, *SD* = 21.91), *t*(174.50) = 7.12, *p* < .001.
- For the hard questions in the ages task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 64.71, *SD* = 15.47) and the two-stage condition (*M* = 49.33, *SD* = 20.59), *t*(179.67) = 5.86, *p* < .001.

The results for the dots task are as follows:

- For the easy questions in the dots task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 78.78, *SD* = 14.47) and the two-stage condition (*M* = 66.14, *SD* = 22.98), *t*(164.67) = 4.57, *p* < .001.
- For the medium questions in the dots task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 67.76, *SD* = 14.12) and the two-stage condition (*M* = 55.21, *SD* = 21.47), *t*(168.65) = 4.80, *p* < .001.
- For the hard questions in the dots task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 64.18, *SD* = 13.75) and the two-stage condition (*M* = 49.02, *SD* = 18.18), *t*(180.16) = 6.52, *p* < .001.

5) Deviations from rationality in the two-stage condition

I examine whether participants violate rationality in the two-stage condition. Their confidence in the favored answer should, in theory, be at least 50 percent. I expected the hard questions might have the most people violating rationality. Indeed, I find that in the two-stage condition, across all tasks and difficulty levels, 161 out of 588 responses (27.34%) had less than 50 percent confidence in the answer selected in the first stage. Figure S 6 depicts average peak confidence for each difficulty level, broken down by those who answered less than 50% confidence and those who answered more.

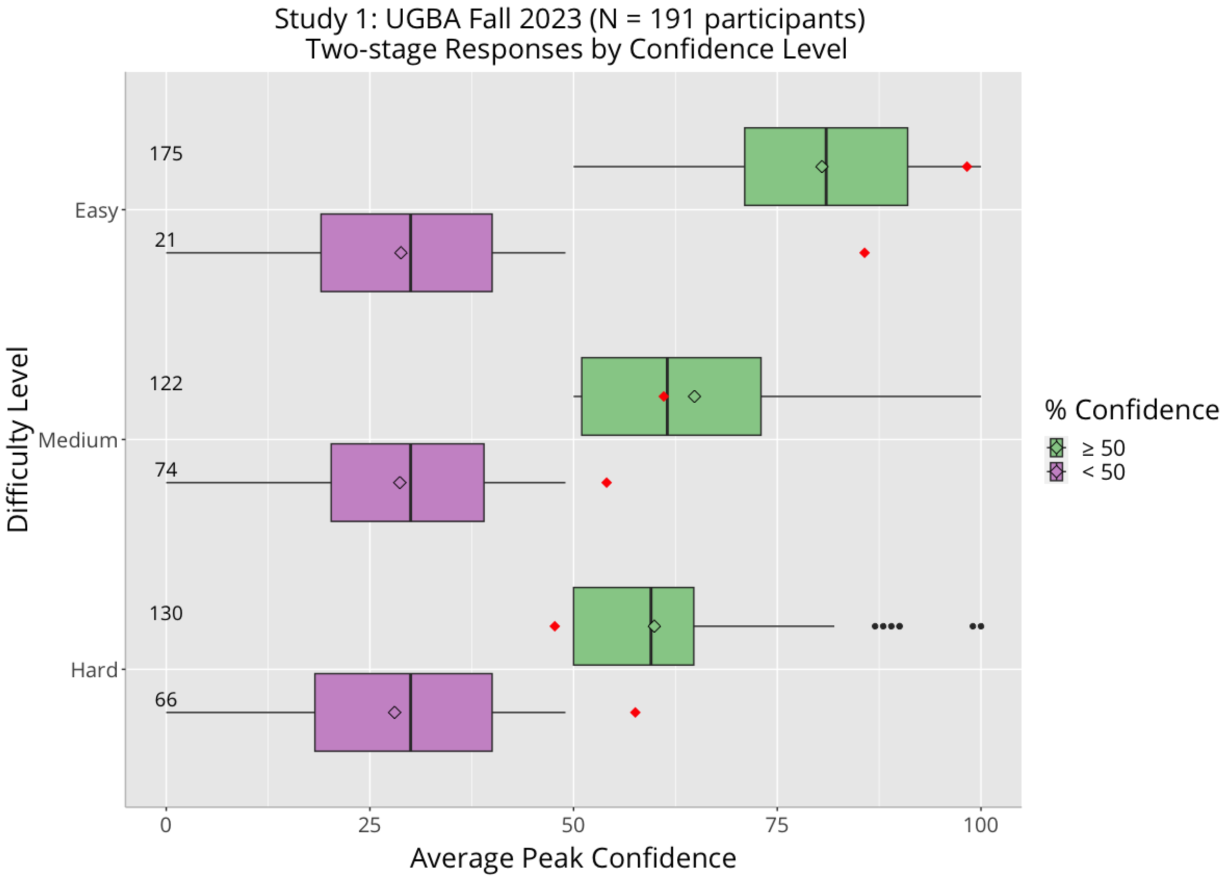


Figure S 6. (Study 1) Average peak confidence by question difficulty and confidence level. Red diamonds indicate average accuracy. Number of responses per confidence level listed on the left-hand side.

I break this result down further by task (see Table S 2 below) and by difficulty (see Table S 3 below). The percentage of responses indicating less than 50% confidence is similar between tasks (27.89% for ages task, and 26.87% for dots task).

	Task	
	Ages	Dots
< 50% confidence	82 responses	79 responses
≥ 50% confidence	212 responses	215 responses

Table S 2. (Study 1) Number of Responses by Task and Confidence Level

The percentage of responses indicating less than 50% confidence differs by difficulty level, with 33.67% of responses indicating less than 50% confidence for hard questions in the two-stage condition, 37.76% for medium, and 10.71% for easy. Given the medium and hard questions were less differentiated in terms of difficulty than desired, it is not surprising that the proportion of responses under 50% does not increase monotonically.

The comparison of interest is between the easy and medium/hard questions, as these two groups were well differentiated in terms of average peak confidence. The results are as expected – fewer responses in the two-stage condition for the easy questions indicate a less than 50% confidence than in the medium/hard questions. In other words, when the answer to a question is fairly easy in the two-stage condition, few participants are uncertain, resulting in fewer violations of rationality, or fewer responses indicating less than 50% confidence. When the answer to a question is fairly difficult (e.g., a coin flip), more participants are uncertain, resulting in more violations of rationality.

	Question Difficulty Level		
	Easy	Medium	Hard
< 50% confidence	21 responses	74 responses	66 responses
≥ 50% confidence	175 responses	122 responses	130 responses

Table S 3. (Study 1) Number of Responses by Question Difficulty Level and Confidence Level

Exploring the data further, I filter the dataset and remove two-stage responses that indicate less than 50% confidence in the selected answer and compare peak confidence by difficulty level against the one-stage responses. With this restricted data set, the observed pattern is mitigated. Average peak confidence is significantly higher in the one-stage than in the two-stage condition for only the easy and hard questions. Figure S 7 depicts these results, collapsed across tasks.

1. For the easy questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 85.94, *SD* = 14.81) and the two-stage condition (*M* = 80.48, *SD* = 13.55), *t*(358.71) = 3.66, *p* < .001.
2. For the medium questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 67.18, *SD* = 14.75) and the two-stage condition (*M* = 64.84, *SD* = 13.77), *t*(271.21) = 1.42, *p* = .156.
3. For the hard questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 64.45, *SD* = 14.60) and the two-stage condition (*M* = 59.91, *SD* = 11.36), *t*(310.36) = 3.10, *p* = .002.

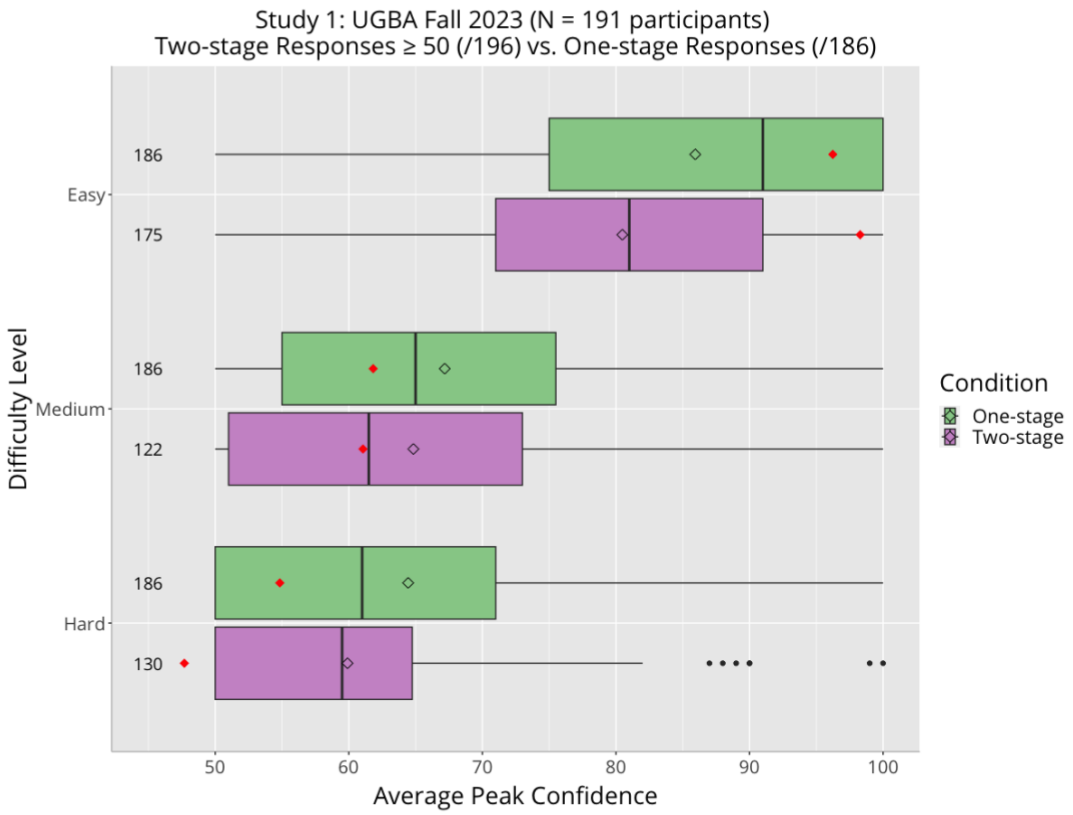


Figure S 7. (Study 1) Boxplot showing average peak confidence by question difficulty level and condition, using filtered dataset. Red diamonds indicate accuracy, transparent diamonds indicate mean. Number of responses per confidence level listed on the lefthand side.

Given that the difficulty levels were more successfully differentiated for the dots task than the ages task, I break down the effect condition has on peak confidence by difficulty and task. I find that the results hold.

The results for the ages task are as follows:

- For the easy questions in the ages task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 93.10, *SD* = 11.33) and the two-stage condition (*M* = 83.74, *SD* = 14.48), *t*(182.48) = 4.99, *p* < .001.
- For the medium questions in the ages task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 66.60, *SD* = 15.41) and the two-stage condition (*M* = 47.15, *SD* = 21.91), *t*(174.50) = 7.12, *p* < .001.
- For the hard questions in the ages task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 64.71, *SD* = 15.47) and the two-stage condition (*M* = 49.33, *SD* = 20.59), *t*(179.67) = 5.86, *p* < .001.

The results for the dots task are as follows:

- For the easy questions in the dots task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 78.78, *SD* = 14.47) and the two-stage condition (*M* = 66.14, *SD* = 22.98), *t*(164.67) = 4.57, *p* < .001.

- For the medium questions in the dots task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 67.76, *SD* = 14.12) and the two-stage condition (*M* = 55.21, *SD* = 21.47), *t*(168.65) = 4.80, *p* < .001.
- For the hard questions in the dots task, there is a significant difference between the mean confidence ratings for the one-stage condition (*M* = 64.18, *SD* = 13.75) and the two-stage condition (*M* = 49.02, *SD* = 18.18), *t*(180.16) = 6.52, *p* < .001.

6) Other Notes

I asked participants "Were you able to pay attention during this section?" at the end of my sections in the survey. No participant reported paying no attention, 7 participants reported they "paid a bit of attention", 42 participants reported paying moderate attention, 55 paying a lot, and 87 paying full attention.

I asked participants in both conditions two open-ended questions at the end of the survey. The first was on what they thought the survey was asking, and served as another check that participants paid attention, and the second was on whether they found anything in the survey weird or confusing.

In both conditions, participants generally reported the survey was asking about confidence, judgement, decision making, etc. Several participants also wrote about "perception and bias," "pattern recognition," and the like – it's possible they read into the ages task too much. This guess is supported by their responses to the question on whether they found anything in the survey weird – several participants commented on the ages task, saying they were concerned it was about racism, that guessing ages was odd, etc. A few responses in the one-stage condition also indicated confusion on the confidence measure – participants reported being unclear on why the percents/totals had to sum to 100.

Discussion

In summary, the exploratory analyses show a main effect of condition and an interaction between condition and difficulty. These patterns hold when two-stage responses that indicate less than 50% confidence in the selected answer are removed from the dataset.

Notably, the undergraduate 105 pre-survey is a somewhat long survey that participants can complete from anywhere. Researchers send the administrator of the survey their materials, and the administrator then randomly orders the survey. Length of the survey varies depending on how many researchers take part in a given semester. Participants also fill out PANAS, OCEAN and various demographic questions at the end of the survey.

Study 2: Does the main effect of condition replicate in a sample of Cloud Research approved Amazon Mechanical Turk participants?

Given the length of the survey and concerns about attention, as well as concerns about data quality with undergraduate participants, I ran exploratory Study 2 to see whether the main effect of condition found in Study 1 would replicate with a sample of top Amazon Mechanical Turk participants.

Method

Participants

The final sample includes 152 Amazon Mechanical Turk participants (57 Female, 88 Male, 3 Gender neutral; mean age = 39; 113 White, 14 Black or African American, 13 Asian, 5 Hispanic or Latino, 2 Other, 2 Prefer not to answer, 1 Native Hawaiian or Other Pacific Islander). These MTurk participants were approved by Cloud Research, had an MTurk approval rating of 95%+, and had to have completed 100 HITs. This study was exploratory and was not pre-registered. 76 participants were in the one-stage condition, and 76 participants were in the two-stage condition.

Design

The design followed that of Study 1, with one change: I did not force confidence to sum to 100 in the one-stage condition, and instead told participants I would normalize to 100. Again, the primary dependent variable was peak confidence.

I note one additional choice made in cleaning the data: If a participant left both sliders at 0 for a given question, normalizing the data would yield NaN for confidence. I take a conservative approach and assign 50% probability to the two answer choices.

With respect to exclusion, responses where a participant did not answer a given question (e.g., skipped the question) were filtered out in the data. Only one response fits this criterion – one respondent answered the easy and medium questions, but not the hard question.

Procedure and Materials

The procedure and materials followed that of Study 1, with one change: Given question difficulty was less differentiable, I dropped the ages task. Therefore, participants in both conditions completed one task (the dot task) with three questions (easy, medium, and hard), such that each participant completed three questions in total.

The survey randomly assigned participants to the two between-subjects conditions and randomized the order of questions.

Results

1) ANOVA model

I conducted a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty). There is a significant main effect of condition, *F*(1, 149) = 4.01, *p* = .047, and of difficulty on peak confidence, *F*(2, 299) = 23.93, *p* < .001. There is no significant interaction between condition and difficulty, *F*(2, 299) = 0.27, *p* = .762. Figure S 8 below depicts peak confidence as a function of condition and difficulty. The following sections break the results of the ANOVA down.

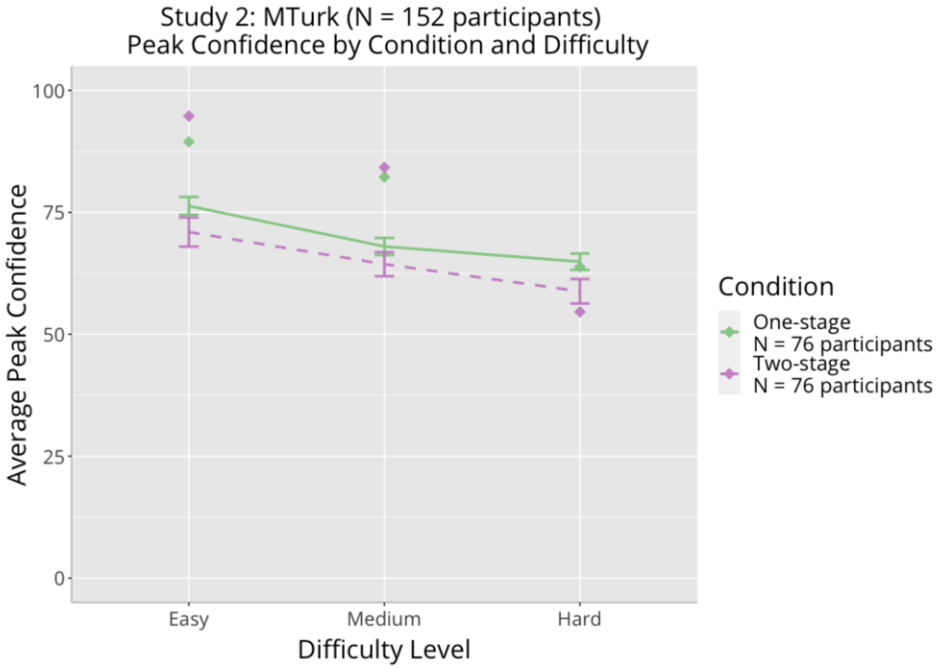


Figure S 8. (Study 2) Average peak confidence by condition and question difficulty level. Diamonds indicate average accuracy.

2) Main effect of question difficulty

I examine whether there are any main effects of question difficulty. Collapsing across questions, I find that average peak confidence is 73.64% (*SD* = 21.71) for the easy question, 66.19% (*SD* = 18.65) for the medium question, and 61.87% (*SD* = 18.69) for the hard question.

I find that there is a significant difference in peak confidence between the easy and medium question, *t*(295.3) = 3.21, *p* = .001, the easy and hard question, *t*(295.05) = 5.06, *p* < .001, and the medium and hard question, *t*(300.98) = 2.01, *p* = .045. Figure S 9 depicts average peak confidence for the three difficulty levels.

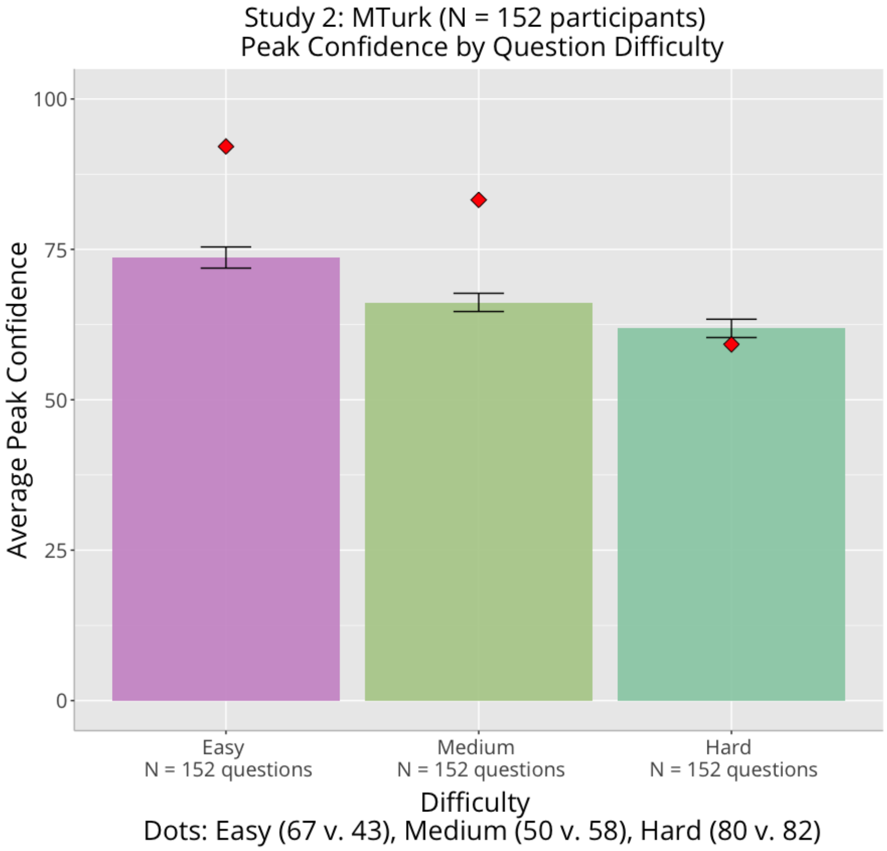


Figure S 9. (Study 2) Average peak confidence by question difficulty level. Red diamonds indicate average accuracy.

As an additional check on question difficulty, I note the average time taken to click the submit button for a given question. For the one-stage questions, the timer measured the time the participants spent on the page allocating confidence across the two answer options. For the two-stage questions, the timer measured the time participants spent on the page where they chose between the two answer options. Table S 4 below provides the summary statistics. Note that data was windsorized at the 5th and 95th percentiles.

Condition	Difficulty	N	Mean	SD	Min	Max
One-stage	Easy	76	17.93	14.32	6.11	61.30
One-stage	Medium	76	17.99	14.50	6.29	55.93
One-stage	Hard	76	19.48	14.25	6.45	54.48
Two-stage	Easy	76	7.99	2.90	5.82	15.54
Two-stage	Medium	76	9.87	5.03	5.81	22.43
Two-stage	Hard	75	9.66	5.98	5.61	27.29

Table S 4. (Study 2) Summary Statistics on Time Per Question in seconds.

3) Main effect of condition

Collapsing across difficulty levels, I find that average peak confidence is 69.74% (*SD* = 16.01) in the one-stage condition, and 64.74% (*SD* = 23.60) in the two-stage condition. This difference is significant with an independent samples t-test, *t*(397.5) = 2.64, *p* = .009. Average accuracy does not differ between conditions (one-stage *M* = 78.51, two stage *M* = 77.8), *t*(453.39) = 0.178, *p* = 0.859. Figure S 10 depicts average peak confidence in each condition, with average accuracy represented by the red diamonds.

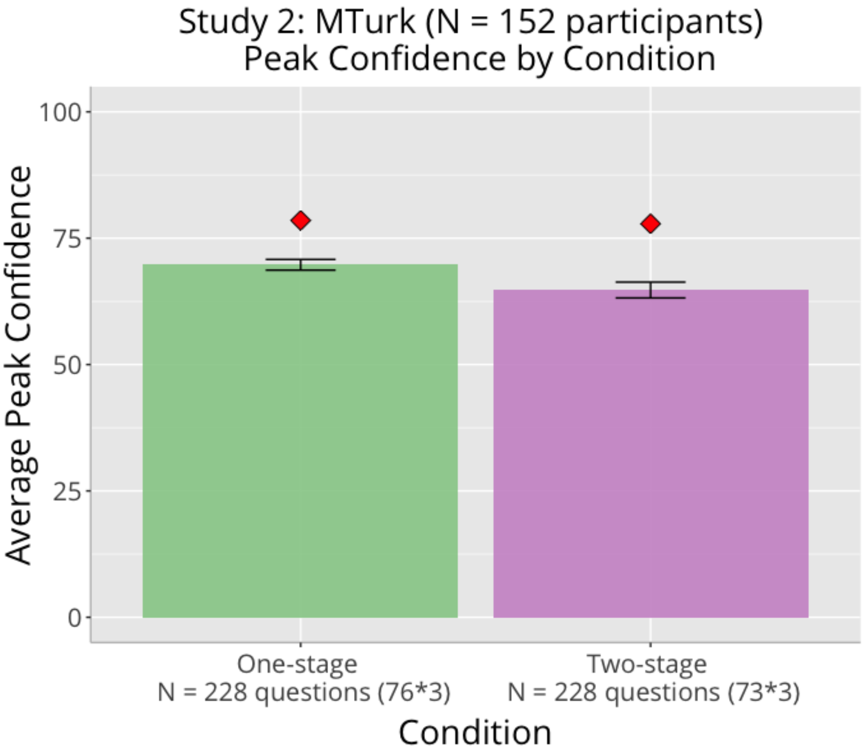


Figure S 10. (Study 2) Average peak confidence by condition. Red diamonds indicate average accuracy.

4) Interaction between condition and difficulty

I examine whether peak confidence differs by condition for each of the difficulty levels, and find a difference only for the hard questions. Figure S 11 depicts these results, with red diamonds indicating average accuracy.

1. For the easy questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 76.33, *SD* = 16.03) and the two-stage condition (*M* = 70.96, *SD* = 26.03), *t*(124.71) = 1.53, *p* = .129)
2. For the medium questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 68.01, *SD* = 15.26) and the two-stage condition (*M* = 64.37, *SD* = 21.47), *t*(135.34) = 1.20, *p* = .231.
3. For the hard questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 64.88, *SD* = 14.68) and the two-stage condition (*M* = 58.81, *SD* = 21.71), *t*(129.77) = 2.01, *p* = .047.

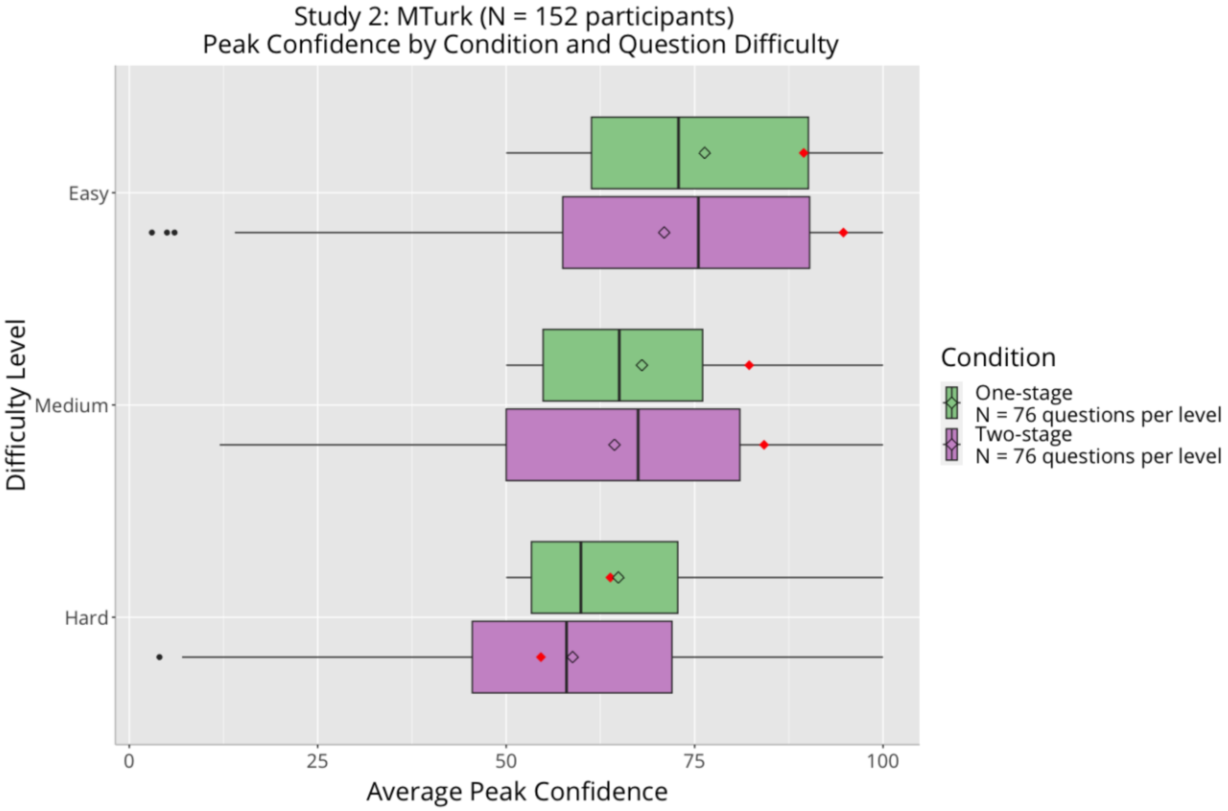


Figure S 11. (Study 2) Average peak confidence by condition and by question difficulty level. Red diamonds indicate accuracy, transparent diamonds indicate mean.

5) Deviations from rationality in the two-stage condition

As in Study 1, I examine whether participants violate rationality in the two-stage condition. I find that across all tasks and difficulty levels, 53 out of 227 responses (23.35%) had less than 50 percent confidence in the answer selected in the first stage. Figure S 12 depicts average peak confidence for each difficulty level, broken down by those who answered less than 50% confidence and those who answered more.

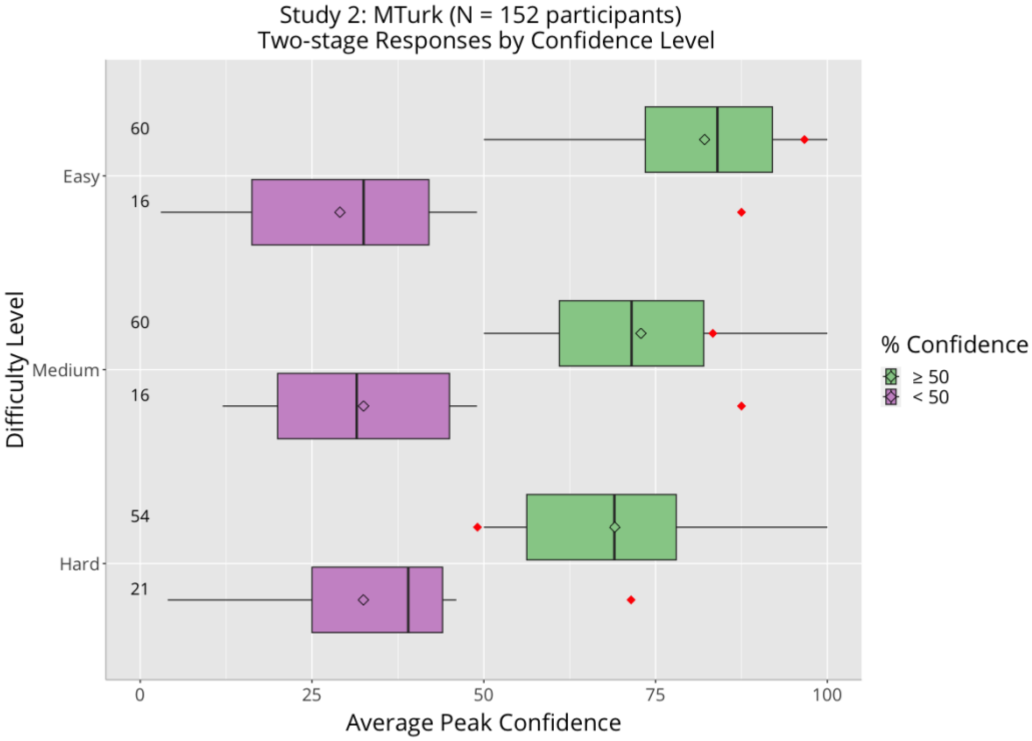


Figure S 12. (Study 2) Average peak confidence by question difficulty and confidence Level. Red diamonds indicate accuracy, transparent diamonds indicate mean. Number of responses per confidence level listed on the lefthand side.

The percentage of responses indicating less than 50% confidence differs by difficulty level, with 28.00% of responses indicating less than 50% confidence for hard questions in the two-stage condition, 21.05% for medium, and 21.05% for easy (see Table S 5 below). As noted above, average peak confidence was significantly different between the medium and hard question, but the effect was very small. Given the medium and hard questions were less differentiated in terms of difficulty than desired, it is not surprising that the proportion of responses under 50% does not increase monotonically.

The comparison of interest is between the easy and medium/hard questions, as these two groups were well differentiated in terms of average peak confidence. The results flip as compared to the first study – more responses in the two-stage condition for the easy questions indicate a less than 50% confidence than for the medium/hard questions. This is difficult to interpret.

	Question Difficulty Level		
	Easy	Medium	Hard
< 50% confidence	21 responses	16 responses	16 responses
≥ 50% confidence	54 responses	60 responses	60 responses

Table S 5. (Study 2) Number of Responses by Task and Confidence Level

Exploring the data further, I filter the dataset and remove two-stage responses that indicate less than 50% confidence in the selected answer, and compare peak confidence by difficulty level against the one-stage responses. With this restricted data set, the observed pattern reverses. Average peak confidence is significantly higher in the two-stage than in the one-stage condition for the easy and medium questions. Figure S 13 depicts these results.

1. For the easy questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 76.33, *SD* = 16.03) and the two-stage condition (*M* = 82.13, *SD* = 13.91), *t*(132.75) = -2.26, *p* = .025.
2. For the medium questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 68.01, *SD* = 15.26) and the two-stage condition (*M* = 72.87, *SD* = 13.95), *t*(131.09) = -1.94, *p* = .055.
3. For the hard questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 64.88, *SD* = 14.68) and the two-stage condition (*M* = 69.06, *SD* = 14.29), *t*(116.14) = -1.62, *p* = .107.

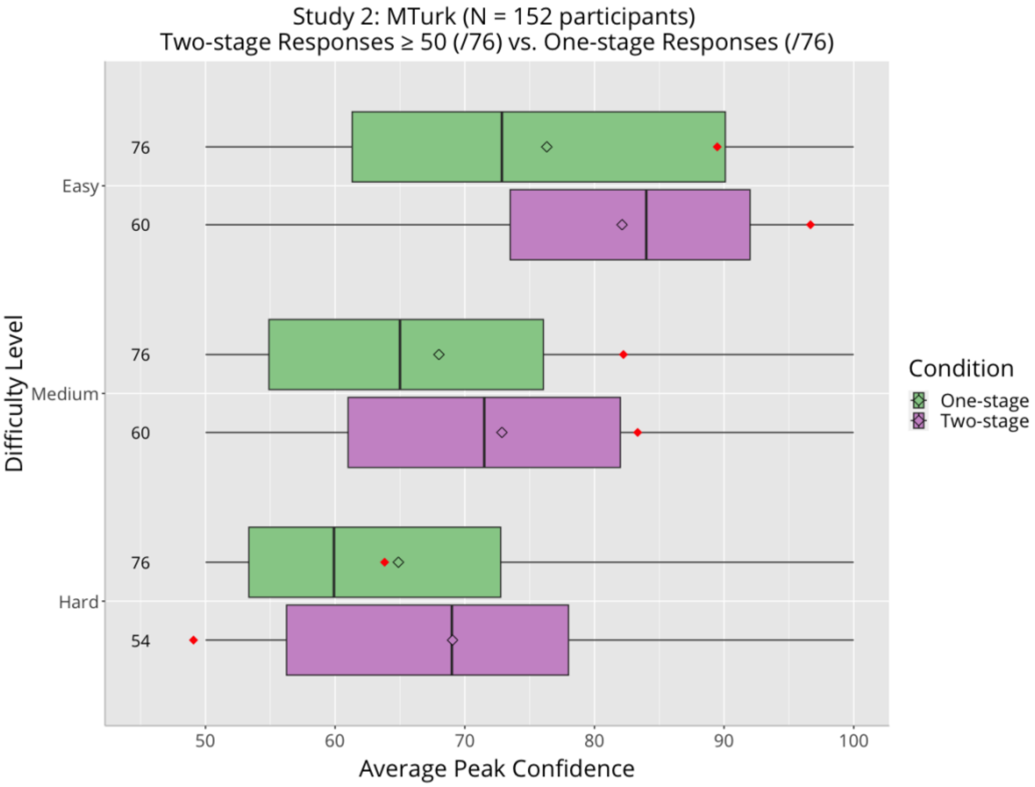


Figure S 13. (Study 2) Boxplot showing average peak confidence by question difficulty level and condition, using filtered dataset. Red diamonds indicate accuracy, transparent diamonds indicate mean.

6) Other Notes

I asked participants to honestly report how much attention they paid during the study, and ask two open-ended questions on what the survey was about and whether they found anything weird or confusing.

For the attention question on MTurk, participants are asked to answer the question, “Were you able to pay attention during this section?” with the note that, “This question will not be used to reject your HIT, or prevent you from receiving future surveys. It is only for our own benefit – if we know how attentive you are it helps us better analyze the data.” No participant reported paying no attention or only a bit of attention, 1 participant reported paying moderate attention, 5 paying a lot, and 144 paying full attention.

In both conditions, participants generally reported the survey was asking about dots, confidence, judgement, decision making, etc. Several participants also wrote about visual judgement, perceptions, visual reasoning, cues, and the like. The vast majority of responses to the question asking whether anything was strange or unclear about the survey was no, with two notable comments:

- Strange is the norm here. If strange isn't your thing, then mturk isn't really for you.
- If I didn't like strange, then I'd be over on Prolific, doing studies about how to suss out turning on a lightbulb.

Discussion

In summary, the exploratory analyses show a small effect of condition, but the interaction between condition and difficulty Is not significant. This does not quite replicate the findings from Study 1, and the difference in average peak confidence between conditions seems to be much smaller.

The patterns in the data do not hold when two-stage responses that indicate less than 50% confidence in the selected answer are removed from the dataset – the effect flips and is significant in the other direction such that average peak confidence is higher in the two-stage condition, but only for the easy questions. This does not replicate my findings from Study 1.

Study 3: (Why) do participants violate rationality in the two-stage condition?

The purpose of Study 3 was to explore the portion of participants who indicated they were less than 50% confidence in the two-stage condition, as this proportion seems to stay fairly consistent across both studies. I wanted to examine the reasoning behind these participants' choices.

Method

Participants

The final sample includes 193 Amazon Mechanical Turk participants (82 Female, 108 Male, 1 Gender neutral; mean age = 41; 148 White, 12 Black or African American, 14 Asian, 12 Hispanic or Latino, 5 Other, 2 Prefer not to answer). With respect to exclusions, I filtered out anyone whose recorded Qualtrics progress was less than 100%. 93 participants were in the one-stage condition, and 100 participants were in the two-stage condition.

Design

The design followed that of Study 2, with the following additions: Two-stage participants who put less than 50% confidence in their selected answer choice in the second stage read on the following page:

> "You indicated you were X% confident in your selected answer 'Image [A/B] has more dots.' This implies you're 1-X% confident in the answer you did not select (e.g., that you're 1-X% confident in 'Image [B/A] has more dots')."

Participants then answered four questions:

1. Is "Image [A/B] has more dots" the answer option you meant to select?
2. Is X% the percent confidence you meant to indicate in your selected answer "Image [A/B] has more dots"?
3. Would you like to change your answer to "Image [B/A] has more dots"?
4. Please feel free to explain your thinking in the text box below.

I note that due to the use of embedded data in showing participants these additional questions, question order could not be randomized. This is a constraint of the Qualtrics survey software. The medium question was always shown first, followed by the easy and then the hard question.

I normalized answers to 100; if participants put 0 on both options in the one-stage condition, the normalized data reflected 50% confidence in each of the two choices. Again, the primary dependent variable was peak confidence.

Procedure and Materials

As in Study 2, participants in both conditions completed one task (the dot task) with three questions (fixed order: medium, easy, and hard), such that each participant completed three questions in total.

The survey randomly assigned participants to the two between-subjects conditions, but did not randomize question order.

Results

1) ANOVA model

I conducted a 2 (condition) x 3 (difficulty) mixed ANOVA (between subjects: condition, within subjects: difficulty). There is no significant main effect of condition, *F*(1, 191) = 2.18, *p* = .142. There is a significant main effect of difficulty on peak confidence, *F*(2, 382) = 64.05, *p* < .001. There is a significant interaction between condition and difficulty, *F*(2, 382) = 3.19, *p* = .042. Figure S 14 below depicts peak confidence as a function of condition and difficulty. The following sections break the results of the ANOVA down.

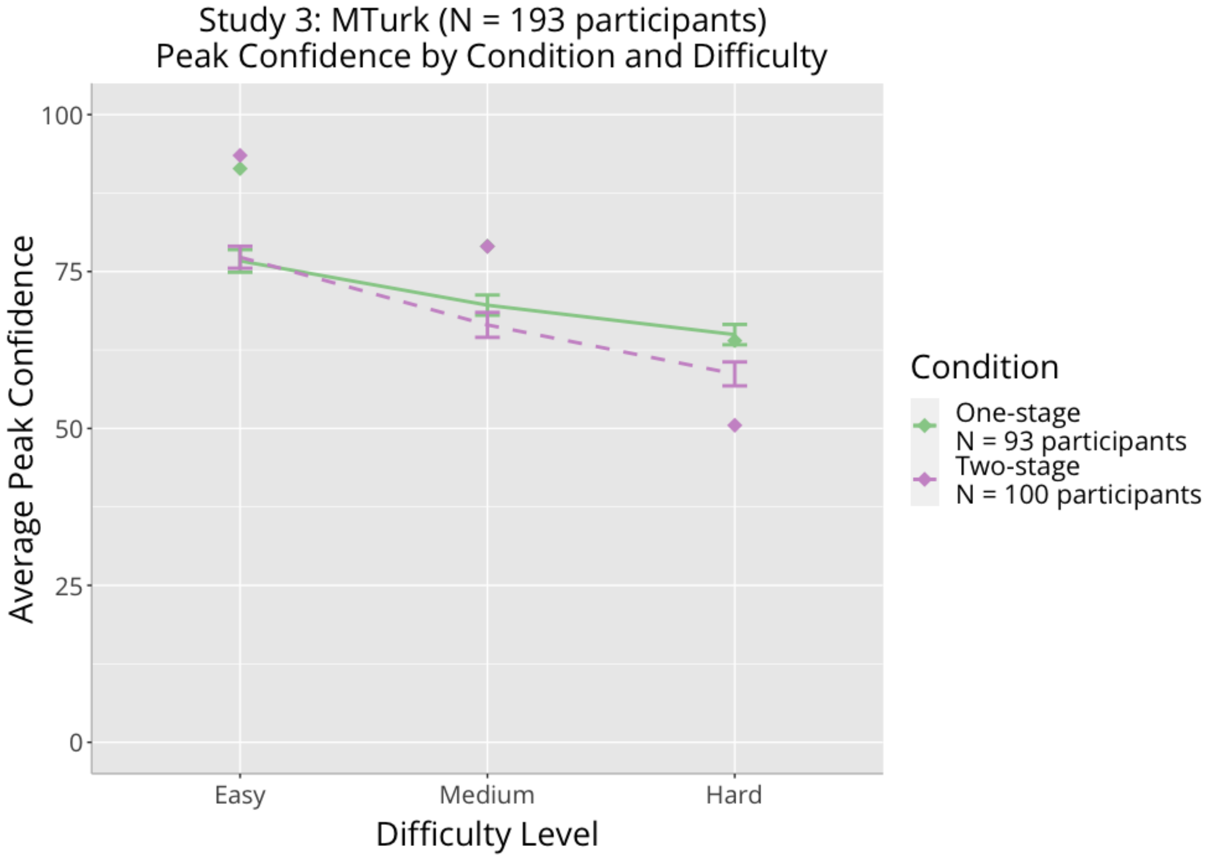


Figure S 14. (Study 3) Average peak confidence by condition and question difficulty level. Diamonds indicate average accuracy. Note that average accuracy for the medium difficulty question in the one-stage condition is 79.03, and in the two-stage condition is 79, so the diamonds overlap almost perfectly in the graph.

2) Main effect of question difficulty

I examine whether there are any main effects of question difficulty to check whether questions were appropriately difficult based on level. Collapsing across questions, I find that average peak confidence is 77.00% (*SD* = 17.40) for the easy question, 68.02% (*SD* = 17.92) for the medium question, and 61.71% (*SD* = 17.71) for the hard question. These numbers are notably stable across both MTurk studies.

I find that there is a significant difference in peak confidence between the easy and medium question, *t*(383.67) = 4.99, *p* < .001, the easy and hard question, *t*(383.88) = 8.55, *p* < .001, and the medium and hard question, *t*(383.95) = 3.48, *p* < .001. Figure S 15 depicts average peak confidence for the three difficulty levels.

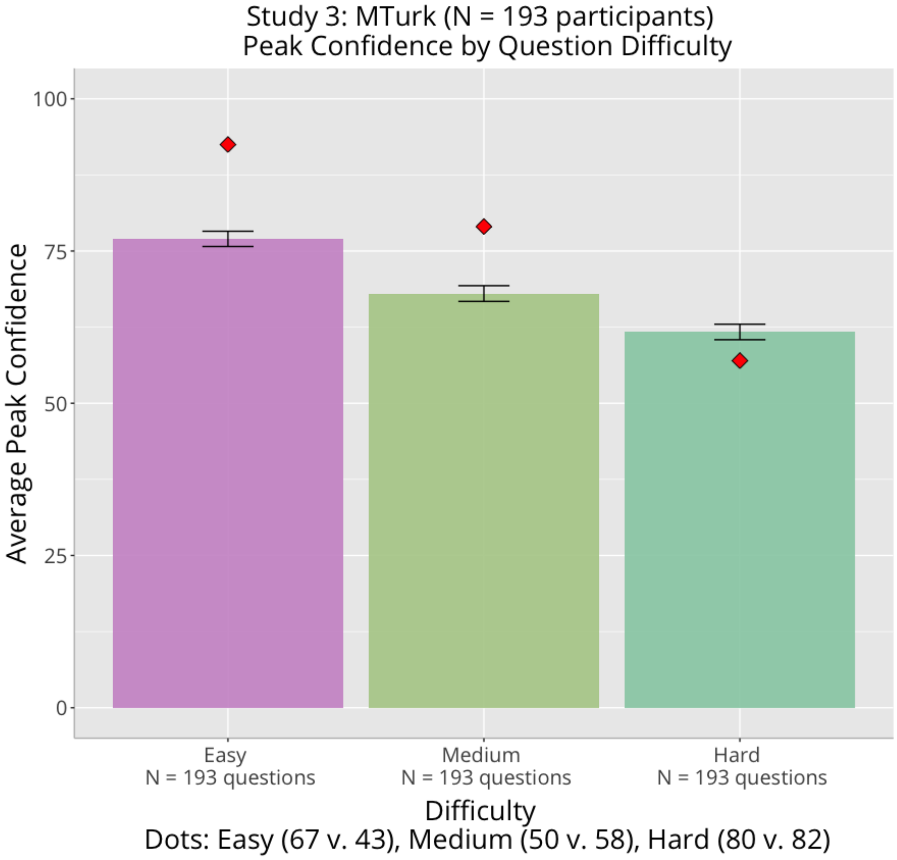


Figure S 15. (Study 3) Average peak confidence by question difficulty level. Red diamonds indicate average accuracy.

As an additional check on question difficulty, I note the average time taken to click the submit button for a given question. For the one-stage questions, the timer measured the time the participants spent on the page allocating confidence across the two answer options. For the two-stage questions, the timer measured the time participants spent on the page where they chose between the two answer options. Table S 6 below provides the summary statistics. Note that data was windsorized at the 5th and 95th percentiles. In this study, there is a bigger difference in time spent on question between the medium and hard questions across conditions, in the opposite direction than one would expect – participants spent on average more time on the medium questions than on the hard questions.

Condition	Difficulty	N	Mean	SD	Min	Max
One-stage	Easy	101	15.92	9.82	5.79	37.04
One-stage	Medium	100	24.40	19.36	6.88	78.93
One-stage	Hard	100	18.86	11.91	6.22	49.24
Two-stage	Easy	104	9.02	6.32	5.63	30.23
Two-stage	Medium	105	10.75	5.90	5.82	27.32
Two-stage	Hard	103	8.66	4.38	5.61	22.61

Table S 6. (Study 3) Summary Statistics on Time Per Question in seconds.

3) Main effect of condition

Collapsing across difficulty levels, I find that average peak confidence is 70.43% (*SD* = 16.86) in the one-stage condition, and 67.49% (*SD* = 20.25) in the two-stage condition; this difference is marginally significant, *t(*570.14) = 1.90, *p* = .057. Average accuracy does not differ between conditions (one-stage *M* = 78.14, two stage *M* = 74.33), *t*(576.33) = 1.132, *p* = .258. Figure S 16 depicts average peak confidence in each condition, with average accuracy represented by the red diamonds.

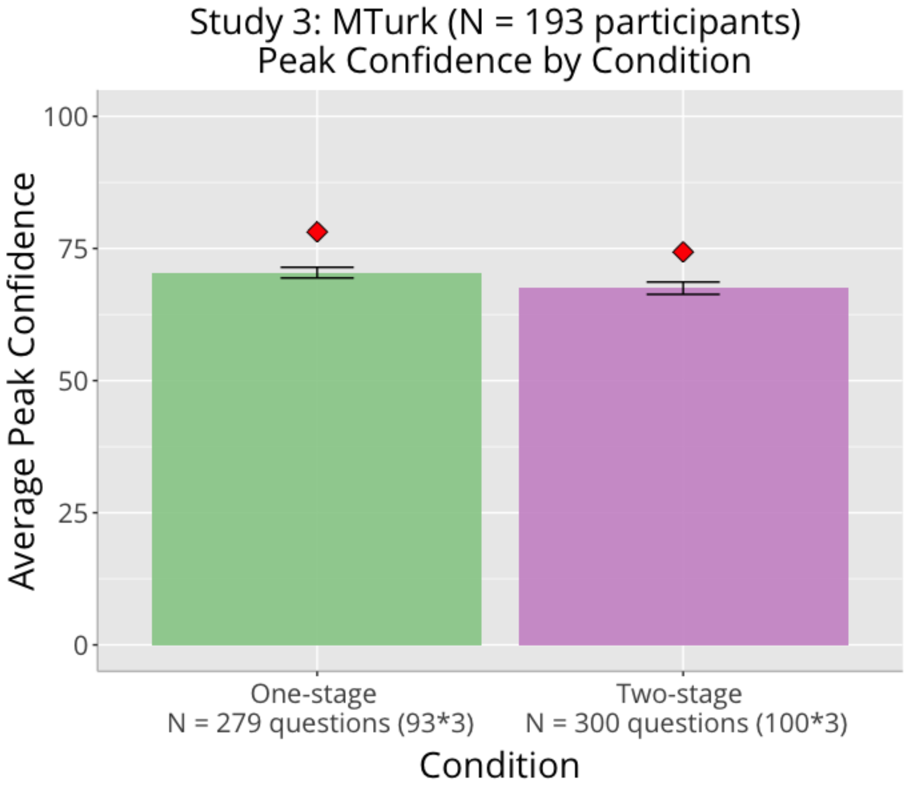


Figure S 16. (Study 3) Average peak confidence by condition. Red diamonds indicate average accuracy.

4) Interaction between condition and difficulty

I examine whether peak confidence differs by condition for each of the difficulty levels, and find a difference only for the hard questions. Figure S 17 depicts these results.

1. For the easy questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 76.69, *SD* = 17.38) and the two-stage condition (*M* = 77.28, *SD* = 17.51), *t*(190.18) = -0.23, *p* = .815.
2. For the medium questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 69.65, *SD* = 15.54) and the two-stage condition (*M* = 66.50, *SD* = 19.84), *t*(185.75) = 1.23, *p* = .220.
3. For the hard questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 64.95, *SD* = 15.63) and the two-stage condition (*M* = 58.69, *SD* = 19.03), *t*(188.16) = 2.51, *p* = .013.

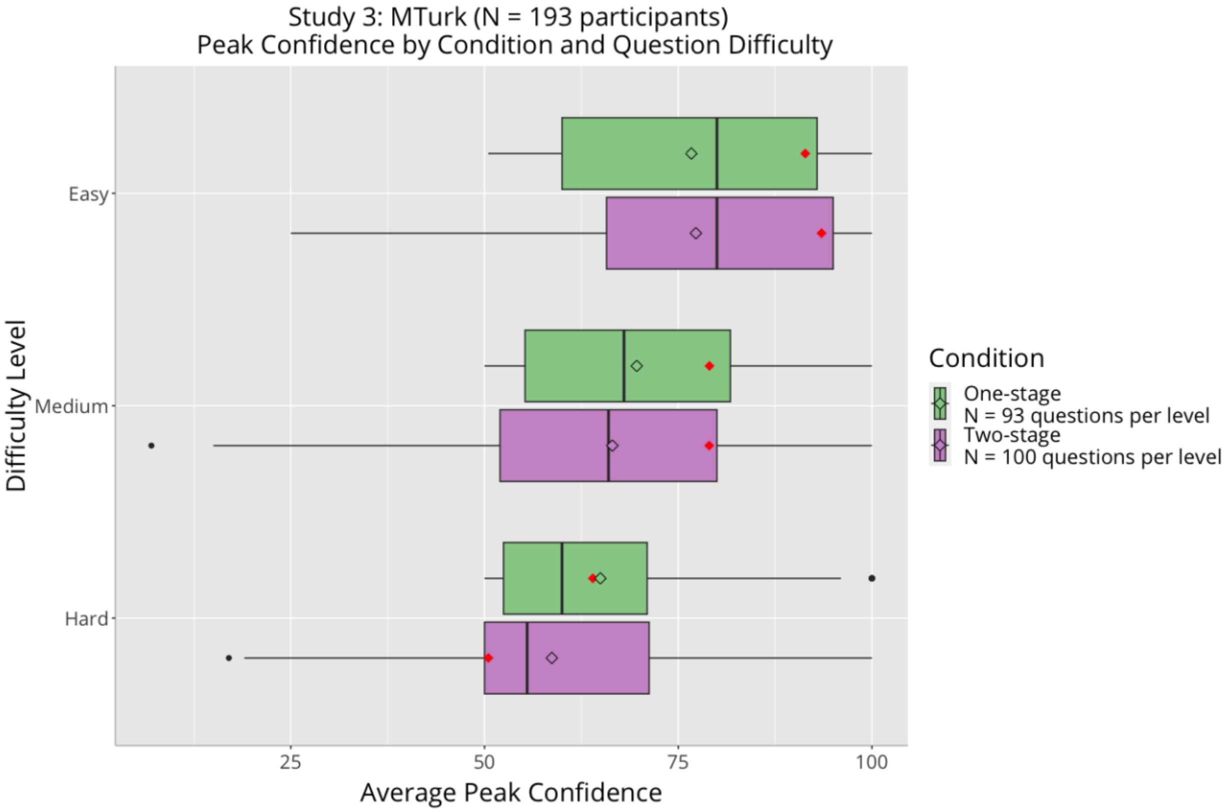


Figure S 17. (Study 3) Average peak confidence by condition and by question difficulty level. Red diamonds indicate accuracy, transparent diamonds indicate mean.

5) Deviations from rationality in the two-stage condition

I examine whether participants violate rationality in the two-stage condition. Figure S 18 depicts average peak confidence for each difficulty level, broken down by those who answered less than 50% confidence and those who answered more. I find that in the two-stage condition, across all tasks and difficulty levels, 30 out of 300 responses (10.00%) had less than 50 percent confidence in the answer selected in the first stage. Notably, this is less than half the proportion of responses with less than 50 percent confidence in prior studies. I attempt to address this discrepancy by examining whether the fixed order may have affected subsequent responses.

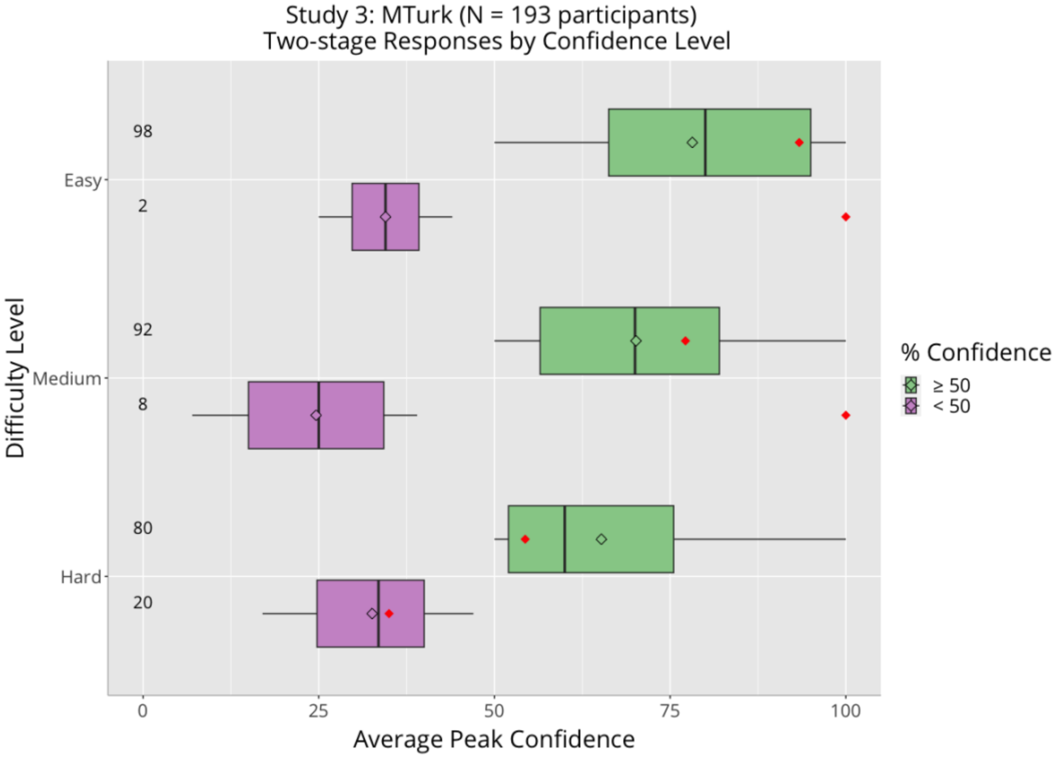


Figure S 18. (Study 3) Average peak confidence by question difficulty and confidence level. Red diamonds indicate accuracy, transparent diamonds indicate mean. Number of responses per confidence level listed on the lefthand side.

The percentage of responses indicating less than 50% confidence differs by difficulty level, with 20.00% of responses indicating less than 50% confidence for hard questions in the two-stage condition, 8.00% for medium, and 2.00% for easy (see Table S 7 below).

	Question Difficulty Level		
	Easy	Medium	Hard
< 50% confidence	2 responses	8 responses	20 responses
≥ 50% confidence	98 responses	92 responses	80 responses

Table S 7. (Study 3) Number of Responses by Task and Confidence Level

Given question order was fixed, I compare these percentages to Study 2, factoring in order. In Study 2, 27 participants in the two-stage condition saw the medium question first, 26 saw the easy question first, and 21 saw the hard question first. Of the participants who saw the medium question first, 22% of responses (6/27) indicate a less than 50% confidence in the two-stage condition. This is still much higher than what I find in Study 3, but of course given the small sample, I hesitate to make any strong conclusions.

13 participants in Study 2 saw the questions in the same order as in Study 3 (i.e., Medium, Easy, Hard). Table S 8 shows the breakdown of responses for these participants: 31% of responses indicating less than 50% confidence for hard questions in the two-stage condition, 24% for medium, and 54% for easy.

	Question Difficulty Level		
	Easy	Medium	Hard
< 50% confidence	7 responses	3 responses	4 responses
≥ 50% confidence	6 responses	10 responses	9 responses

Table S 8. (Study 3) Response Breakdown from Study 2 Participants who saw Study 3 Question Order

Participants in the two-stage condition who put less than 50% confidence in their selected answer choice in the second stage read on the following page:

> "You indicated you were X% confident in your selected answer 'Image [A/B] has more dots.' This implies you're 1-X% confident in the answer you did not select (e.g., that you're 1-X% confident in 'Image [B/A] has more dots')."

They were then asked to answer three follow-up questions. The breakdown of responses is shown in Table S 9 below.

Two-stage Responses to Follow-up Questions						
	Easy (Q2)		Medium (Q1)		Hard (Q3)	
	Yes	No	Yes	No	Yes	No
Is "Image [A/B] has more dots" the answer option you meant to select?	1	1	7	1	17	3
Is X% the percent confidence you meant to indicate in your selected answer "Image [A/B] has more dots"?	0	2	6	2	17	3
Would you like to change your answer to "Image [B/A] has more dots"?	0	2	2	6	4	16
Number of Responses <50	2		8		20	

Total Responses (N = 100)

Table S 9. (Study 3) Responses to Follow-ups by Question Difficulty Level

Given there are so few responses that indicate less than 50% confidence for the easy questions, I hesitate from extrapolating too much from the responses to the follow-up questions. I do note that across the medium and hard questions, it looks like the majority of respondents indicate that yes, they meant to select their indicated answer option, that yes, they meant the percent confidence they indicated, and that no, they would not like to change their answer.

I examined what participants who answered less than 50% confidence for the first question (medium difficulty) answered for the following two questions. The 8 respondents, 6 of whom say they wouldn't change their answer, answer 50% or more in the subsequent two questions, including the hard question. This might imply they are correcting themselves for subsequent questions, despite not wanting to change their answer for the first question. This could contribute to the lower total proportion of responses indicating less than 50% confidence in this as compared to prior studies.

Interestingly, this also implies that the 20 responses for the hard question that indicate less than 50% confidence are responses from 20 new participants – these are participants who answered 50% or more for the medium and easy questions, then saw the hard difficulty question and answered less than 50% confidence. This may relate to research showing that responses can be influenced by normatively irrelevant order variations (Ariely et al., 2005); in this case, responses on the easy and medium questions may have influenced participant responses on the hard question through anchoring.

After the three follow-up questions, participants were given a text box where they could choose to explain their thinking. I highlight the responses I find intriguing below:

1. Medium Question (N = 8): "**I still think B has more dots but my confidence level is very, very low.** ; I think Image B has more dots but without actually counting them it's really hard to be sure.; I still think Image B has more dots.; **I have very low confidence in my choice, so I chose a low percentage, but I did mean to choose Image B as appearing to have more dots.** I hindsight, I should have chosen something like 65% confidence.; According to my opinion I choose this option ; **I have never thought about confidence rating in that way before, that it would imply I'm the remainder confident in the other answer.** I do think B has more, it just looked very similar in number **so I was not very confident, which is what I was trying to express by the 15% rating.**"
2. Easy Question (N = 2): "The clusters always make me think there's more but i am uncertain of my answers; I still think that A has more dots, but **im just not super sure about it**"
3. Hard Question (N = 20): "I was 38% confident, meaning not very confident, but my choice was B; I judged by what looked more populated at a glance.; I think I am right but now I am not sure but I will stick with my first gut choice.; **I indicated I had 40% confidence in Option A, because I am only somewhat confident in my answer.** I still lean towards Image A having more dots, but I would not be surprised at all if the correct answer was Image B. It was a toss up for me but I was leaning towards image A. **I have low confidence that I was correct <u>but that doesn't mean I am more confident that image B has more dots.</u>**; my thought process was that image be had more dots, but i was not completely sure. i was not fully trusting my judgement.; The images were hard to tell as they both had a lot of dots so I as not sure; I guess it's a matter of interpretation - it seemed like a toss-up to me (i.e. they both looked to have the same amount of dots) - so my confidence that my choice was correct was rather low. **I realize that in that case, you could argue that the % chance my choice was correct should be 50% at least, but I see that as slightly different from \"confidence\".**; I feel like A has more dots but

> confidence is low; great; **I didn't realize the confidence was relative to each other I just was saying out of 100 I'm not super confident that I'm right** ; I feel that the grouped dots amounted to more, but it looked fairly close to being the same without counting each individual dot.; I just felt like I'm not confident at all it's either one but I guess I should have just said 50% in that case but **I kind of feel like 0% would mean I have no idea and I felt like 30% confidence doesn't necessarily mean that I think the other is 70% if that makes sense.**; I will stick with my gut.; I indicated my percentage of being sure in my answer- I meant that I was not very sure at all. **It does NOT mean that the other answer is more likely to be right.**; I should have said 67% for that not 33% sorry; **I didn't take the confidence measure that the other had more of an implication of having more, I took it as I was only 21% sure of my answer since they were very similar**; Those looked as though they had about the same number of dots, **but there wasn't a selection for that**. I suppose that next time I'll just select 50."

Exploring the data further, I filter the dataset and remove two-stage responses that indicate less than 50% confidence in the selected answer, and compare peak confidence by difficulty level against the one-stage responses. Notably, this should not influence results much if at all for easy and medium level questions, where there were only 2 and 8 responses under 50% respectively, but may influence results for the hard level question, given there were 20 responses with less than 50% confidence. With this restricted data set, the observed pattern goes away. Average peak confidence is not significantly different between conditions, for any level of difficulty. Figure S 19 depicts these results.

1. For the easy questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 76.69, *SD* = 17.38) and the two-stage condition (*M* = 78.15, *SD* = 16.51), *t*(186.97) = -0.60, *p* = .552.
2. For the medium questions, there is a significant difference between average peak confidence in the one-stage condition (*M* = 69.65, *SD* = 15.54) and the two-stage condition (*M* = 70.14, *SD* = 15.79), *t*(182.87) = -0.21, *p* = .831.
3. For the hard questions, there is no significant difference between average peak confidence in the one-stage condition (*M* = 64.95, *SD* = 15.63) and the two-stage condition (*M* = 65.21, *SD* = 14.75), *t*(169.51) = -0.11, *p* = .911.

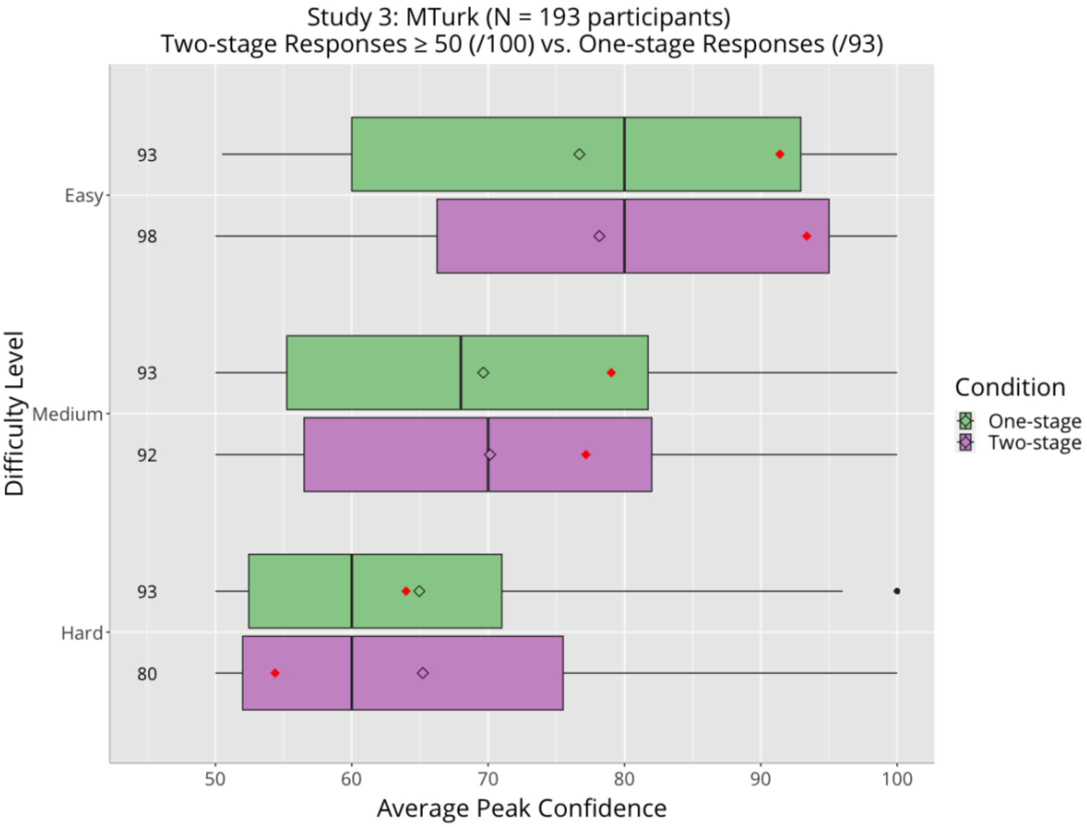


Figure S 19. (Study 3) Boxplot showing average peak confidence by question difficulty level and condition, using filtered dataset. Red diamonds indicate accuracy, transparent diamonds indicate mean.

6) Other Notes

With respect to the standard attention question, 1 participant reported paying no attention, 12 a lot of attention, and 180 paying full attention.

In both conditions, participants generally reported the survey was asking about dots, confidence, judgement, decision making, estimation, etc. Several participants also wrote about visual judgement, perceptions, visual reasoning/skill, cues, and the like. The vast majority of responses to the question asking whether anything was strange or unclear about the survey was no, with three notable comments:

- I was tempted to choose 50:50 because that what I usually thought at first, but then I tried to evaluate better and chose different numbers. *(one-stage condition)*
- **Yeah the questions about how confident I was in my answer it shouldn't have been relative to the other answer, I HAD to select one but sometimes I wasn't sure so if I said 30% confidence it shouldn't mean I was 70% confident the other option was right.** *(two-stage condition)*
- **Well maybe how the fact that just because I was less than 50% confident it meant I believe the other one more.** *(two-stage condition)*

Discussion

In summary, the exploratory analyses show no main effect of condition, and an interaction between condition and difficulty such that average peak confidence is higher in the one-stage condition, but only for hard questions. This partially replicates the results from Studies 1 and 2, and implies that future studies might benefit from a focus on hard questions.

These patterns do not hold when two-stage responses that indicate less than 50% confidence in the selected answer are removed from the dataset – the effect goes away such that there is no difference in peak confidence between conditions for either difficulty level. This does not replicate our findings from Studies 1 or 2, suggesting that though worth exploring more, subsetting the data in this way may not produce a predictable and reliable effect.

Supplemental Analyses for Pre-registered Studies

Study 4: Main effect of difficulty

Collapsing across question formats, I find that average peak confidence is 70.06% (*SD* = 21.50) for the easy question, 62.16% (*SD* = 20.94) for the medium question, and 56.41% (*SD* = 21.99) for the hard question. Figure S 20 depicts average peak confidence by difficulty level, with questions collapsed across conditions.

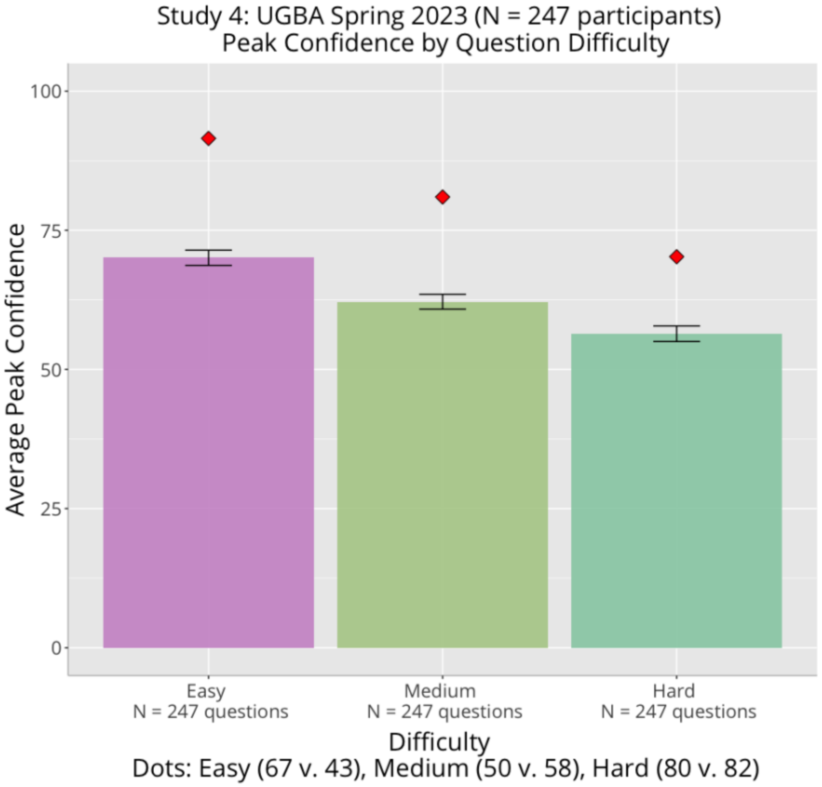


Figure S 20. (Study 4) Average peak confidence by question difficulty level. Red diamonds indicate average accuracy.

As an additional check on question difficulty, I note the average time taken to click the submit button for a given question. For the one-stage questions, the timer measured the time the participants spent on the page allocating confidence across the two answer options. For the two-stage questions, the timer measured the time participants spent on the page where they chose between the two answer options. Table S 10 below provides the summary statistics. Note that data was windsorized at the 5th and 95th percentiles.

Condition	Difficulty	N	Mean	SD	Min	Max
One-stage	Easy	124	16.79	13.33	5.89	56.91
One-stage	Medium	124	19.80	15.31	5.91	59.08
One-stage	Hard	124	20.28	18.43	5.77	71.85
Two-stage	Easy	123	7.82	3.08	5.58	17.01
Two-stage	Medium	123	8.94	4.44	5.72	21.45
Two-stage	Hard	123	11.02	5.93	5.75	26.45

Table S 10. (Study 4) Summary Statistics on Time Per Question in seconds.

Study 5: Main effect of difficulty

Collapsing across questions, I find that average peak confidence is 76.92% (*SD* = 17.72) for the easy question, 66.63% (*SD* = 18.32) for the medium question, and 61.56% (*SD* = 17.85) for the hard question. Figure S 21 depicts average peak confidence by difficulty level, with questions collapsed across conditions.

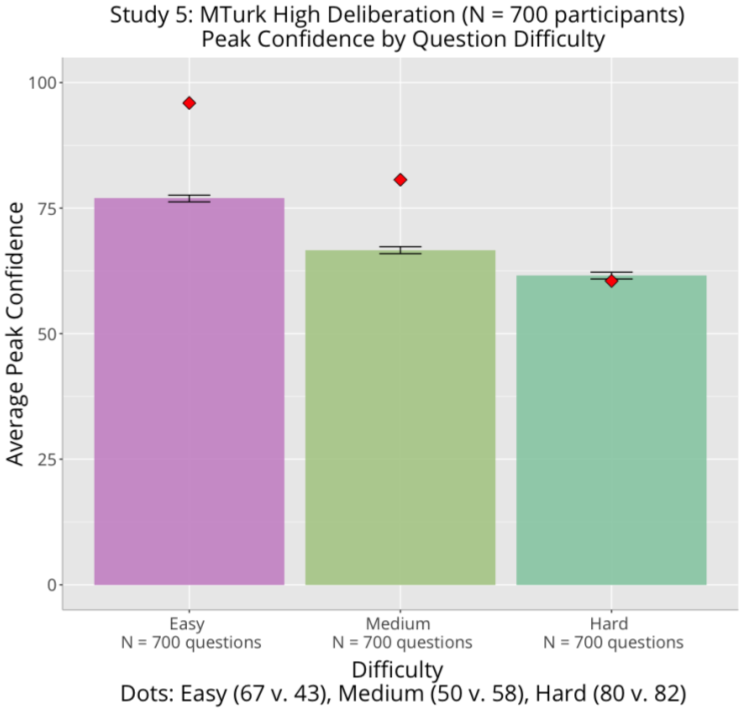


Figure S 21. (Study 5) Average peak confidence by question difficulty level. Red diamonds indicate average accuracy.

As an additional check on question difficulty, I note the average time taken to click the submit button for a given question. For the one-stage questions, the timer measured the time the participants spent on the page allocating confidence across the two answer options. For the two-stage questions, the timer measured the time participants spent on the page where they chose between the two answer options. Table S 11 below provides the summary statistics. Note that data was windsorized at the 5th and 95th percentiles.

Condition	Difficulty	N	Mean	SD	Min	Max
One-stage	Easy	346	17.63	11.04	6.10	45.93
One-stage	Medium	346	19.38	12.81	6.18	53.35
One-stage	Hard	346	20.66	14.32	6.28	60.15
Two-stage	Easy	354	8.66	3.98	5.73	21.10
Two-stage	Medium	354	10.41	6.48	5.72	30.10
Two-stage	Hard	354	10.94	5.86	5.75	26.11

Table S 11. (Study 5) Summary Statistics on Time Per Question in seconds.

Study 6: Main effect of difficulty

Collapsing across questions, I find that average peak confidence is 72.13% (*SD* = 15.76) for the easy question, 64.43% (*SD* = 17.24) for the medium question, and 59.71% (*SD* = 17.31) for the hard question. Figure S 22 depicts average peak confidence by difficulty level, with questions collapsed across conditions.

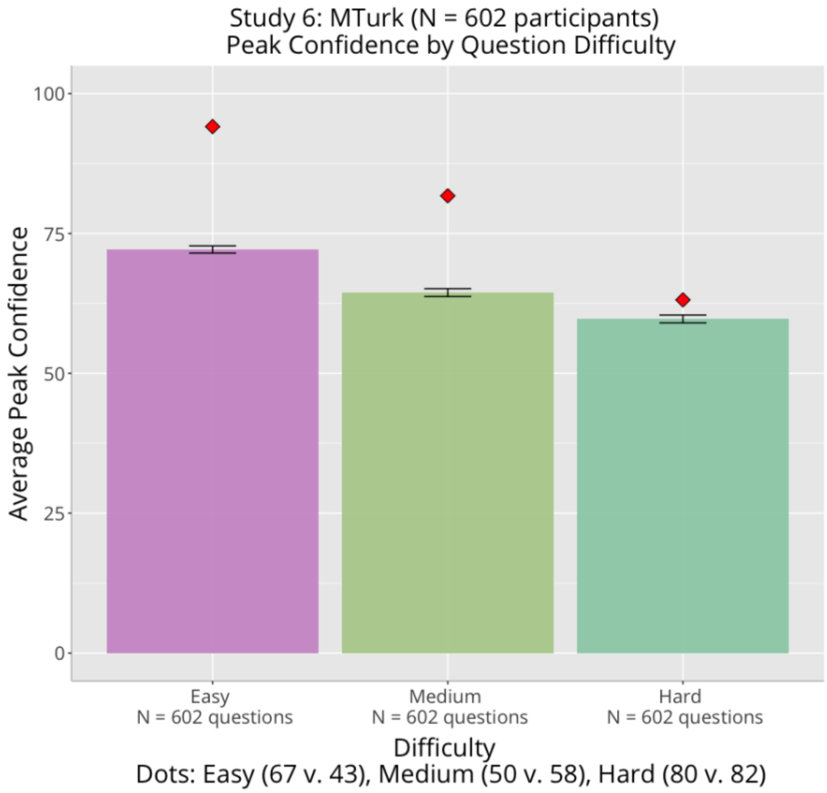


Figure S 22. (Study 6) Average peak confidence by question difficulty level. Red diamonds indicate average accuracy.

As an additional check on question difficulty, I note the average time taken to click the submit button for a given question. For the one-stage questions, the timer measured the time the participants spent on the page allocating confidence across the two answer options. For the two-stage questions, the timer measured the time participants spent on the page where they chose between the two answer options (Step: Choice), and another timer measured the time participants spent on the page where they reported confidence in their chosen answer (Step: Confidence). Table S 12below provides the summary statistics. Note that data was windsorized at the 5th and 95th percentiles.

Condition	Step	Difficulty	N	Mean	SD	Min	Max
One-stage	-	Easy	187	15.22	9.96	4.41	42.28
One-stage	-	Medium	187	17.50	11.86	4.90	49.85
One-stage	-	Hard	187	17.54	11.24	4.50	44.54
Two-stage	Choice	Easy	415	7.42	4.22	3.73	19.75
Two-stage	Choice	Medium	415	7.90	4.23	3.75	19.05
Two-stage	Choice	Hard	415	9.53	6.05	3.78	27.10
Two-stage	Confidence	Easy	415	6.43	3.14	3.67	14.76
Two-stage	Confidence	Medium	415	5.87	2.33	3.62	12.26
Two-stage	Confidence	Hard	415	6.39	3.07	3.66	14.49

Table S 12. (Study 6) Summary Statistics on Time Per Question in seconds.

Study 7: Main effect of difficulty

Collapsing across questions, I find that average peak confidence is 91.24% (*SD* = 12.04) for the easy question and 67.16% (*SD* = 19.22) for the hard question. Figure S 1 depicts average peak confidence by difficulty level, with questions collapsed across conditions.

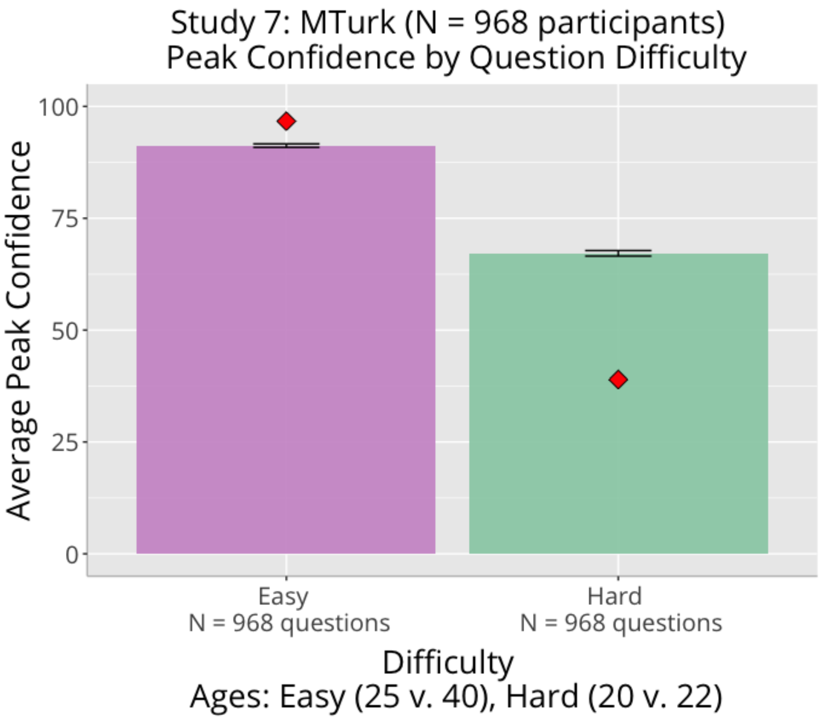


Figure S 23. (Study 7) Average peak confidence by question difficulty level. Red diamonds indicate average accuracy.

As an additional check on question difficulty, I note the average time taken to click the submit button for a given question. For the one-stage questions, the timer measured the time the participants spent on the page allocating confidence across the two answer options. For the two-stage questions, the timer measured the time participants spent on the page where they chose between the two answer options (Step: Choice), and another timer measured the time participants spent on the page where they reported confidence in their chosen answer (Step: Confidence). Table S 13 below provides the summary statistics. Note that data was windsorized at the 5th and 95th percentiles.

Condition	Step	Difficulty	N	Mean	SD	Min	Max
One-stage	-	Easy	489	20.68	11.50	6.52	50.35
One-stage	-	Hard	489	26.68	15.35	7.29	60.89
Two-stage	Choice	Easy	479	9.95	5.49	4.08	24.91
Two-stage	Choice	Hard	479	15.52	9.92	4.69	42.68
Two-stage	Confidence	Easy	479	6.64	3.27	3.78	15.85
Two-stage	Confidence	Hard	479	6.97	3.64	3.77	17.06

Table S 13. (Study 7) Summary Statistics on Time Per Question in seconds.