MASTERARBEIT | MASTER'S THESIS

Titel | Title

Confidence and Accuracy

Exploring the Effects of Confidence Judgement Methods
in Two-Alternative Forced-Choice Tasks

verfasst von | submitted by

Filippus Nemestothy BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt	Degree programme code as it appears on the student record sheet:	UA 066 914
Studienrichtung lt. Studienblatt	Degree programme as it appears on the student record sheet:	Masterstudium Internationale Betriebswirtschaft
Betreut von	Supervisor:	Assoz. Prof. Steffen Keck PhD

Acknowledgments

Dedicated to my parents who are a source of support, inspiration and love.

My name stood in the dedications of my father's thesis already a long time ago. It is a pleasure to put his name in mine: Thank you, dear Petsku.

Without the creativity, humor, laughter, inspiration and love of my mother this world would be less interesting and colorful: Kiitos, rakas Pia.

Immensely thankful for all those patient and supportive. Be it family, partner, friends, the state or professors.

Abstract

Doctors need it just as well as business managers: Confidence calibration, the measure of certainty with which a judgement has been made. Confidence calibration methods are used to hone forecasts, identify correct medical treatments and elicit expert decisions. The topic is reported to have implications in most fields of research and areas of everyday life, yet the body of research on measurement methods of confidence calibration accuracy shows inconsistent results. The traditional half-range method (50-100) and the full-range method (0-100), distinct by the range of their applied scale and commonly applied in two-alternative tasks, have each been reported superior to the other. This study proposed the half-range method to be superior, thus yielding more accurate calibration. Furthermore, a novel alteration of the traditional methods is developed with the intent to improve calibration accuracy: the dual-opposite (100-50-100) method. In an experimental study with 232 participants, all three confidence judgment methods were tested using two-alternative forced-choice general knowledge questions, and Brier scores were calculated for each method. The analysis confirmed the superiority of the half-range method, though the novel method did not improve nor worsen calibration accuracy. This study concludes with a critical discussion of the findings and offers suggestions for future research directions.

Keywords:

Metacognition, confidence calibration, human judgement, confidence judgements, half-range method, full-range method, anchoring, framing, performance monitoring, decision processing

Table of contents

1. Introduction

Ranging from a "bad boy"–persona in economic research (Litvinova et al., 2019, p. 3; Griffin & Brenner, 2004), "one of the oldest and most intriguing problems in experimental psychology" (Baranski & Petrusic, 1995, p.388), often introduced in context with some of the most disastrous political wrong decisions in history (see Walters et al., 2017) and yet still prevailing in academic interest, is the topic of confidence calibration.

Due to the fact that the research on confidence calibration, the relationship between the correctness of a judgement and the degree of certainty with which it was made, is an already aged topic in experimental psychology (see e.g. Peirce & Jastrow, 1884), we may question why we would still care about research on confidence calibration and what reason deems it to be relevant? Confidence calibration relates to any domain of judgements and decision making: From managerial, political, medical decisions to educational insights, forensic and criminal investigations and weather forecasts to artificial intelligence – just to name a few. Thus, new or updated insights on calibration can have a plethora of applications in academics as well as in practice (Litvinova et al., 2019).

However, the study of confidence calibration faces many challenges due to the biased human nature. Two of the undoubtedly best known names in the academia of this field, Daniel Kahneman and Amon Tversky, proclaimed that "*subjective confidence judgments violate coherence norms of rationality*" (Kahneman & Tversky, 1982; found in Litvinova et al., 2019) and yet the noble laureates[1] found that "errors of judgment are often systematic rather than random" (Kahneman & Tversky, 1977, p. I). Exploring this system and investigating a possible measure for improvement in the process of judgement and confidence calibration is the motivation for this master thesis.

[1] Despite Daniel Kahneman receiving the 2002 Nobel Prize in Economic Sciences together with Vernon L. Smith, does Kahneman point out, that had Tversky not already passed away, would they have been awarded Nobel Prize together for their work on Prospect Theory. (Kahneman, 2011, p. 10)

In order to investigate this system, we need measures. The goodness of calibration as the subjective probability of being right can be measured with a mean probability score, the Brier score (Brier, 1950). While this study's experiment will focus on the Brier score, further relevant measures for calibration to be mentioned are resolution and overconfidence. Resolution is the ability of assessors to distinguish the correctness of a task or occurrence of an event (Murphy, 1973). Overconfidence is the definition of subjective probability of being right exceeding the objective event of being correct. Different tasks and experiment setups naturally produce results with variably well calibrated respondents. As general knowledge questions are frequently used for research in calibration (Luna & Martín-Luengo, 2012), this masters' thesis will retain with this tradition in conducting a survey with general knowledge questions while applying three different methods of confidence judgement: The traditional half-range and full-range method are compared with a method developed for this study which will go by the working title of the "dual-opposite" method. The goal is to examine an improved confidence judgement method to produce more accurate calibration judgements.

To establish a line of argument along a common thread we will introduce the status quo of selected human biases which may affect the outcomes of confidence calibrations. Well-proven research methods and experiment set-ups will be presented – introducing the underlying for research gap for this work. Finally, the results of a conducted survey experiment will be analyzed and discussed.

"There are two kinds of forecasters: those who don't know, and those who don't know they don't know." Maurizio Bovi (2009) borrowed these words once from John Kenneth Galbraith (Wall Street Journal, Jan 22, 1993) to prelude his own paper on economic versus psychological forecasting. Maybe this master's thesis will help to understand both kinds of forecasters better. Those that are well calibrated and those that are not.

2. Literature Review

This chapter will provide an overview of the current state and history of research on confidence calibration in the realm of human judgement. The research set ups which have been employed with focus on those utilizing general knowledge questions and the biases in human judgement which affect level of confidence.

Building on this basis, this chapter will conclude by identifying a research gap, which this work will treat subsequently.

2.1 Confidence Calibration

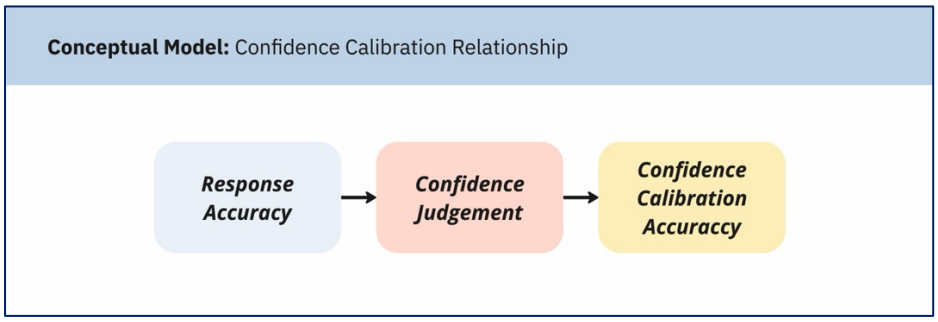
Overall confidence calibration is home within the field of metacognition and human judgement. However, corresponding to the miscellaneous domains where confidence calibration matters, a respective number of scholars from various academic fields have delved into the topic, sometimes using different names for confidence calibration. Lichtenstein et al. (1982) name just a few of the previously used definitions for confidence calibration, as "it has also been called *realism* (Brown & Shuford, 1973), *external validity* (Brown & Shuford, 1973), *realism of confidence* (Adams & Adams, 1961), *appropriateness of confidence* (Oskamp, 1962), *secondary validity* (Murphy & Winkler, 1971), *and reliability* (Murphy, 1973)." In the psychophysical comparison literature we find the *accuracy-confidence relationship* (Baranski & Petrusic, 1995).

This thesis will employ the terms that appear as the standard in the literature (Lichtenstein et al., 1982; Litvinova et al., 2019; Luna & Martín-Luengo, 2012; Walters et al., 2017), confidence calibration, confidence judgements and response accuracy.

Generally, confidence calibration describes the relationship between the correctness of a judgement and the level of certainty with which it was made (Baranski & Petrusic, 1995). We call the statement about one's certainty level the confidence judgement.

In simple terms, one possibility to extrude confidence calibrations is thus asking a person a general knowledge question and in a second step asking how sure about the answer the person is, where the latter is the confidence judgement. If a person gives a correct answer and is absolutely sure about being correct, we speak of good calibration. At this level we have a simple relationship between the components of response (and its accuracy) and confidence judgement, resulting in confidence calibration with a certain calibration accuracy, see Figure 1. In the following chapters this relationship will be successively expanded by more factors into a more complex model.

Figure 1 – Confidence Calibration Relationship - Conceptual Model



When rating the quality of calibration accuracy scholars have produced a variety of models and normative indices in different context, especially with the popularity of machine learning and the necessity to calibrate the output of artificial intelligence models as well. To name a few we find indices like the Expected Calibration Error, Maximum Calibration Error and the Brier score (Huang et al., 2020). The Brier score emerges as a commonly applied measure in the domain of human judgement and probabilistic forecasts (Ferro & Fricker, 2012), which we will apply in our study.

2.1.1 Brier Score

The Brier score or mean probability score, is a mean squared error function which Glenn Brier proposed in the Monthly Weather Review in January 1950. Initially developed to improve weather forecasts by evaluating the confidence by which meteorologists predict e.g. rain for tomorrow, "[the Brier score] is used to index the accuracy of a person's assessment of probabilities" (Baranski & Petrusic, 1995).

$$\boldsymbol{BS} = \frac{1}{N}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(p_i - x_i)^2$$

The Brier score measures the mean squared deviation between confidence judgements p_i that events x_i will occur for *N* items. p_i can take values between 0 and 1, depicting the forecasters subjective probability assigned in percent. x_i takes the value 0 if an event did not occur and the value 1 if an event occurred. Thus, the smallest error (or deviation) expresses the best calibrated judgements and yields the Brier score value of 0, whereas a value of 1 depicts the worst possible score. Suggest, a forecaster makes a random a guess between two alternative forecasts and gives a confidence judgement of 0.5, the Brier score would yield 0.25. See *Table 1* for calculation examples of the Brier Score.

Table 1 – Example of Brier scores in prediction

Respondent	**Question *i***	**Prediction p_i**	**Occurrence x_i**	**Brier Score**
A	Q1	1.00	1	0.00
A	Q2	0.00	0	0.00
A	Q3	0.50	1	0.25
A	Q4	0.21	0	0.04
A	Q5	0.68	1	0.10
			Final Brier Score	**0.08**

Apart from the application for predictions the Brier score has found further use in measuring confidence calibration in metacognitive tasks like general knowledge questions with binary or categorical outcomes. In this case the component x_i represents the correctness of an answer instead of the occurrence of an event, thus it can take the binary values 1 for a correct answer and 0 for a false answer. The confidence judgement p_i represents the subjective confidence of being correct instead of the subjective assessment of the probability for the occurrence of an event. A respondent who has all answers correct with 100% confidence, would yield a Brier Score of 0, thus being perfectly calibrated. Note, that this logic is oppositional to the English language. A respondent with all answers wrong and no confidence at all in knowing the answers (0% in all confidence judgements) would yield the same Brier Score of 0, thus "knowing well that he or she does not know". In Table 2 a paradigmatic calculation for a case with general knowledge questions is shown.

Table 2 - Example of Brier Scores for Metacognitive Tasks

Respondent	**Question i**	**Confidence p_i**	**Correctness x_i**	**Brier Score**
B	Q1	0.08	1	0.85
B	Q2	1.00	0	1.00
B	Q3	0.76	1	0.06
			Final Brier Score	**0.63**

Allan H. Murphy (1973) proposed a decomposition of the Brier score into three additive parts: uncertainty, reliability and resolution. This decomposition allows more detailed interpretations on potential origins of prediction errors and can help to identify where models need improvement.

$$BS = Uncertainty + Reliability - Resolution$$

In Murphy's decomposition uncertainty explains the inherent difficulty of the prediction of an event itself and is defined by the base rate of the actual outcomes appearance. In other words, uncertainty can be understood as the true probability of an events' occurrence. Hence, for binary events with even

probability of occurrence (50%), uncertainty would be at maximum, as it would always be a 50% chance to guess the occurrence correctly. Applied on general knowledge questions, we could therefore say, that uncertainty is minimal, as long as the answer to the general knowledge questions hold true under all circumstances.

Reliability measures how close the predictions of respondents come to the actual event frequencies. For this component predictions with a similar forecasted probability are grouped together. Hence, a model is reliable when the observed frequency of occurrence of an event matches the predictions of grouped predictions. If we imagine an example, where we group all forecasts which expected rain with a 75% chance, the model is perfectly reliable when rain appeared really in 3 of 4 cases.

Finally, resolution looks at the ability of forecasters to make the right probability forecast for different events with different occurrence probabilities. In other words, how well forecasters can differentiate their forecasts for varying events. For resolution, forecasts with similar forecast probabilities are grouped. The difference of each groups' average prediction probability to the overall base rate of event probability is calculated and all differences added up. The higher the resolution term, the better.

Ferro & Fricker (2012) suggest a further improved bias-corrected Brier score decomposition. However, for terms of comparability and applicability with the previous literature, which has applied the original Brier score, and for reasons of parsimony this thesis will remain with the traditional Brier score and refrain from the decompositions. The observance of the Brier score decomposition would be recommended for studies on forecasts. In the case of general knowledge questions and models which do not differ in probability of occurrence the traditional Brier score will suffice for the comparison of the confidence judgement method.

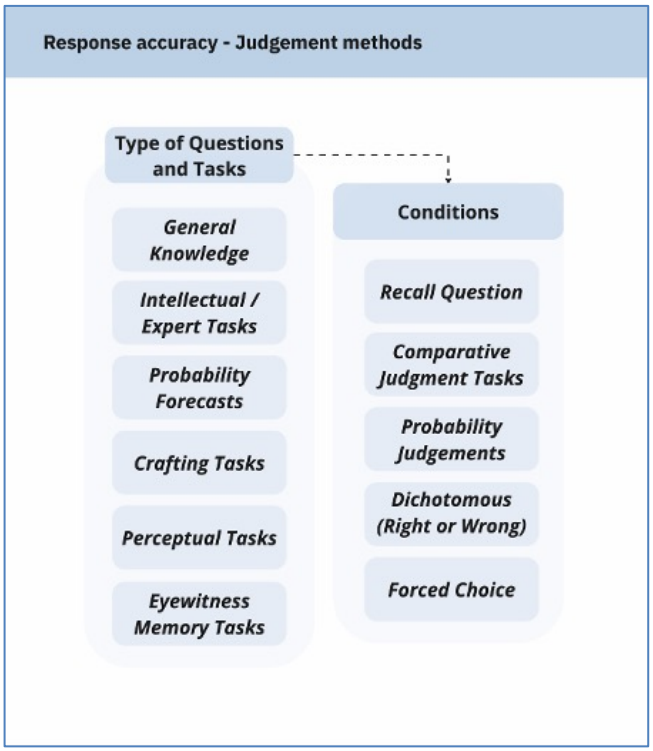
In order to produce a Brier score, as the measure of goodness of confidence calibration, we will nead to look at two components: The elicitation of respondents' response accuracy through a task and a confidence judgement to get a level of certainty with which a respective task was solved.

2.1.2 Response accuracy

The tasks which can be followed up by a confidence judgment are versatile and scholars have applied an impressively creative bandwidth of different study setups to assess response accuracy, the level of how correct an answer to a question or the solution to a task was.

"From what fruit is wine normally made? Grapes, Oranges, Cherries or Apples?" From 1364 general knowledge questions normed for a Spanish population, all 400 respondents knew the answer to this multiple-choice question is Grapes (Buades-Sitjar et al., 2021). This type of question is an example for a 4-alternative forced-choice question about general knowledge. General knowledge questions constitute a paradigmatic task for eliciting response accuracy and can be setup in research experiments in many ways. For our study we will also deploy general knowledge tasks. However, scholars have used a variety of other tasks to measure the correctness or goodness of a solved task. These can take the form of perceptual, eye-witness and crafting tasks as well as more intellectual metacognitive tasks like general knowledge questions, probability forecasts or expert knowledge assessments. In the next chapters we will embark on short introductions to several research setups and confidence judgement methods with the objective of finding an opportunity for improvement within this research. Figure 2 sketches an overview of different approaches for testing response accuracy.

Figure 2 - Response methods



Forecasts of future events

Confidence calibration has a long history, if not even starting grounds in the meteorological academia around weather forecasts. Therefore, weather forecasts represent a classical event of forecasting future events. Apart from these further classical examples are predictions electoral outcomes and financial market developments. As reported by Lichtenstein et al. (1982) as early as in 1906 an Australian official astronomer proposed to combine weather forecasts with an indicator of subjective probability by the forecasters themselves. Thus, Cooke (1906, p. 23) suggested already in 1906 the use of confidence calibration for forecast models, when he wrote: "It seems to me that the condition of confidence or otherwise forms a very important part of the prediction, and ought to find expression." Following Cooke's suggestion a number of meteorologists began taking confidence judgements into account, including Cooke himself, Williams (who noted expressions of overconfidence), Murphy and Winkler, Sanders among others (comp. Lichtenstein et al., 1982, p. 310). One of the more important figures from the group of meteorologists for underlying work is Glenn W. Brier, who in 1950, proposed the scoring rule for the accuracy of predictions, which we now know as the Brier score (see Brier, 1950). Taking confidence judgements into account improved weather forecasts significantly according to Lichtenstein et al. (1982).

Perceptual judgements

Different perceptual tasks have been tested by scholars. In early experiments Adams (1957) showed respondents 40 different words on computer screens with differently well or badly lit screens. Respondents were asked to describe the word they recognized and give a probability judgement on the certainty that they recognized the word correctly. Respondents generally performed better than they expected themselves, as the comparison of the correct answers and the subsequent confidence judgement showed. In other words, the respondents showed underconfidence in this perceptual word

recognition task. This introduces the concept of over- and underconfidence, which plays an important role in many studies in the underlying field. We will give attention to this concept in a separate chapter.

Similarly to Adams (1957) study, subjects showed over- or respectively underconfidence in a different perceptual task setup by Baranski & Petrusic (1994). In three experiments, the researchers showed respondents visual stimuli in the form of two differently long lines on a computer screen. Respondents were tasked to recognize the longer line with varying difficulty (smaller or bigger length difference). In the case of bigger differences in length, subjects were rather underconfident, for smaller differences they were overconfident. Thus, for easy task judgements underconfidence was registered, and respectively for difficult tasks overconfidence (Baranski & Petrusic, 1994a). The hard-easy effect will later be covered separately.

Baranski & Petrusic (1994, p. 418) further reported the "well-known inverse relation between response time and confidence […]" in settings where respondents where incentivized to maximize accuracy. Longer response times were matched with less confident assessments, whereas short response times were met with higher confidence. This effect was accentuated by the difficulty category of the tasks, which the authors called the *response time difficulty effect* (p. 418). This effect showed the easer the question category the shorter were response times. Notably, when participants are incentivized to maximize response speed, "caution is sacrificed" and the response time difficulty effect disappears (Baranski & Petrusic, 1994, p. 425).

Eyewitness-memory

Confidence calibration has been studied extensively in eyewitness-memory and recognition memory tests often with the aim of supporting the distinguishment of correct and incorrect answers in interrogation (Luna & Martín-Luengo, 2012). Their research comparing the confidence-accuracy relationship across general knowledge and eyewitness memory tasks demonstrated a stronger

correlation between confidence and accuracy in general knowledge tasks than in eyewitness memory tasks.

In another face recognition study Weber & Brewer (2003) compared confidence judgement methods. 48 study participants had to recognize viewed pictures of faces as old or new and next indicate their confidence estimates by clicking on buttons on either the half-range scale (50%-100%, in 10% increments) or the full-range scale (0%-100%, in 10% increments). What Weber & Brewer (2003) found was that the half-range method was superior to the full-range method in terms of produced calibration accuracy.

General Knowledge Questions

As mentioned above scholars like Weber & Brewer (2003) among many others (Walters et al. 2017; Attali et al., 2020; Coane et al., 2023; Alba & Hutchinson, 2000; Whitcomb et al., 1995; Gigerenzer et al., 1991; Luna & Martín-Luengo, 2012; Ronis & Yates, 1987) conducted studies on confidence calibration with the application of general knowledge questions. As we are going to apply these for the experiment for this thesis we will expound on characteristics of general knowledge questions.

Overall, for intellectual tasks like general knowledge questions, we can differentiate tasks by the number of discrete propositions (comp. Lichtenstein et al., 1982, p. 307).

No alternatives: Recall questions demand the respondent to reproduce the answer from memory and in a second step give an assessment of how probably she believes the answer to be correct. The latter probability judgement is appropriately answered on a range between 0 and 1. An example question for a no-alternative recall question is “Who invented the printing press around 1440?”.

One alternative: In this type of question one answer alternative is provided and assessors are asked to respond if the given answer option is true or false. A confidence judgement can wether be provided in a second step or within the first question itself by asking for the probability of the

answer option being true. Hence, the example would read “Guttenberg invented the printing press around 1440. How sure are you, that this statement is true?” or “What is the probability that Gutenberg invented the printing press?”

Two alternatives: A question is offered with two answer options, one correct and the other as a lure, and the respondents are asked to choose the correct answer and then give a confidence judgement for their certainty of being correct on a range between 0.5 and 1. In this *half-range* called method choosing 0.5 would mean respondents have no clue and are guessing, values below 0.5 are not eligible after having chosen one of the options already, as this would mean, that the respondent tends to the other answer option to be correct. Optionally, in the *full-range* method one of the answer options can be prechosen and the respondent asked on a range of 0 to 1 about the probability that the prechosen answer is correct. In this case a value of 0 would mean confidence in the other option to be correct. The above example question would reform in the half-range method to “The printing press was invented around 1440 by a) Gutenberg or b) Erasmus. Which answer do you think is correct? How confident are you about the correctness of your answer on a scale from 0.5 to 1.0?“ or accordingly with the full-range method “How sure are you that option a) Gutenberg is correct on a scale from 0.0 to 1.0?”

Three or more alternatives: When a question comes with more than two answer options respondents can be given two ways of providing an answer. On the one hand, respondents, like in the half-range method, can first select their most likely correct answer and in the second step give a probability judgement for their correctness in the range of $\geq 1/n$, where n is the number of given answer options. On the other hand, respondents can assess a probability for all answer options in the range of 0 to 1. See our example: “The printing press was invented around 1440 by a) Gutenberg; b) Erasmus; c) Luther or d) Gauss. Which answer do you think is correct? How confident are you about the correctness of your answer on a scale from 0.25 to 1.0?“

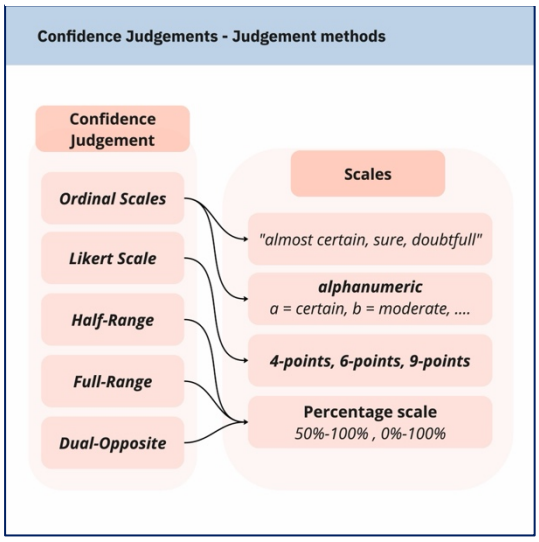
Litvinova et al. (2019) and Gigerenzer et al. (1991) point out that miscalibration in confidence judgements does not only result from respondents characteristics but from shortcomings in the

representativeness of general knowledge questions. Hence great attention will have to be paid to selecting a unbiased battery of questions for our experiment. The consideration put into the selection of the general knowledge questions for our survey are explained in later chapter on the survey process.

2.1.3 Confidence judgments

Next to the different possibilities to elicit response accuracy assessments the other essential element for confidence calibration are confidence judgements.

Figure 3 - Confidence Judgement Methods



Confidence judgements can be made with above mentioned methods like the half-range and the full-range method on percentage scales from 50%-100% and 0%-100% (Adams, 1957; Luna & Martín-Luengo, 2012; Weber & Brewer, 2003). Other scholars have used ordinal confidence judgements with different degrees like *"almost certain", "normal probability or "doubtful"* (Cooke, 1906). As stated by (Baranski & Petrusic, 1994a), "*many of the classic psychophysical comparison studies employed alphanumeric-based confidence scales (e.g., a, b, c, d; where a = certain, b = moderate certainty, c = little certainty, and d = guess).*" For calibration of the feeling-of-knowing some scholars ask respondents for the percentage of confidence of the right answer (see e.g. Walters et al., 2017), while others use a Likert-scale (Kawaguchi, 1992), however the concept of feeling-of-knowing differs from the concept of confidence judgements. The feeling-of-knowing refers to the likelihood to recall a piece of information in the future, whereas confidence calibration refers to the accuracy of how well a judgement meets the true value. According to Cummings (2006) the majority of calibration studies are conducted with the use of Likert scales, entailed with the difficulty of not being able to translate Likert scales directly into probabilities of being correct.

Correspondingly, confidence judgements on Likert scales, and even more so in ordinal categories, cannot be directly compared with the true probabilities of occurrence of an event in order to produce a confidence calibration score.

It was mentioned above, that Weber & Brewer (2003) found the half-range method to be superior to the full-range method. Yet Baranski & Petrusic (1994) note respondents' desire to apply a full-range confidence judgement. The scholars continued in their study on perceptual judgements to "[employ] *a confidence scale ranging from 0% (certainty of an error) to 100% (certainty of a correct response) in order to permit an investigation of the relationship between subjective errors (i.e., confidence between 0%and 49%) and the actual error rate associated with such confidence reports […]"* (Baranski & Petrusic, 1994, p. 415). Their investigation however lacks a direct comparison with two subject groups, where for one the half-range method is employed and for the other the full-range method. Furthermore, Ronis & Yates (1987) had on the other hand found the full-range method to produce more accurate calibrations in earlier studies. The findings from Schoenherr & Petrusic (2015) attribute the "poorest calibration" to the full-range method. In this ambivalence and in the shortcoming of direct comparison tests among some scholar we find a research gap which will later be considered in the development of our research question and hypotheses. Prior to that, regarding the design of the traditional full-range method we shift our focus to the potential biases of anchoring and framing.

Anchoring

Anchoring is an effect that has gained awareness through the scientists who conducted the leading research on judgement heuristics and biases: Amos Tversky and Daniel Kahnemann (Tversky & Kahneman, 1974). The anchoring effect describes the influence of random numbers on human individuals' decisions. In the book "Thinking, fast and slow", where Kahnemann summarized decades of work on economic psychology, an example of the anchoring effect is described: Judges were tasked to roll two dices and read the pips before passing a sentence on a fictional case. Despite no existing

context of the number of pips on the dice and the juristic evaluation of the given case, judges with loaded dice (producing higher pips on average) significantly condemned the accused to more years in prison compared to judges with uninfluenced dice. A completely random number had altered an expert judgement.

For the half-range method Carroll & Petrusic (2007) suggested anchoring effects when limiting the range scale-range (f.e. to 55%-95%). We suspect the traditional full-range method to influence the confidence judgement of a respondent in two-alternative forced-choice tasks too. When a two-alternative task has a prefixed answer choice, then the 100%-end of the scale defines full confidence in that answer being correct, while 0% defines full confidence that the answer is wrong. In other words, at 0% the alternative is believed to be correct with full confidence. In the light of the anchoring effect, we suspect that confidence in the alternative from the fixed choice is enfeebled by indicating that confidence with "0%".

Framing

The second relevant cognitive bias is framing. The framing effect influences a judgement based on how a question's answer alternatives are presented (Kahneman, 2011, p. 87, p. 363; Plous, 1993). An example for the framing effect is the equivalent description of the consistence of a food product: "90% fat-free" or "10% fat". Costumers will normally prefer positive framings ("fat-free"). When the full-range method is applied in two-alternative forced-choice tasks, framing bias might come into effect. The indication of being sure of the alternative choice to be true is then given at "0%". This not only contradicts the linguistic logic but is likely prone to the framing effect. To treat this problem, a new alternation of the full-range design will be proposed and tested in the course of this thesis.

Duration

The relationship of response time and response accuracy is another typical subject in classical psychophysical comparison literature (Baranski & Petrusic,1995). The findings of Schoenherr et al. (2010) suggest that longer response times enhance response and confidence accuracy. This may be thought of in connection of the "slow" and the "fast" thinking processes, which are described by Kahneman (2011). The fast "System 1" is rather impulsive, emotional and unconscious while the slow "System 2" is conscious, calculating and more logical. This can have implications on the design of calibration judgement methods, when they rather invoke one or the other system.

Overconfidence and the Hard-Easy effect

Following Gigerenzer et al. (1991) and Baranski & Petrusic (1994) calibration research has put a major focus on the confidence-accuracy relationship and the "subjective probability assessment on questions of intellectual knowledge". Judgements are considered overconfident if the mean proportion confidence exceeds the mean proportion correct, and underconfident if the reverse is true (Liechtenstein & Fischhoff, 1977). It has consistently been found that individuals are typically overconfident when judgements are difficult and reversely underconfident when judgements are easy which is known under the term "difficulty effect" or "hard-easy effect" (Griffin & Tversky, 1992; Gigerenzer et al., 1991; Attali et al., 2020).

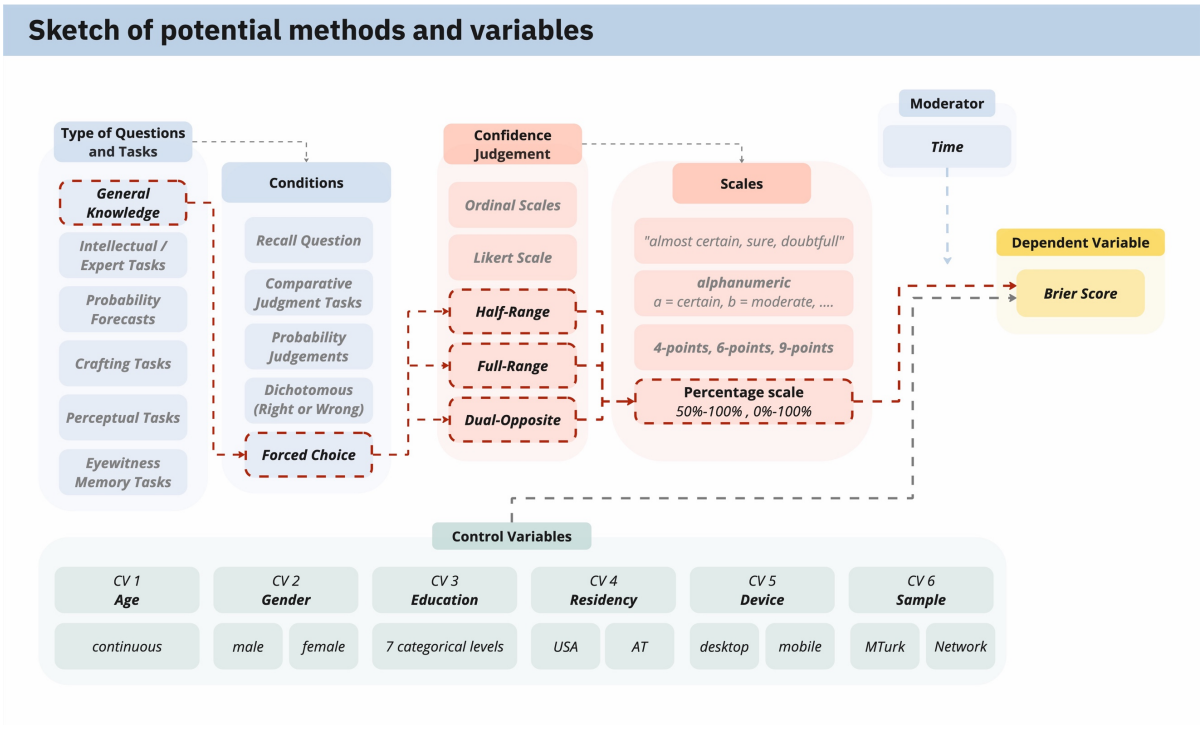
Due to the fact that the studies on overconfidence have produced consistent and very robust results – including study setups with two-alternative intellectual knowledge tasks – we will refrain from including an investigation into our research and focus instead on the Brier score levels dependence on the calibration judgements. We will however control for effects by level of education.

Overall, Baranski & Petrusic (1994) collected several studies promoting the fact that individuals tend to be overconfident about how much they know. In a 2013 study Hadwin & Webster found a significant

relationship between students' calibration accuracy and academic success. Interestingly, Hadwin and Webster (2013) further note that students provided more accurate confidence judgements for current tasks compared to such lying back in time. While this time component will not be able to integrate in our study, we will control for an effect of the level of education on response and calibration accuracy.

Figure 4 - Sketch of methods and variables

In the effort to explore the literature in order to identify and isolate a research gap this final model was sketched. It combines the reviewed methods for response accuracy elicitation and confidence judgements as well as the included moderators, control variables and the final DV Brier score. The chosen methods for this study are highlighted with the dash-dotted red line.



2.2 Research Gap

In a short special issue on the “calibration of calibration” the researcher Patricia A. Alexander (2013; p. 2) asks “How can researchers realistically and effectively measure such judgments and calculate calibration from the resulting data?”. Coming from a distinguished professor of educational psychology, strategic processing and quantitative methodology ("Dr. Patricia A. Alexander Named a Distinguished University Professor" | UMD College of Education, 2022) her question underlines the relevance of further research on improving methods for confidence calibration.

Based on the reviewed literature on confidence calibration and human judgement we find that confidence scales and scale sizes have only partially been investigated like Schoenherr & Petrusic (2015), Carroll & Petrusic (2007) and (Weber & Brewer, 2003) with a focus on the comparison of the half-range and full-range methods by scholars. However, results on the superiority of one or the other confidence judgement method have varied, suggesting further investigation.

As confidence judgements on ordinal scales and Likert scales are suboptimal for examining the response accuracy relationship (Cummings, 2006) we will employ a percentage scale. The strong correlation of the confidence-accuracy relationship with general knowledge tasks (Luna & Martín-Luengo, 2012) and the suitability of general knowledge questions for a two-alternative forced choice setup ensue their application in our study ideal.

Furthermore, we suspect the full-range method prone to the anchoring and framing bias (Carroll & Petrusic, 2007; Tversky & Kahneman, 1974; Kahneman, 2011; Plous, 1993). This issue will receive treatment through the suggestion of an improved judgement method, which will be called the dual-opposite method.

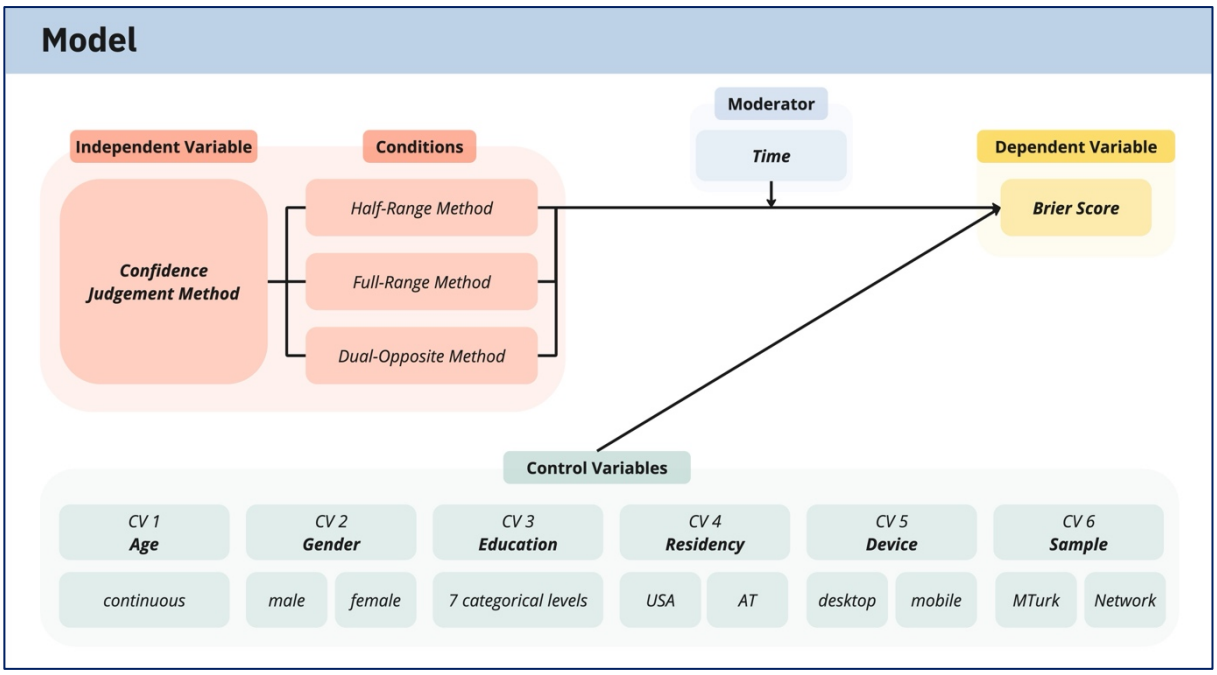
Based on this background we formulate the general research question: *“How do the confidence half-range, full-range and dual-opposite judgement methods influence the calibration accuracy of respondents in two-alternative forced-choice general knowledge tasks.”*

3. Research Design

3.1 Model and Hypotheses

Building on the reviewed research on confidence calibration the following model in Figure 5 was constructed for the research question with our independent variables “Confidence Judgement Method” and the moderating independent variable “Completion Time”, the dependent variable “Brier score” for the measurement of calibration accuracy and all included control variables.

Figure 5 - Model



With the aim to improve calibration accuracy compared to the full-range method through a new and improved calibration method on a discrete 0%-100% scale, which tries to reduce a possible anchoring and framing bias, we develop Hypothesis (H1a):

Hypothesis (H1a) *The dual-opposite method leads to smaller Brier scores than the full-range method.*

While scholars Ronis & Yates (1987), Baranski & Petrusic (1994), Weber & Brewer (2003) and Schoenherr & Petrusic (2015) had found contradicting conclusion on the superiority of one of the

two confidence judgement methods, we assume the status of the latest publication stating that the half-range method is superior and investigate if this can be assumed for our two-alternative forced-choice experiment. The according hypothesis is formulated as:

Hypothesis (H1b) The half-range method leads to smaller Brier scores than the full-range method.

In aspiration that the improvement of the new confidence judgement method would be substantive that it would constitute a better method in comparison to the half range method, we propose the third hypothesis.

Hypothesis (H1c) *The dual-opposite method leads to smaller Brier scores than the half-range method.*

In case, that all initial hypotheses are not refuted, the dual-opposite calibration judgement method would ascent as the superior method in comparison to the traditional methods. We express this in the fourth hypothesis .

Hypothesis (H1d) *The dual-opposite method leads to smaller Brier scores than both the traditional half-range and full-range methods.*

Finally, in accordance to the findings of Schoenherr et al. (2010) that longer response times improve calibration, we test if this hypothesis can be confirmed in our models as well. Therefore we propose the last hypothesis:

Hypothesis (H2) *Longer completion time leads to better calibrated answers, thus to smaller Brier scores in all method groups.*

3.2 Method

To investigate our research question, an experimental survey was conducted via Amazon Mechanical Turk, the survey tool Qualtrics and a link-distribution tool. Two samples were recruited, one from Amazon Mechanical Turk and a second sample was from a personal network of international students, friends and family. With the collected data an empirical analysis was undertaken. Details for the setup in each tool, the selection of participants, the instructions given as well as arisen pros and cons of the method and chosen tools will be discussed in the following chapter.

3.2.1 Data Collection

Three variations of the survey were setup in Qualtrics and distributed as Human Intelligence Tasks (HITs) via the online platform Amazon Mechanical Turk (MTurk) and via a private network with use of the weblink distribution tool *Nimble Links*.

Amazon Mechanical Turk

Amazon Mechanical Turk is an on-demand crowdsourcing online-platform that allows access to a "global, 24x7 workforce" (*Amazon Mechanical Turk*, n.d.). This workforce consists of voluntary participants, that are called workers or "Turkers" (Fort et al., 2011). Researchers, called requesters, can design Human Intelligence Tasks (HITs) directly within MTurk or embed tasks created with external tools like e.g. Qualtrics or LimeSurvey for workers to complete. These HITs are generally constructed as microtasks, sets of elementary tasks which can be completed within a short amount of time online. Requesters assign a payment for the completion of HITs. The amount of this payment can be designed to be paid out dependent on specified goals that have to be met, like e.g. completion time. The allotted amount of payment calls for caution for good practice of scientific ethical values. While it has been noted, that MTurk enables experiments in a faster and cheaper manner than laboratory experiments (Berinsky et al., 2012; Buhrmester et al., 2011), this can come at the cost of the workers through often unfair wages, as criticized by Fort et al. (2011). Possible discrimination of workers could in this case be even more exploited, as "Turkers" are reported to belong to the lower income population in the U.S

(Paolacci et al., 2010). However, this issue does not apply to all domains of usage of the platform, as researchers often pay equal to or above hourly minimum wage of the respective country (compare e.g. Keck & Tang, 2020). Yet this issue remains worth noting for awareness for fair use.

Fair pay raises further questions about how quality of data is affected by the motivation of earning money for the completion of HITs. As the main motivation of most workers has been reported to be additional income (Paolacci et al., 2010), their motivation would naturally be the completion of as many task in as little time as possible for maximum pay. There was however no effect of the amount of compensation on the quality of data found in the research of Buhrmester et al. (2011). The time it took until a HIT was selected and completed by workers was speed up by higher assigned rewards.

One further problem with online experiment setups with MTurk come by the lack of control over the activity of workers during the completion of tasks. Workers could e.g. be distracted, use the internet for answers or aided by neighbors sitting with them on a couch. In fact, Chandler et al. (2014) confirm that "Turkers" engage in HITs often as byline, like during watching television. Nevertheless, it could not be confirmed that workers were distracted.

Apart from concerns for data quality through money incentives and uncontrolled experiment situations, the demographic characteristics of MTurk samples have been investigated by several scholars. It has been shown, that MTurk samples represent the U.S. population better than college student samples (Berinsky et al., 2012), despite generally being younger and better educated than the general U.S.-American public (Paolacci et al., 2010). Further, Kees et al. (2017) found that MTurk samples outperformed data from panels of other professional marketing research companies in marketing research.

Summarized, the negligible effects of money incentives and uncontrolled experiment setups, as well as the superior MTurk samples over student samples qualify Amazon Mechanical Turk as the chosen data collection platform for this study.

Network Sample

Prior to the launch of the HIT on Amazon Mechanical Turk another sample was recruited within a private network of students, acquaintances, friends, family, work colleagues and colleagues from a professional network around the European Forum Alpbach. These participants were recruited through private messaging channels and partook without monetary compensation. This sample was recruited for several reasons: Prior of publishing on the paid platform of Amazon Mechanical Turk the private network sample allowed to test the prepared survey for understandability of the survey instruction and methods, as well as pretest the completion time in order to estimate the necessary fair payment for workers on Amazon Mechanical Turk. Furthermore, if the survey process was tested successfully, as was the case, the participants from the private network could be included into the study to increase the total sample size, broaden the demographic versatility of the sample and add the opportunity to test for country-level differences.

Qualtrics

Qualtrics was chosen after encountering difficulties in programming the appearance and data collection of the confidence judgements in Amazon Mechanical Turk other programs like LimeSurvey. The main issue was clarity for each confidence judgement method's instruction and application. Especially the dual-opposite method was met with barriers for the appearance and data collection, as it needs to add percentage points from 50 to 100 in both directions of the 50%-threshold in the middle. We found after excruciating trials with other programs, upon recommendation, that Qualtrics provided the best fitting solution for our research design.

Figure 8 - Test Setup in Qualtrics

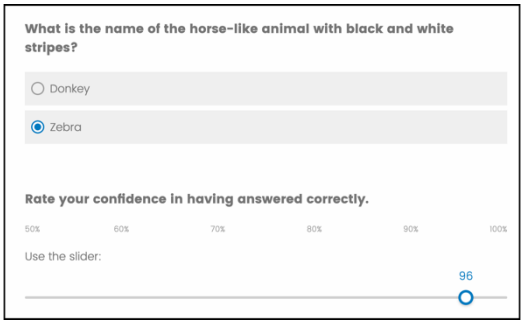


Figure 6 - Test Setup in Amazon Mechanical Turk

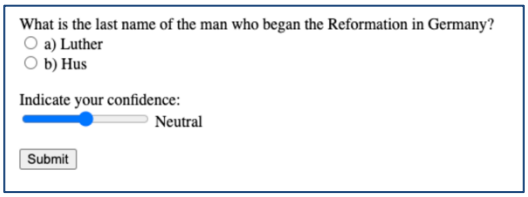
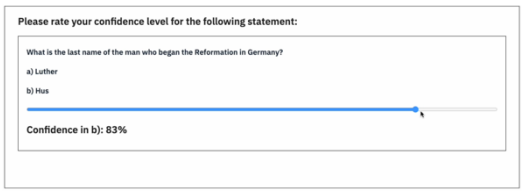


Figure 7 - Test Setup in LimeSurvey



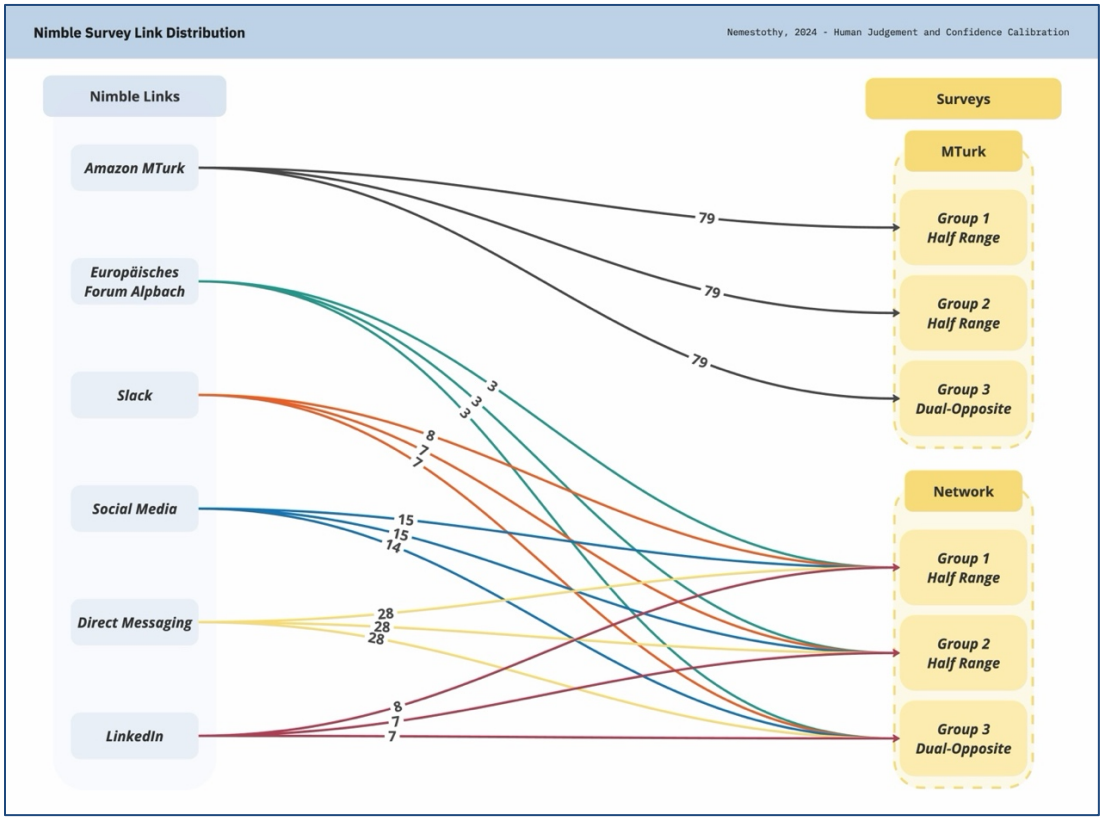
Nimble Links Inc.

Nimble Links Inc. (*Nimble Links: Links, Unleashed.*, n.d.) is a software-as-a-service provider for a weblink distribution tool which offers several options for redirecting links. Basically, links can be redirected evenly, randomly and by individual redirection formulas. Further, redirections can be set based on country, users' browser language settings, users' devices or operating systems or based on specific browsers. *Nimble* provides the operator basic web analytics tools for retraceability purposes, like number of referrers (pages from where users clicked on embedded links), destinations (the links to which users were redirected to), country of users and device used. It is important to note, that none of the tracked data is collected, stored or sold by the service provider (*Privacy - Nimble Links*, n.d.). *Nimble* offers a free and paid plan, where the latter was necessary for this study.

For this study six links were created in *Nimble*. Five links were created for the precedent circulation of the survey across the private network. This way all recruited participants from the private network were redirected and spread evenly into three Qualtrics survey groups respectively for each applied confidence judgement method. Figure 9 gives an overview of this distribution. The separate links allowed tracking the progress of recruitment across all different channels like e.g. direct messaging, Slack[2] and social media. Apart from that, this procedure initially allowed the use of the free version of *Nimble.*

The last link was embedded as the survey link in Amazon Mechanical Turk, such that all recruited participants were spread randomly and evenly among the three different survey groups and redirected to the respective Qualtrics-survey. This required the paid plan to surpass the number limit of clicks on the distribution link.

Figure 9 - Nimble Link Distribution (incl. number of redirected link clicks)



[2] Communication and messaging program commonly used at work in the digital work sector.

3.2.2 Procedure

One HIT was published on Amazon Mechanical Turk with a redirection tool to distribute the respondents randomly and evenly among three surveys. Another link was prior shared via private channels using the same distribution tool. Hence, participants who selected the HITs in MTurk just as via the privately shared link were redirected randomly to the respective questionnaires on Qualtrics. There they would find instructions (see Appendix, Figure 26; Figure 27 and Figure 28), which would explain the tasks that had to be completed: Answering general knowledge questions and giving an assessment on how confident they felt about having answered each question correctly. Note, that each of the three surveys applied a different method for the confidence judgement.

The given task instructions did not point towards the focus of this study, the difference of confidence calibration accuracy dependent on input method, to avoid experimenter demand effects (EDEs). EDEs are a bias triggered by respondents realizing the intent of a study and trying to answer in a way that helps confirm that intent (Mummolo, 2019).

Respondents were not incentivized for particularly fast completion nor was the survey programmed with time limits. This procedure took into account the response time difficulty effect (Baranski & Petrusic, 1994a). The response time difficulty describes the moderating effect of time-constraints in decision processes on confidence levels, thus specific decision times could be linked to specific confidence levels. The necessary average response time was estimated at 6 minutes and respondents were given no time restraints nor implications on effects of completion time, therefor we disregard the response time difficulty effect.

Before heading towards the core part of the survey respondents had to fill out demographic questions on their age, gender, education, residency and solve a reCAPTCHA to ensure them being humans instead of bots. The survey tracked the respondent's duration time and device used in the metadata in the background.

Next, in the core section of the survey, respondents had to answer 21 general knowledge questions and assess their confidence of being correct for each question. The chronology of the questions was randomized to reduce risks of question order effects.

While the content of the general questions remained identical, the method to provide a confidence judgment for each question varied, such that for each group one of the three methods were applied: the half-range method, the full-range method and the dual-opposite half-range method.

The following chapters will dive deeper into the topic of general knowledge questions and the confidence judgement methods.

3.2.3 General Knowledge Questions

Following the tradition of many studies on confidence calibration, the present study required participants to answer general knowledge questions.

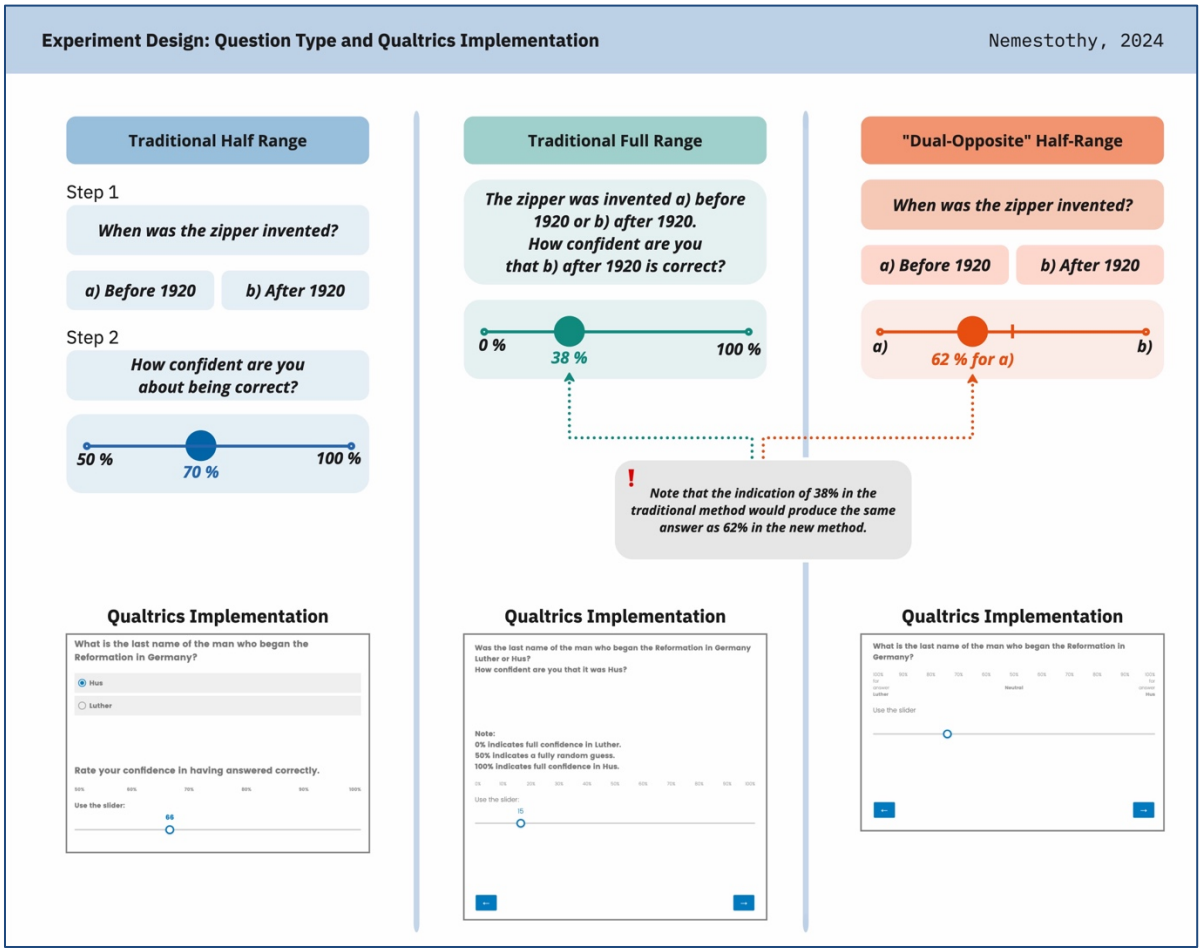
Concerning the curation of a representative set of general knowledge questions, internal and external validity face numerous challenges. Naturally, general knowledge is not static. Currently the majority of US citizens can be expected to identify Barack Obama as former president of the United States. The same cannot be asserted about Gerald Ford (in office from 1974 to 1977) or even less so about Andrew Jackson (in office from 1829 to 1837). Apart from timeliness and age constraints the generality of general knowledge questions is further bounded by territory (Buades-Sitjar et al., 2021; Buades-Sitjar et al., 2021; Coane & Umanath, 2021) and culture (Wertgen & Richter, 2023; Martín-Luengo et al., 2024; Whitcomb et al., 1995). One “normed battery of general knowledge questions” (Wertgen & Richter, 2023) stands out in the literature: Nelson and Narens (1980) compiled 300 general knowledge questions with the intent of timelessness of these questions. Regardless some questions are antiquated. Take for example “What is the capital of Czechoslovakia?” (Nelson & Narens, 1980) which has two answers since the Czech Republic (Prague) and Slovakia (Bratislava) have split in 1992. Further, the curation of general knowledge questions can evoke a gender gap. Scholars like Pallier (2003) and Exley & Kessler (2022) found an explicit gender gap in self-evaluation in context of male-typed tasks, which

did not occur in female-typed tasks despite similar performance in respective tasks. Male-type tasks were defined as such from STEM fields like math and science, whereas female-typed tasks where defined in the verbal communicative domain. In the case of Exley & Kessler (2022) the authors borrowed 20 questions from the Armed Services Vocational Aptitude Battery (ASVAB) from the subjects General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects to provide the possibility to compare the previous literature using ASVAB measure as well. For the female-typed tasks the authors chose a test for a verbal skills assessment which was described as stereotypically connoted to be more female (Exley & Kessler, 2022). Finally, the normative difficulty of general knowledge questions matter, as questions need to be distributed equally in levels of difficulty to allow coherence of judgement accuracy and calibration in the following experiment. Offering a big number of study participants only the questions, that have been normed "easy" will very likely produce mostly overconfident responses. An updated study on the questions of Nelson and Narens was conducted by Tauber et al. (2013) offering an updated succession of hard to easy general knowledge questions. "Easy" questions will be considered as those with a mean probability of recall of over 50 %, "hard" questions with below 50 % mean probability of recall. Summarizing the listed boundaries for an eligible set of general knowledge questions – which is certainly expandable – the curation of respective questions took into account the factors of timeliness, territory, culture, gender and variation in normative difficulty. The final set of 21 questions used for this thesis' experiment can be found in the appendix, Table 10.

3.2.4 Applied Confidence Judgments

The focus of this study is the comparison of the three confidence judgement methods: the half-range, full-range and dual-opposite method. These enable respondents to rate how confident they feel about being correct with the answer on the respective question. The traditional methods are the half-range method and the full-range method. The underlying work examines if an adapted method, the dual-opposite method, produces more accurate results. More accurate means better calibrated, hence a smaller Brier score would be expected for a better calibrated result. The scheme in Figure 10 gives an overview of the methods with examples of their implementation in the survey program.

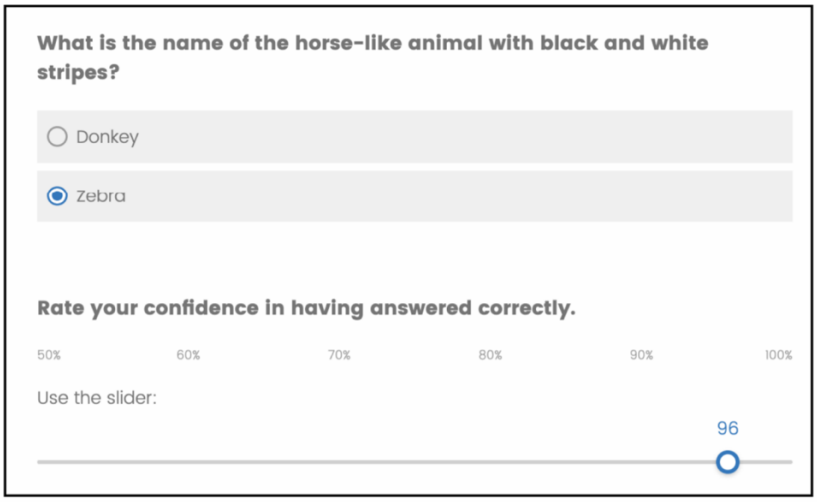
Figure 10 - Method Design for Qualtrics



Half-Range Method

The half-range method appears as the most common method for two-alternative questions. It traditionally produces a confidence calibration measure of respondents in two steps. In this study, participants first had to choose one of two alternative answer options for a general knowledge question. Next they had to give a confidence judgment on the "half-range" of a 100%-scale to indicate how confident, they felt about being correct on the general knowledge question. See Figure 11 below for the introductory example from the survey.

Figure 11 - Confidence Judgement - Half-Range Method



The exported data from Qualtrics would hence produce two data entries per question (see Table 3 - Half Range Survey Data Output Example below) from which the Brier Score was calculated by a formula in MS Excel for all questions. Therefor answer cells (Q1.1) were compared with the list of true answers and computed respectively into 0's (incorrect) and 1's (correct).

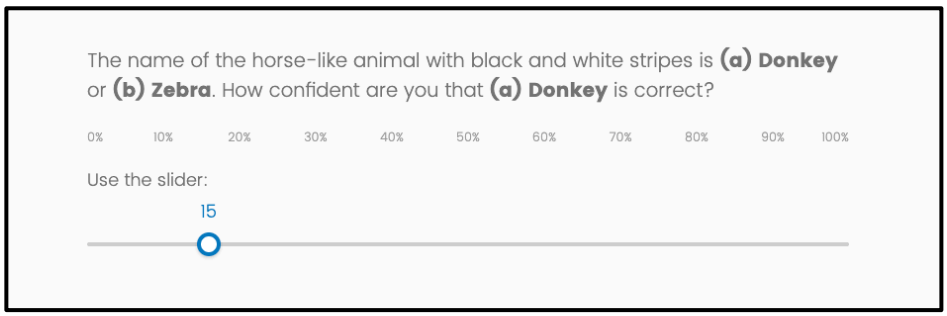
Table 3 - Half Range Survey Data Output Example

Q1.1	Q1.X (Computed)	Q1.2	Q1.BS (Computed)
General Knowledge Question	Correct State	Confidence Judgement	Brier Score
Zebra	1	96%	0,0016

Full-Range Method

The full-range method was established to improve the accuracy of the half-range method. It enabled respondents to correct their belief below the 50%-threshold and with the premise that the uncut scale could be easier to understand for respondents. Hence the full-range method deploys the entire 100%-scale for a predefined answer option. The indication of a value below 50% translates to the respondents believe that the alternative answer option is correct. In our study the answer on the general knowledge question and the confidence judgement could therefore be given in a single step, as seen in Figure 12.

Figure 12 - Confidence Judgement - Full-Range Method



The data output from the full-range method hence produced only one data entry per question. The Brier score was then computed by checking if the correct answer or the lure had been the predefined answer option. If the given answer option was the lure and the respondent indicated 15%, the response was computed as “correct” with a confidence level of 85%, therefor the Brier score would yield 0,0225, a well calibrated example (see Table 4).

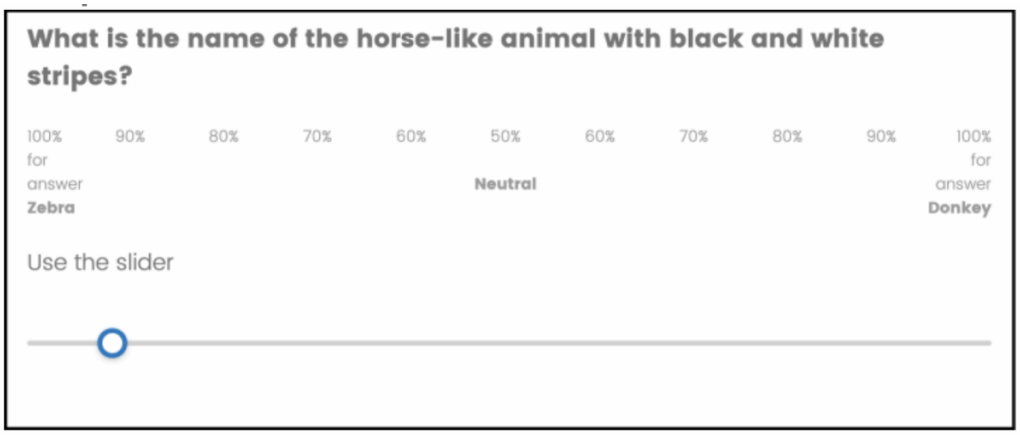
Table 4 - Full Range Survey Data Output Example

Q1	Q1.X (Computed)	Q1.2 (Computed)	Q1.BS (Computed)
Combined Gen. Knowl. Q. and Confidence Judgment	Correct State	Confidence Judgement	Brier Score
15%	1	85%	0,0225

Dual-Opposite Method

The dual-opposite method was established for this study in the endeavor to improve the comprehensibility and thus the accuracy of the full-range method. It combines the principles of both traditional methods by offering the clarity of choice of the half range method and yet deploying a 100%-scale with a threshold at 50%. On the left of the threshold the indicator would show the "positive" confidence judgement in the left answer alternative. In our study, respondents were hence able to give their general knowledge answer and confidence judgement in one step like in the full-range method, however with more clarity for the meaning of their indication. Regarding the scale in Figure 13 below, we see the indication of the believe that the answer option "Zebra" is correct with 85% confidence.

Figure 13 - Confidence Judgement - Dual-Opposite Method



The data output from this method is similar to the full-range method with a computed correct state, confidence judgement and Brier score, as the improvement in indication clarity was only possible to program in the frontend of the survey in Qualtrics.

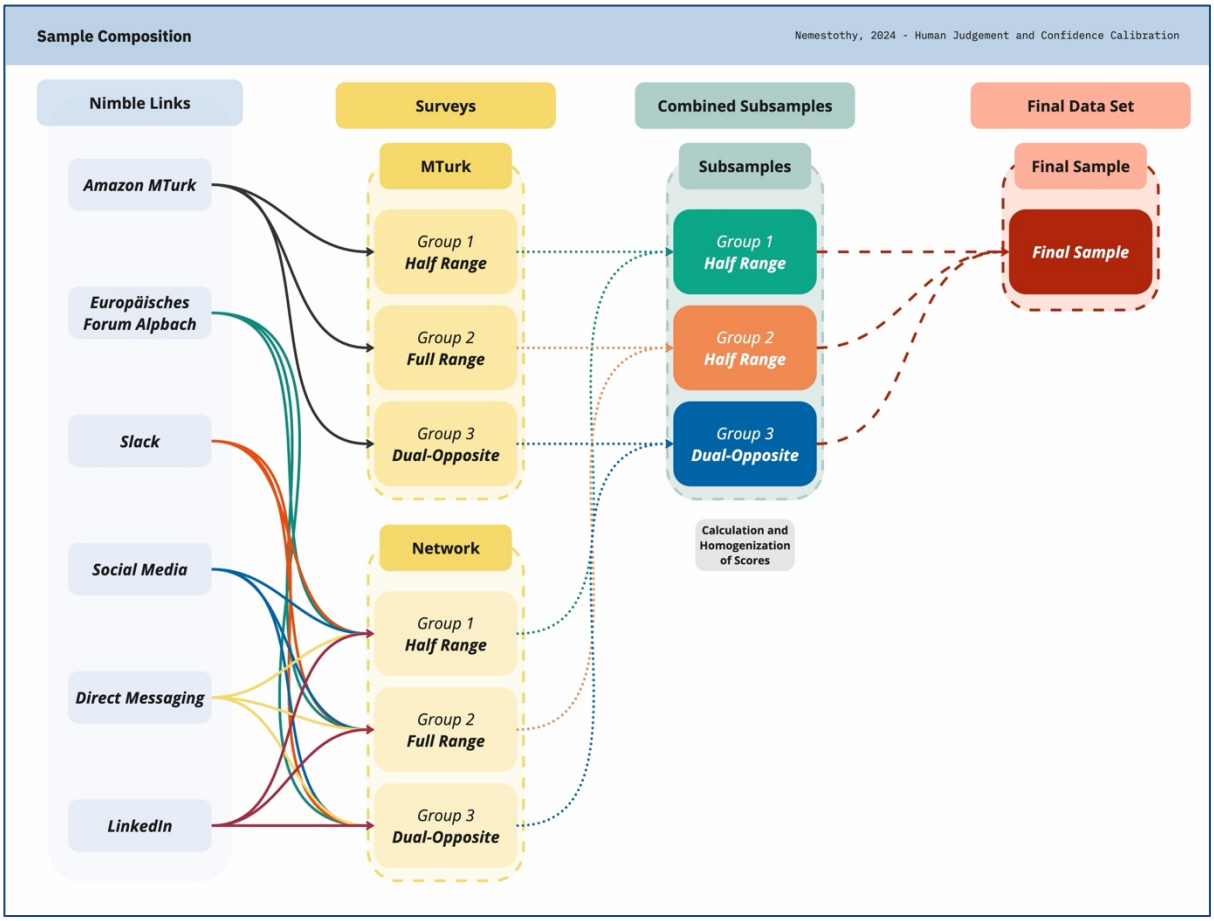
Table 5 - Dual-Opposite Survey Data Output Example

Q1	Q1.X (Computed)	Q1.2 (Computed)	Q1.BS (Computed)
Comb. Gen. Knowl. Q. and CJ	Correct State	Confidence Judgement	Brier Score
15%	1	85%	0,0225

3.2.5 Sample Composition

The redistribution of the separate surveys and the differences in each survey data output required multiple steps of treatment of the data in order to transform all results into one comparable data set. The six independent survey groups were appended into three groups according to the applied confidence judgement method in the survey. Within these combined samples results had to be transformed and computed to allow comparisons among all three groups. The computation of the relevant variables (correct state, confidence judgement and brier score) was already mentioned in the chapter above. The three computed subsamples were next appended into the final data set. Figure 14 provides an overview of the composition of the data, beginning with the redistribution of the links and leading through all junctions to the final appended data set.

Figure 14 - Sample Composition



3.2.6 Participants

In total 184 workers from MTurk chose the survey HIT on 26th August 2024 in a timeframe of 4 hours during which the survey was online. Prior, from 16th August 2024 to 24th August 2024 another sample counting 135 participants had been recruited from a private network of international students, friends and family. Notably this network sample was recruited on completely voluntary basis without monetary compensation or other incentives. All participants were randomly allocated to one of the three surveys through the paid link-distribution tool Nimble (see Figure 9 above).

Two participants from the MTurk sample submitted incorrect completion codes and where thus expelled and not paid, further two submitted successfully completed surveys yet did not enter the completion code in given restriction time and where thus not able to be compensated due to system limitations of Amazon Mechanical Turk. The remaining 180 Turkers were compensated with $ 0.73 for their participation after a check-up on the properness of their completion codes.

From the total 319 submissions of both samples 58 (22 MTurk / 36 Network) participants did not finish the given tasks correctly and were therefor deleted from the data set. The success status of 12 respondents showed an "unfinished" status, however upon inspection revealed that all data entries were successfully and correctly filled out, thus they were included. Despite the set requirement in Amazon Mechanical Turk to only permit

Table 6 - Demographic information

Demographic information	
	Number (percent)
N	232
Age	31.086
Gender	
male	157 (67.7%)
female	75 (32.3%)
Educational Level	
Prefer not to say	1 (0.4%)
Some high school or less	6 (2.6%)
High school diploma or GED	28 (12.1%)
Some college, but no degree	9 (3.9%)
Associates or technical degree	4 (1.7%)
Bachelor	108 (46.6%)
Graduate or professional	76 (32.8%)
Country of residence	
Austria	83 (35.8%)
United States of America	149 (64.2%)
Device	
desktop	150 (64.7%)
mobile	82 (35.3%)
Sample	
mTurk	150 (64.7%)
network	82 (35.3%)
Confidence Judgement Method	
HR	76 (32.8%)
FR	73 (31.5%)
DO	83 (35.8%)

US residents to enter the survey a handful of residents from other countries were permitted to partake by Amazon Mechanical Turk. 29 respondents from the MTurk- and private network sample were deleted based on residence to minimize the risk for location- or culture-based biases, to avoid interference of language barriers on the data quality and for the sake of parsimony in further analyses. Seven participants showing a completion time of more than three interquartile ranges from the 25th and 75 percentile were removed from the analysis, following Hoaglin & Iglewicz (1987)**.** 14 participants with a completion time below 180 seconds which raised the doubt in their conscientious survey completion. 180 seconds as a minimum completion time was decided based on two factors: Firstly, trials for fast completion could not be finished diligently in shorter time and secondly the fastest completion time of the network sample accounted 180 seconds. A detailed breakdown of the data treatment can be found in the appendix with Table 11 - Data treatment and cleaning.

After cleansing of the data, the number of successful completions of the study was accomplished by 233 participants. One further respondent was later disregarded for statistical analyses as explained under section Gender Ratio below. A summarizing table about the demographic information of the participants is found in Table 6.

Completion time

Expected completion time was evaluated with the precedent private network sample. Based on Chandler et al. (2014) who found Amazon Mechanical Turk respondents often working on HITs in distractive environments and the likeliness of respondents interrupting a survey task in order to finish it shortly later, the duration average value was expected to be right-skewed, hence looking at the median completion time is anticipated to give a more realistic value for the required time for the task. The median duration of the three private network groups were 315 seconds, 377 seconds and 360 seconds (see detailed reports in appendix: Figure 31, Figure 32 und Figure 33), thus the average median duration time of all three private network samples before data cleansing was 5min 51 seconds (350,67 seconds). The minimum wage in the U.S. is 7.25 $ per hour (U.S. Department of Labor), hence the median duration of the survey was expected to be require a minimum compensation of € 0,71. The final median duration of the treated data of the Amazon Mechanical Turk respondents was however 8min 50seconds (533 seconds) as seen in Figure 15, thus the paid compensation of $ 0,71 for the finished HIT came short by 36 cents to meet the level of minimum wage. This needs to be taken into account in future studies to improve duration estimation and guarantee fair pay.

Figure 16 - Boxplot: Total Completion Time

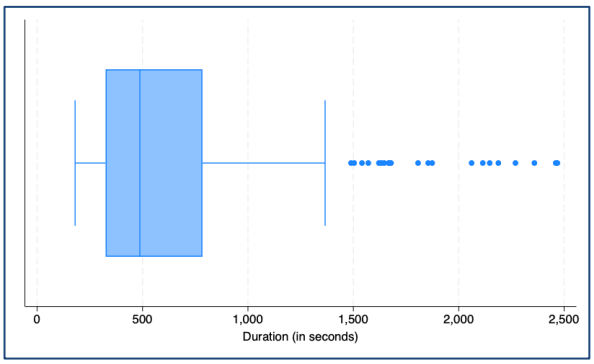
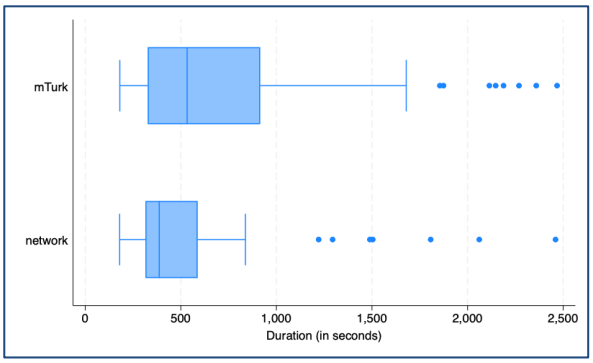


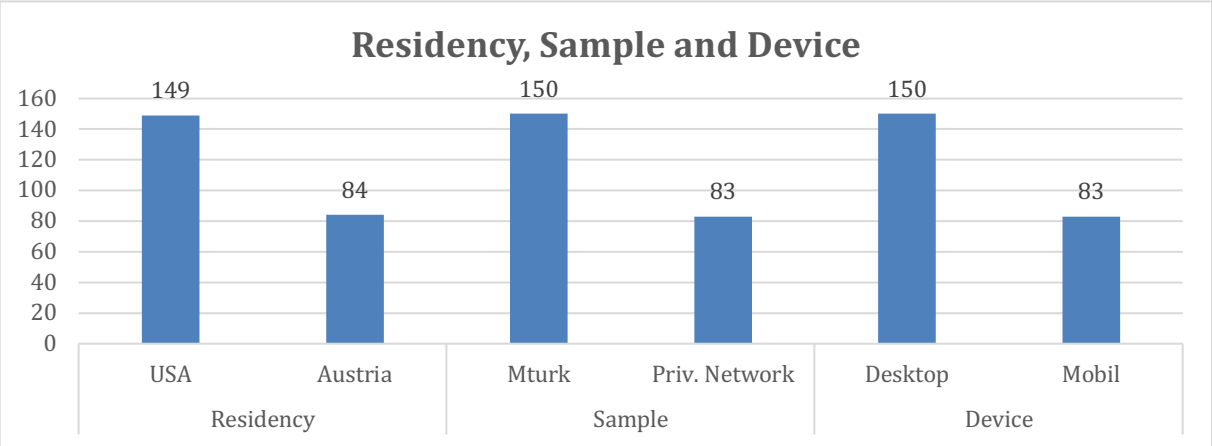
Figure 15 - Boxplot: Completion Time by Sample



Residency, Sample & Device

84 respondents from the final data set declared their residence to be Austria, while 149 were residents from the U.S.A. With the exception of one Austrian Turker, these numbers represent the ratio of the sample size from MTurk (150) and the private network (83). All respondents from the Amazon Mechanical Turk sample entered the survey in Qualtrics using a desktop device, while all respondents from the private network sample used a mobile device. This obviously entails total correlation of these variables and will require treatment in the statistical analysis. Further this circumstance limits our ability to draw conclusions from these characteristics of the respondents, thus limiting the possibility for interpretation.

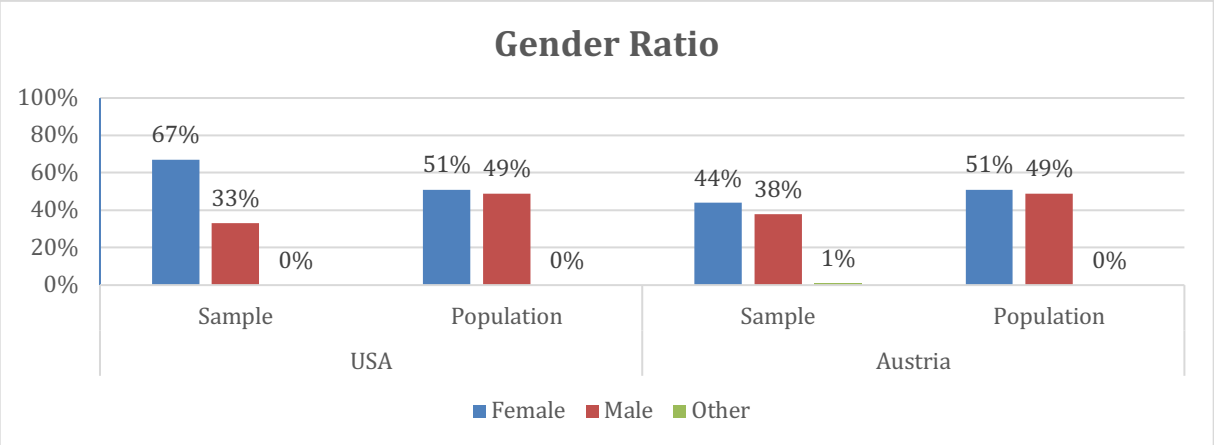
Figure 17 - Diagram: Residency, Sample and Device



Gender Ratio

There was a gender ratio of 67 % male participants with 75 workers identifying as female, 157 as male and one as other in the cumulative data set. This percentage of men is far higher than of the resident population of the U.S., where 51% are female against 49 % male in 2020 (US Census Bureau, 2020; Blakeslee et al., 2020). This contrast is even higher in the MTurk sample, which is composed exclusively of U.S. residents (after data cleansing). In the MTurk sample 75% of respondents are male and 25% are female. From the private network sample, composed of Austrian residents the gender ratio is closer to factual Austrian demographics yet also different: From the respondents 53 % were male, 46 % female and 1 % defined as other. By 01.01.2024 51% of Austrians were female, 49 % were male while other genders were not recorded (Statistik Austria, 2022.). It is worth noting, that the Austrian ratio equals that of the USA. However, the respective questions of the survey were designed to avoid gender bias and later data analysis showed no significant differences between the female and the male respondents results. Hence, the gender ratio poses a neglectable constraint to external validity of the underlying study. Note, that the data entry of the singular person who identified as other was later dropped to aide simplicity of statistical analyses.

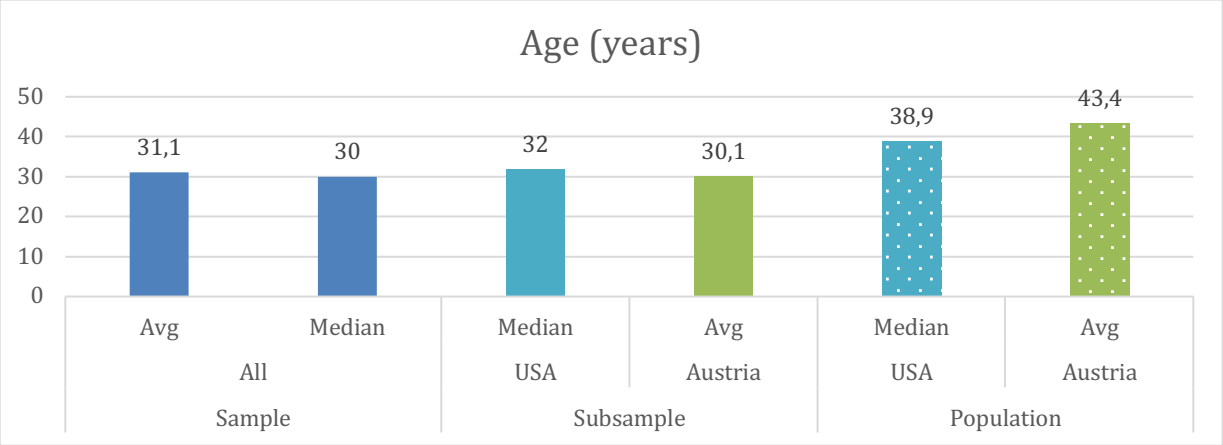
Figure 18 - Diagram: Gender Ratio



Age

The average age of the sample was 31.1 years, the median age was 30 years. The average age among U.S. residents in the sample was 31.7 years and the median age 32. The average age among Austrians in the sample was 30.1 years and the median age 29. This is far from the U.S. populations' median age of 38.9 years in 2022 (US Census Bureau, 2022.) and the Austrian average age of 43.4 years. The sample was hence younger than the population.

Figure 19 - Diagram: Age in years

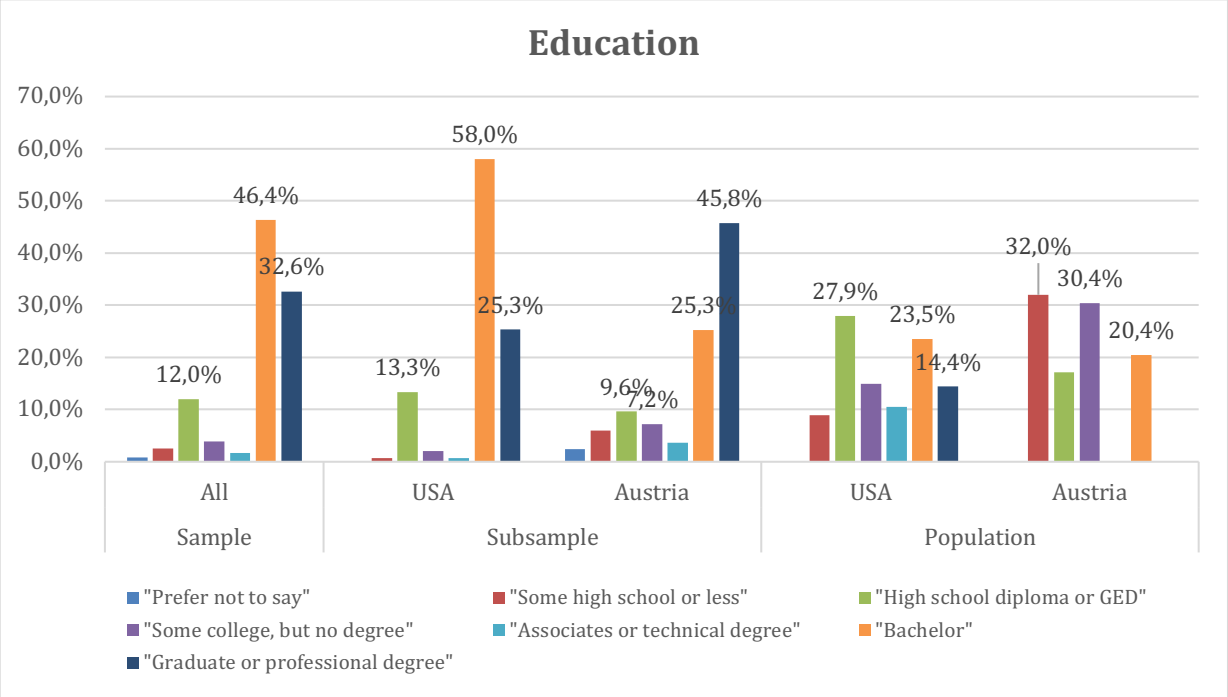


Education

Education was measured as a categorical variable with 7 possible answer options, where a majority of 79% of the respondents showed an academic education (Bachelor or Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.). The mode in the MTurk subsample was the entry “Bachelor” (87 respondents) and in the private network subsample “Graduate or professional degree” (38 respondents). This is far higher than the educational level of the general population of both the USA (37,9%) in 2021 (US Census Bureau, 2021.) and Austria (20,4% tertiary education) in 2022 (*Statistik Austria*, 2022). Thus, the sample shows a higher average educational level than the population.

The small number of respondents in categories lower than the “Bachelor” level and later found multicollinearity do not allow for proper statistical analyses. Therefore, the variable will later be excluded from statistical tests.

Figure 20 - Diagram: Education



4. Analysis

To get a feeling for the data before delving into statistical tests, three graphs were created. A scatter plot of the Brier scores by each method group over completion time (Figure 23) already shows several potential insights: While the data points are rather scattered vertically for the Brier score, they are right skewed with most entries on the left end of the x-axis for completion time. The fitted lines clearly show a difference in the completion time depending on the method group. The (blue) full-range method appears to have produced the shortest completion time, while the half-range and the dual-opposite method show rather similar completion times. This circumstance will require interacting the method and duration variables in further analyses.

The fitted prediction lines in Figure 23 indicate different average brier scores for each method group. This becomes apparent in Figure 22 and Figure 23, as we see a smaller brier score median and mean for the traditional half-range group.

These graphs suggest different results than anticipated in the proposed hypotheses and require further analyses.

Figure 23: Scatterplot - Brier Score over Completion Time

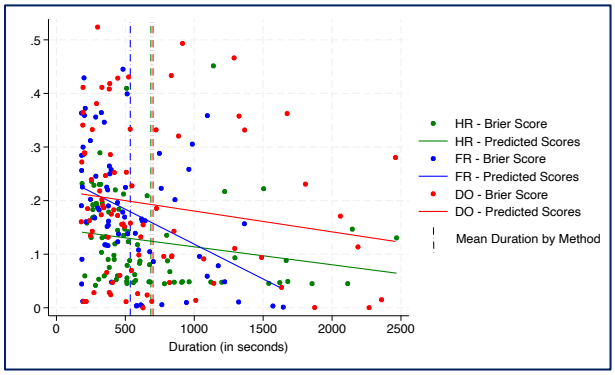


Figure 22: Boxplots - Brier Score by Confidence Judgement

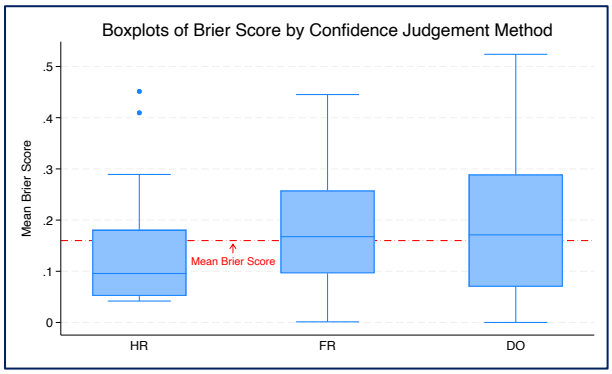
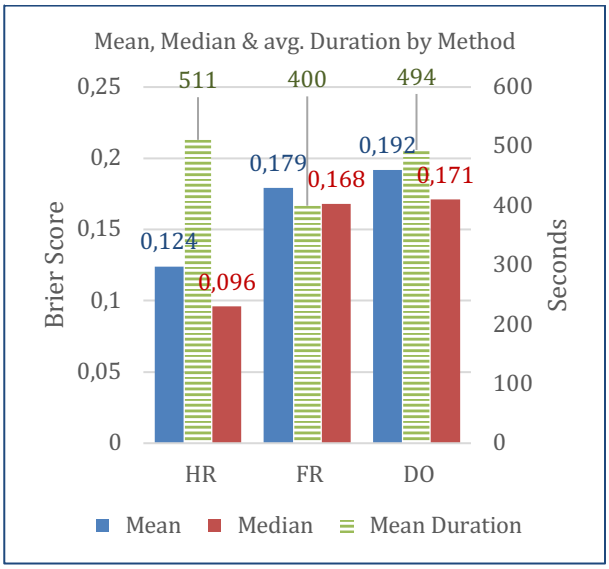


Figure 21: Bar Chart - Mean, Median & Duration



4.1 Tests for significance

Statistical analyses are conducted to support the assumptions drawn from the graphs.

Levene's tests for the homogeneity of variances (Appendix, Table 12) resulted in a p-value (p=0.00001486) smaller than 0.05 in all cases, thus the null hypothesis is rejected, meaning that variances are significantly different across all groups. This is a violation of the assumption of homogeneity over variances. Therefor a non-parametric test, like the Kruskal-Wallis, is suggested instead of an ANOVA test. While ANOVA tests are conducted to test for significant differences in the mean of a variable of multiple groups, the Kruskal-Wallis compares the difference in the median of multiple groups and is suggested in the case of not normal distributed data (Kruskal & Wallis, 1952). The p-value of the Kruskal-Wallis test (p=0.0045) is less than 0.05, hence we find a significant difference across the median values of the method groups (Table 7).

Table 7 - Kruskal-Wallis Test (Brier Score by Method)

Kruskal-Wallis Equality-of-populations rank test

method	Obs	Rank sum
HR	76	7278.000
FR	73	9187.000
DO	83	10563.000

chi2(2) = 10.807
Prob = 0.0045
chi2(2) with ties = 10.807
Prob = 0.0045

Table 8 - Dunn's Test (Brier Score by Method)

Dunn's Pairwise Comparison of brier by method (Bonferroni)

Col Mean – Row Mean	HR	FR
FR	-2.735	
	0.009	
DO	-2.956	-0.131
	0.005	1.000

With the post-hoc Dunn's Pairwise Comparison it can be specified, which groups have significantly different values. The half-range method shows significantly different values to the full-range (p=0.009) as well as the dual-opposite (p=0.005) method (Table 9).

While the Kruskal-Wallis test is suggested, in face of met criteria to assume the Central Limit Theorem applicable (>30 observations per group), ANOVA tests were conducted as well (see Table 21). Their results are in line with above findings.

Hence, we can conclude that there is a difference in the Brier Scores of the half-range group to the full-range-group as well as to the dual-opposite group.

4.2 Regression analyses

To produce further insights on the strength and direction of influence of the confidence judgement method on the dependent variable Brier score multiple linear regressions are conducted. Before testing the hypotheses, assumption for classical linear models (CLM) were checked for each model (listed as a table in Appendix, Table 20). All models are linear in parameters. Respondents have completed the experiment independently and were distributed equally. Control variables (age and gender) were included to avoid omitted variable bias and support the zero conditional mean. Multicollinearity was treated after testing, such that regressors which were perfect linear functions of another regressor were dropped from the model (see Variance Inflation Factor test in Appendix, Table 14). The final models show no multicollinearity. At this level, the models can be considered unbiased. Heteroskedasticity was however found with three of four models (see regressions and Bausch-Pagan-Tests in Appendix, Table 18), hence beta coefficients are assumed valid but suboptimal. Robust standard error regressions were applied for the respective models (Model 1, 3 and 4) to treat heteroskedasticity, as these do not rely on the assumption of homoskedasticity. The residuals off all models were not distributed normally, violating the last assumption for classical linear models. It must hence be concluded that confidence intervals and significance tests can be invalid while the β-coefficients (model estimates) of the models remain valid but not optimal.

While the models are thereby not robust and "BLUE" (Best Linear Unbiased Estimators), they remain unbiased. Therefor we continue with robust regression analyses, which do not rely on assumptions of homoskedasticity with the final models. Further, quantile regressions are conducted for all models in the 50%-quantile to support the consistency of the findings (see Appendix, Table 19).

Table 9 - Robust regressions

	VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
x_1	Full Range Method	**0.052*** **		**0.046*** **	**0.101*** **
		(0.017)		(0.016)	(0.027)
x_2	Dual Opposite Method	**0.064*** **		**0.066*** **	**0.069** **
		(0.019)		(0.019)	(0.027)
x_3	Duration (in seconds)		**-0.000*** **	**-0.000*** **	**-0.000***
			(0.000)	(0.000)	(0.000)
x_4	HR * Duration				0.000
					(0.000)
x_5	FR * Duration				**-0.000*** **
					(0.000)
x_6	DO * Duration				-0.000
					(0.000)
x_7	Age	0.001	0.001	0.001	0.001
		(0.001)	(0.001)	(0.001)	(0.001)
x_8	Gender = Female	0.002	0.008	0.004	0.008
		(0.016)	(0.016)	(0.016)	(0.016)
x_9	Sample = Network	**0.032** **	**0.028***	0.022	0.021
		(0.015)	(0.015)	(0.016)	(0.015)
β_0	Constant	**0.077***	**0.162*** **	**0.126*** **	**0.109** **
		(0.042)	(0.048)	(0.045)	(0.045)
	Observations	232	232	232	232
	R-squared	0.079	0.061	0.113	0.134

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

All models are successfully tested for significance, Model 1 ($F(5, 226) = 7.37$; $p < 0{,}01$; $R^2=0.079$), Model 2 ($F(4, 227) = 6.61$; $p < 0{,}01$; $R^2=0.061$), Model 3 ($F(8, 223) = 6.61$; $p < 0{,}01$; $R^2=0.134$) and Model 4 ($F(8, 223) = 6.61$; $p < 0{,}01$; $R^2=0.134$). In regard of the explained variability of the dependent variable through the independent variables Model 4 shows the highest $R^2 = 0.134$. To provide a different view the regressions respective regression equations can be found in Appendix (p. 79).

In Hypothesis (H1a) it was predicted that participants of the dual-opposite method group would produce smaller Brier scores compared to participants in the full-range method group. Tests on significance of the difference in the two groups did however not confirm a significant difference (Dunn's Pairwise Comparison test in Table 8; z-Score=-0.131, p-value = 1.00). In contrast with the hypothesis, the

regression shows an even stronger negative effect of the dual-opposite method on the confidence calibration level than the full-range method, thereby increasing the Brier score (Model 1 in Table 9; β=0.064, p=0.001; $CI_{95\%}$[0.0270, 0.100]). Hence, Hypothesis (1a) is refuted.

In Hypothesis (H1b) we proposed that participants of the half-range method group would produce smaller Brier scores compared to participants in the full-range method group. The regression model confirms the proposed effect in our model. Hence, being in the full-range method group would increase the Brier score (Model 1 in Table 9; β=0.052, p=0.002; $CI_{95\%}$[0.014, 0.090]). Accordingly, Hypothesis (1b) is validated.

Hypothesis (H1c) suggested the dual-opposite method would lead to smaller Brier scores than the half-range method. Consistent to above results, regarding the beta-coefficients of Model 1 in Table 9 the inverse holds true. The dual-opposite method (β=0.064) affected the brier score stronger than the half-range method (β=0.052). The Kruskal-Wallis and Dunn's Test did however not confirm the difference to be significant. Hence, Hypothesis (1c) is refuted.

Hypothesis (H1d) assumed that the novel dual-opposite method would outperform the traditional methods. This assumption was expected to be satisfied, provided that the first three hypotheses are validated. However, as Hypothesis (1a) and (1c) are refuted, we must refute Hypothesis (H1d) too.

Additionally, the control variable "Sample" emerged as a significantly influential (Model 1 in Table 9; β=0.032, p=0.031; $CI_{95\%}$[0.003, 0.061]) suggesting that participants from the Network sample have higher Brier Scores, thus are less accurate in their confidence calibration. This effect was only consistent through Model 1 and Model 2 but did not show significance in Model 3 and Model 4. Further, due to the correlation of the variables "Sample", "Residence" and "Device" and the circumstance of them having been dropped from the model, it cannot be determined which of these conditions causes the effect as it could be attributed to each of the variables.

The influence of other control variables, such as age and gender, was not statistically significant across all models.

The intercept β_0 appears significant across models, however on different significance levels. The constant indicates the baseline value of the Brier score, when all other variables are set to zero or their reference category. In context of significant variables like duration, assuming a value of zero is not meaningful, thus we only conclude that the intercept is significantly different from zero (Model 4 in Table 9; β_0=0.109, p=0.017; $CI_{95\%}$[-0.197, 0.197]).

When we turn to our Hypothesis (H2) we predicted that longer completion times lead to better calibrated answers, thus smaller Brier scores in all method groups. Again, we find significant effects of completion time on the brier score throughout all models and contrary to the previous predictions, this hypothesis is supported as expected. The longer participants took to complete the survey, the better calibrated their results were. However, while duration has an decreasing effect on the Brier score across all models and seems to not appear by chance, the marginal effect is so small that it is negligible (Model 2 in Table 9; β= -0.0000467, p=0.003; $CI_{95\%}$[0.00, 0.00] ; Model 3 in Table 9; β= -.0000469, p=0.003; $CI_{95\%}$[0.00, 0.00] ; Model 4 in Table 9; β= -.0000281, p=0.003; $CI_{95\%}$[0.00, 0.00]). Hence, Hypothesis (H2) is confirmed, however with a neglectable small effect.

Obvious differences in the mean duration time for each method group were shown earlier in graphs (Figure 23 and Figure 22). This raises reasonable suspicion, that duration time has an even stronger effect in the full-range method group (steeper slope in the Figure 23) compared to its' effect on the other groups (flatter slopes for HR and DO in Figure 23). To test this assumption the variables "method" and "time" were interacted in Model 4. Indeed, the effect was only significant in the interaction of the full-range method and duration (Model 4 in Table 9; β= -0.0000988, p=0.017; $CI_{95\%}$[-0.197, 0.197]). While the beta-coefficient can be described as more than two times stronger in the interaction model than in the previous models, yet the effect remains so small that again it is neglectable.

5. Discussion

Discussion of Hypothesis (H1a)

The dual-opposite method was introduced with the intent to improve the calibration accuracy of respondents using this method instead of the traditional half-range and full-range methods. In particular, it was intended to improve calibration accuracy compared to the full-range method, as proposed in Hypothesis (H1a). We suspected the nature of design of the full-range method to hold probable weaknesses and susceptibility to the anchoring and framing bias. The anchoring bias was expected to be the main driver for our assumption, as respondents were expected to struggle with the concept of indicating full confidence in the alternative answer-option at the 0%-end of a scale. In terms of framing the interchangeable statements in both methods are:

> Full-Range: *"I am 0% sure that option a) is correct. (Hence alternative option a) is 100% correct.)"*
>
> Dual-Opposite: *"I am 100% sure that alternative option b) is correct."*

We expected respondents to be able to overcome a potential anchoring and framing bias in the second option. The results from our analysis however did not confirm our assumptions and there was no significant difference found between the results of the traditional full-range method and the dual-opposite method. Hence, the dual-opposite method did not improve the Brier score compared to the full-range method.

Several suggestions for the lack of an improvement can be attempted. In this regard, attention has to be turned to the duration variable. The duration variable had an independent influence on the dependent variable Brier score and a moderating effect on the categories of the variable confidence judgement method. Though the effect was very small, we found that respondents in the full-range method completed the survey generally faster and the positive interaction effect of shorter completion time on the confidence calibration level was only significant in this group. Thus, it could be assumed, that

despite our assumptions the comprehensibility of the full-range method was better and hence might have led to both faster completion time and more accurate responses.

Regardless of the time variable and in consideration of no significant difference between the dual-opposite and half-range group it could be assumed that participants may be biased due to the concept of a 100%-scale to assess their confidence judgement.

Concluding, the novel dual-opposite method could not be validated to improve calibration accuracy compared to the traditional full-range method.

Discussion of Hypothesis (H1b)

Hypothesis (H1b) assumed that the half-range method would produce better calibrated results than the full-range method. We based our assumption for this hypothesis on the latest statement referring to this in the literature review by Schoenherr & Petrusic (2015), despite disparate findings by other scholars (Baranski & Petrusic, 1994b; Ronis & Yates, 1987; Weber & Brewer, 2003).

The analyses validate the hypothesis, as the results from half-range method showed the best calibrated results across all groups.

A potential reason aligns with the assumptions of anchoring and framing effects in prior hypothesis. It was expected that the quality of the confidence calibration would improve through the simplicity of the survey flow of the half-range, which seems easier instructed and less prone to a anchoring and framing bias. Respondents could easily separate their initial answer and after focus on the judging their confidence. The separated general knowledge question and confidence judgement task could possibly also inspire clearer focus on the distinction of the tasks. Confidence judgements were set up as a self-contained task in the half-range group, this may have supported the focus of respondents on this part of each survey question. Consequently, it could be argued that separating the general knowledge questions' judgement and the confidence judgement instead of combining them in a singular input method like in

the full-range and dual-opposite method may have a beneficial effect on the confidence calibration level.

Concluding, this result adds to the literature another example, where the half-range method was confirmed to produce superior calibration accuracy compared to the full-range method.

Discussion of Hypothesis (H1c)

Turning our focus to Hypothesis (H1c), the proposition was made that the dual-opposite method would yield smaller Brier scores than the half-range method. The results could not confirm this and the analyses showed the best confidence calibration levels for the half-range method, instead of the dual-opposite method.

The explanation of the difference through special focus on the confidence judgement as a discrete self-contained task remains probable. This is a consistent difference of both the dual-opposite and the full-range method towards the half-range method.

Discussion of Hypothesis (H1d)

Since the previous hypotheses have not all been validated, Hypothesis (H1d) is refuted. This hypothesis assumed that the novel confidence judgement method would produce superior calibration results compared to both traditional methods. Since it was found that it did not differ significantly from the usual full-range method and was inferior to the half-range method, the dual-opposite method can not be recommended for future research. This recommendation is further encouraged in light of the relatively complex setup and instruction of the dual-opposite method. Despite a lot of effort to set the new method up in standard survey tool, many programs could not support the method design. In contrast however, the half-range and the full-range method are simple to arrange in all survey tools.

Discussion of Hypothesis (H2)

With Hypothesis (H2) we expected longer survey completion time to improve the confidence calibration level of respondents. Following Kahneman's (2011) concepts of a fast “system 1” thought process and a slow “system 2” thought process and the findings of Schoenherr et al. (2010), we imagined respondents, who took more time to reflect on their answer and their confidence of being correct to achieve better calibration. In fact, we found a significant positive effect of duration on the Brier score, however with such a marginal influence that we can hardly interpret the hypothesis as confirmed by the data.

6. Conclusion

The literature on confidence calibration is rich and the application of the traditional confidence judgement methods is correspondingly frequent, while findings on the superiority of confidence judgement-methods have shown inconsistencies. Alexander (2013) stressed the importance to further “calibrate calibration”. This work tends to Alexanders’ request and contributes significant results supporting the thesis that the half-range method produces superior calibration accuracy in two-alternative forced-choice study setups.

In addition, the motivation for this thesis was to attempt an improvement of the traditional confidence judgement methods, by introducing a novel improved method, the dual-opposite method. This endeavor was not successful, and no significant improvement compared to the full-range method was found and what-is-more the half-range method excelled the dual-opposite method significantly.

6.1 Practical implications

Situations where people must decide for the right choice between two alternatives are pervasive in everyday life. This holds true for managers, for politicians, investors, physicians, teachers, students,

criminal investigators just as forecasters, just to name a few. Hence practical implications of the general topic of confidence calibration can be drawn for any domain, where decisions are made. However educated a decision is, decision makers in any domain often rely on their confidence to make the right decision (Baranski & Petrusic, 1994a). Sometimes the correct alternative is obvious and can be chosen with full confidence, sometimes it is very difficult to rate one's chance to be correct. The importance of understanding the metacognitive processes of decision making and confidence judgments has been emphasized by scholars (Baranski & Petrusic, 1995; Lichtenstein et al., 1982). The underlying thesis work adds to the understanding how scholars can measure these processes with particular focus on confidence judgement methods.

The distinctive contribution of this study to the field is its examination of a novel variation in confidence judgment methods: the dual-opposite method. While the dual-opposite method did not enhance respondents' calibration accuracy in a two-alternative forced-choice task, it reinforces existing practices and offers a research framework that can be utilized in future studies and comparisons of confidence judgment methods. Summing up, the results from this study recommend the use of the half-range method in two-alternative forced-choice surveys.

Furthermore, the preparation, participant recruitment and conduction of the survey experiment was followed by the setup of an automated response system for the private network respondents. Based on the individual participants responses the response tool automated an e-mail dispatch providing the respondents with their calculated confidence calibration scores, ratio of correct answers and comparisons to the general average. Such tools may possibly be applied as well in business practice. For example, this process can be useful in the development of a concordant strategic decision among a group of managers. With a previous round query of confidence judgements on a number of strategic questions a common baseline as well as individual standpoints could help gain a better assessment of an imminent strategic decision.

6.2 Limitations and future directions

This study faces some statistical and structural limitations. The recruited samples and participant demographics constituted several limitations.

External validity and generalizability beyond the sample is impaired due to the demographic difference between the samples and the populations. The samples were in total younger, “more male” and better educated than the Austrian and US-American population.

Statistical tests showed that there is a significant difference in the mean brier scores between the MTurk sample and the private network sample (see Appendix, Table 22, Table 23). Due to the near perfect correlation between the variables sample, device and residence (see Appendix, Table 24) the same effect on the brier score can be assumed by all of these variables. A causality can therefore however not be explained, despite many options for interpretation being possible: Despite great care in the curation of the general knowledge question to cover subjects known in all western cultures, unidentified cultural or residential differences could still take effect.

While Buhrmester et al. (2011) found that the level of monetary compensation did not influence the results of the quality of data from MTurk samples, this could still be a possible source for the differences in our sample and in the particular case of confidence judgements, which would need further research.

While a positive effect of the level of education could be assumed due to the large share of correctly answered questions, this assumption was not examined and would require further studies. The MTurk sample had a 11 percentage points higher share of academics than the private network sample with 71% academics (see Figure 20), which supports the assumption that higher educated respondents led to better calibrated Brier scores. Notably, the private network sample had a bigger share of respondents with a higher level than bachelor’s degree.

Simultaneously, the effect of the difficulty of the questions on the Brier score among different methods could be studied in the future. The curation of the survey questions for this study potentially produced

a drawback for this study. Even though the questions were selected from previously studied and standardized general knowledge questions catalogues, they were selected by the criteria that "easy" questions had been answered correctly by a minimum of 50% of participants from the question database and hard questions with by less than 50%. This approach neglected the risk of the "overeducated" sample group, which might thus know the correct answer to more than the half of the questions with fairly high certainty. Further, we assumed that confidence judgements would still vary enough to produce significant findings, when there was a higher percentage of probably knowable questions. Therefor two thirds of the questions were picked as rather easy and one third as rather hard. The results however showed that the majority of the questions was answered correctly and with high confidence, thus skewing the data for statistical analysis. Concluding, future studies are recommended to adopt a higher number of difficult and very difficult questions in order to receive a better diversified data set and probably more normally distributed data. Moreover, further research could take specific interest in the effect of the difficulty of general knowledge questions on new confidence judgement methods like the dual-opposite method.

Another improvement of the study could have been made by including an initial assessment of the confidence in their profoundness of general knowledge by participants prior to answering all questions. Added with a retrospective assessment of how many questions participants thought to have answered correctly, this could allow for insights in the effects of over- or underconfidence on confidence calibration. A hypothesis such as "Higher overconfidence leads to a higher Brier score than a neutral confidence level." could be proposed.

Further future research possibility is a comparison of the chronology of the (general knowledge) question response and confidence judgement within the half-range method.

To test the anchoring bias solely for the full-range method, a possibility would have been to conduct a survey with two respondent groups where one group would receive questions with the correct answer option always on the right end of the judgement scale at 100% and the other respondent group would receive questionnaires with the lure always on the right end of the judgement scale at 100%. The

question would hence go into the direction of "can respondents resist the anchoring bias and choose the right answer at 0% of a 100%-judgement scale?"

If the many possible future research opportunities describe one common thing, it is the complexity of the human decision process and complex possibilities for its research. In that sense, this work contributed one little piece to the big puzzle of the human mind and judgement.

"There are two kinds of forecasters: those who don't know, and those who don't know they don't know" (Bovi, 2009). This thesis contributes to recognizing the latter better.

7. References

Adams, J. K. (1957). A Confidence Scale Defined in Terms of Expected Percentages. *The American Journal of Psychology*, *70*(3), 432. https://doi.org/10.2307/1419580

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review*, *68*(1), 33–45. https://doi.org/10.1037/h0040274

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge Calibration: What Consumers Know and What They Think They Know. *Journal of Consumer Research*, *27*(2), 123–156. https://doi.org/10.1086/314317

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, *24*, 1–3. https://doi.org/10.1016/j.learninstruc.2012.10.003

Amazon Mechanical Turk. (n.d.). Retrieved 3 August 2024, from https://www.mturk.com/

Attali, Y., Budescu, D., & Arieli-Attali, M. (2020). An item response approach to calibration of confidence judgments. *Decision*, *7*(1), 1–19. https://doi.org/10.1037/dec0000111

Baranski, J. V., & Petrusic, W. M. (1994a). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. https://doi.org/10.3758/BF03205299

Baranski, J. V., & Petrusic, W. M. (1994b). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. https://doi.org/10.3758/BF03205299

Baranski, J. V., & Petrusic, W. M. (1995). On the Calibration of Knowledge and Perception. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *49*, 397–407. https://doi.org/10.1037/1196-1961.49.3.397

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.

Blakeslee, L., Caplan, Z., Meyer, J. A., Rabe, M. A., & Roberts, A. W. (2020). *Age and Sex Composition: 2020*.

Bovi, M. (2009). Economic versus psychological forecasting. Evidence from consumer confidence surveys. *Journal of Economic Psychology*, *30*(4), 563–574. https://doi.org/10.1016/j.joep.2009.04.001

Brier, G. W. (1950). VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brown, T. A., & Shuford, E. H. (1973). *Quantifying Uncertainty Into Numerical Probabilities for the Reporting of Intelligence: Vol. Santa Monica: The RAND Corp.*

Buades-Sitjar, F., Boada, R., Guasch, M., Ferré, P., Hinojosa, J. A., Brysbaert, M., & Duñabeitia, J. A. (2021). The thousand-question Spanish general knowledge database. *Psicológica Journal*, *42*(1), 109–119. https://doi.org/10.2478/psicolj-2021-0006

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Carroll, S. R., & Petrusic, W. M. (2007). Numerical Distance and Anchoring Effects in Judging confidence. *Proceedings of Fechner Day*, *23*(1).

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods (Online)*, *46*(1), 112–130.

Coane, J. H., Cipollini, J., Beaulieu, C., Song, J., & Umanath, S. (2023). The influence of general knowledge test performance on self-ratings of and perceived relationships between intelligence, knowledge, and memory. *Scientific Reports*, *13*(1), 15723. https://doi.org/10.1038/s41598-023-42205-y

Coane, J. H., & Umanath, S. (2021). A database of general knowledge question performance in older adults. *Behavior Research Methods*, *53*(1), 415–429. https://doi.org/10.3758/s13428-020-01493-2

Cooke, E. (1906). FORECASTS AND VERIFICATIONS IN WESTERN AUSTRALIA. *Monthly Weather Review*, *34*(1), 23–24. https://doi.org/10.1175/1520-0493(1906)34<23:FAVIWA>2.0.CO;2

Cummings, A. (2006). *The Use of Item Response Theory to Assess Adults' Postdiction Accuracy* [Georgia State University]. https://doi.org/10.57709/1061127

Dr. Patricia A. Alexander Named a Distinguished University Professor | UMD College of Education. (2022, May 10). https://education.umd.edu/news/08-26-19-dr-patricia-alexander-named-distinguished-university-professor

Exley, C. L., & Kessler, J. B. (2022). The Gender Gap in Self-Promotion*. *The Quarterly Journal of Economics*, *137*(3), 1345–1381. https://doi.org/10.1093/qje/qjac003

Ferro, C. a. T., & Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, *138*(668), 1954–1960. https://doi.org/10.1002/qj.1924

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, *37*(2), 413–420. https://doi.org/10.1162/COLI_a_00057

Galbraith, J. K. (1993, January 22). *Wall Street Journal*.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Griffin, D., & Brenner, L. (2004). Perspectives on Probability Judgment Calibration. In *Blackwell Handbook of Judgment and Decision Making* (pp. 177–199). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470752937.ch9

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-R

Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction*, *24*, 37–47. https://doi.org/10.1016/j.learninstruc.2012.10.001

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-Tuning Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association*, *82*(400), 1147–1149. https://doi.org/10.1080/01621459.1987.10478551

Huang, L., Zhao, J., Zhu, B., Chen, H., & Broucke, S. V. (2020). An Experimental Investigation of

Calibration Techniques for Imbalanced Data. *IEEE Access*, *8*, 127343–127352. https://doi.org/10.1109/ACCESS.2020.3008150

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed). Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1977). *Intuitive Prediction: Biases and Corrective Procedures*. https://apps.dtic.mil/sti/citations/ADA047747

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123–141. https://doi.org/10.1016/0010-0277(82)90022-1

Kawaguchi, J. (1992). Measurement of the feeling of knowing: Norms of 216 general-information questions. *Shinrigaku Kenkyū*, *63*(3), 209–213.

Keck, S., & Tang, W. (2020). Enhancing the Wisdom of the Crowd With Cognitive-Process Diversity: The Benefits of Aggregating Intuitive and Analytical Judgments. *Psychological Science*, *31*(10), 1272–1282. https://doi.org/10.1177/0956797620941840

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In A. Tversky, D. Kahneman, & P. Slovic (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 306–334). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.023

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., & Hertwig, R. (2019). *How the "Wisdom of the Inner Crowd" Can Boost Accuracy of Confidence Judgments* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ngweh

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

Martín-Luengo, B., Zinchenko, O., Dolgoarshinnaia, A., & Alekseeva, M. (2024). Normative study of 500 general-knowledge of true-false questions for Russian young adults. *PloS One*, *19*(4), e0300600. https://doi.org/10.1371/journal.pone.0300600

Mummolo, J. (2019). Demand Effects in Survey Experiments: An Empirical Assessment. *The American Political Science Review*, *113*(2), 517–529.

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, *12*(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Murphy, A. H., & Winkler, R. L. (1971). forecasters and probability forecasts: Some current problems. *Bulletin of the American Meteorological Society*, *52*(4), 239–247.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 338–368. https://doi.org/10.1016/S0022-5371(80)90266-2

Nimble Links: Links, unleashed. (n.d.). Retrieved 30 September 2024, from https://www.nimblelinks.com/

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General and Applied*, *76*(28), 1–27. https://doi.org/10.1037/h0093849

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles*, *48*(5/6), 265–276. https://doi.org/10.1023/A:1022877405718

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419. https://doi.org/10.1017/S1930297500002205

Peirce, C. S., & Jastrow, J. (1884). On Small Differences in Sensation. *Memoirs of the National Academy of Sciences*, *3*, 75–83.

Plous, S. (1993). *The psychology of judgment and decision making*. McGraw-Hill.

Privacy—Nimble Links. (n.d.). Retrieved 6 October 2024, from https://www.nimblelinks.com/privacy

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*(2), 193–218. https://doi.org/10.1016/0749-5978(87)90012-4

Schoenherr, J. R., Leth-Steensen, C., & Petrusic, W. M. (2010). Selective attention and subjective confidence calibration. *Attention, Perception, & Psychophysics*, *72*(2), 353–368. https://doi.org/10.3758/APP.72.2.353

Schoenherr, J. R., & Petrusic, W. M. (2015). Scaling internal representations of confidence: Effects of range, interval and number of response categories. *Proceedings of the 31st Annual Meeting of the International Society for Psychophysics*, 36–37.

Statistik Austria. (2022). *Bevölkerung nach Alter/Geschlecht*. STATISTIK AUSTRIA. Retrieved 3 October 2024, from https://www.statistik.at/statistiken/bevoelkerung-und-soziales/bevoelkerung/bevoelkerungsstand/bevoelkerung-nach-alter/geschlecht

Statistik Austria. (2022). *Bildungsstand der Bevölkerung*. STATISTIK AUSTRIA. Retrieved 3 October 2024, from https://www.statistik.at/statistiken/bevoelkerung-und-soziales/bildung/bildungsstand-der-bevoelkerung

Tauber, S., Dunlosky, J., Rawson, K., Rhodes, M., & Sitzman, D. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods*, *45*. https://doi.org/10.3758/s13428-012-0307-9

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

US Census Bureau. (2022). *Census Bureau Releases New Educational Attainment Data*. Retrieved 3 October 2024, from https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html

US Census Bureau. (2023). *America Is Getting Older*. Retrieved 3 October 2024, from https://www.census.gov/newsroom/press-releases/2023/population-estimates-characteristics.html

US Department of Labor. (n.d.). *Minimum Wage*. U.S. Department of Labor. Retrieved 4 August 2024,

from https://www.dol.gov/general/topic/wages/minimumwage

Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2017). Known Unknowns: A Critical Determinant of Confidence and Calibration. *Management Science*, *63*(12), 4298–4307. https://doi.org/10.1287/mnsc.2016.2580

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology*, *88*(3), 490–499. https://doi.org/10.1037/0021-9010.88.3.490

Wertgen, A. G., & Richter, T. (2023). General knowledge norms: Updated and expanded for German. *PLOS ONE*, *18*(2), e0281305. https://doi.org/10.1371/journal.pone.0281305

Whitcomb, K. M., Önkal, D., Curley, S. P., & George Benson, P. (1995). Probability judgment accuracy for general knowledge. Cross-national differences and assessment methods. *Journal of Behavioral Decision Making*, *8*(1), 51–67. https://doi.org/10.1002/bdm.3960080105

8. Appendix

The Appendix is structured in chronological order with the respective figures' and tables' mention in the body of text.

8.1 List of Figures

8.2 List of Tables

8.3 Figures & Tables

Table 10 - List of Survey Questions

	Question	Answer	Lure	Category	Difficulty	Source
1.	What is the last name of the man who began the Reformation in Germany?	Luther	Hus	religion	easy	Coane & Umanath, 2021
2.	Of which country is Budapest the capital?	Hungary	Romania	world	easy	Coane & Umanath, 2021
3.	What is the last name of the first person to climb Mount Everest?	Hillary	Brown	sports	easy	Coane & Umanath, 2021
4.	What is the name of the Chinese religion founded by Lao Tse?	Taoism	Confucianism	religion	easy	Coane & Umanath, 2021
5.	What is the capital of Canada?	Ottawa	Toronto	world	easy	Coane & Umanath, 2021
6.	What is the last name of the man who invented dynamite?	Nobel	Oppenheimer	invention	easy	Coane & Umanath, 2021
7.	What is the name of the brightest star in the sky excluding the sun?	Sirius	Polaris	astronomy	easy	Coane & Umanath, 2021
8.	What is the capital of Australia?	Canberra	Sydney	world	hard	Coane & Umanath, 2021
9.	What animal runs the fastest?	Cheetah	Ostrich	zoology	easy	Tauber et al., 2013
10.	How long is the gestation time of an elephant?	22 months	18 months	zoology	hard	Litvinova et al., 2019
11.	From what fruit is wine normally made?	grapes	cherries	food	easy	Buades-Sitjar et al., 2021
12.	In which Disney movie does the heroine enlist in the military and save China from the Huns?	Mulan	Moana	culture	easy	Buades-Sitjar et al., 2021
13.	What is the belief that maximizing pleasure is the way to achieve happiness called?	Hedonism	Stoicism	philosophy	easy	Buades-Sitjar et al., 2021
14.	Who painted "The Last Supper"?	Leonardo da Vinci	Michelangelo	culture	easy	Buades-Sitjar et al., 2021
15.	Approximately when did Gutenberg introduce the printing press in Europe?	Before 1500	After 1500	invention	hard	Buades-Sitjar et al., 2021
16.	From what type of milk is feta cheese primarily made?	Sheep	Goat	food	hard	Litvinova et al., 2019
17.	When was the zipper invented?	Before 1920	After 1920	invention	easy	Litvinova et al., 2019
18.	How many letters are there in the English alphabet?	26	28	linguistics	hard	Buades-Sitjar et al., 2021
19.	What does the Japanese word "origami" mean?	Folding paper	Flying birds	linguistics	hard	Buades-Sitjar et al., 2021
20.	Which word best describes a thankless, excruciating, repetitive, and endless task?	Sisyphean	Otreran	linguistics	hard	Buades-Sitjar et al., 2021
21.	Edgar Allan Poe was:	American	Englishman	literature	easy	Buades-Sitjar et al., 2021

8.4 Method Design and Qualtrics Examples

Figure 24 - Method Design for Qualtrics

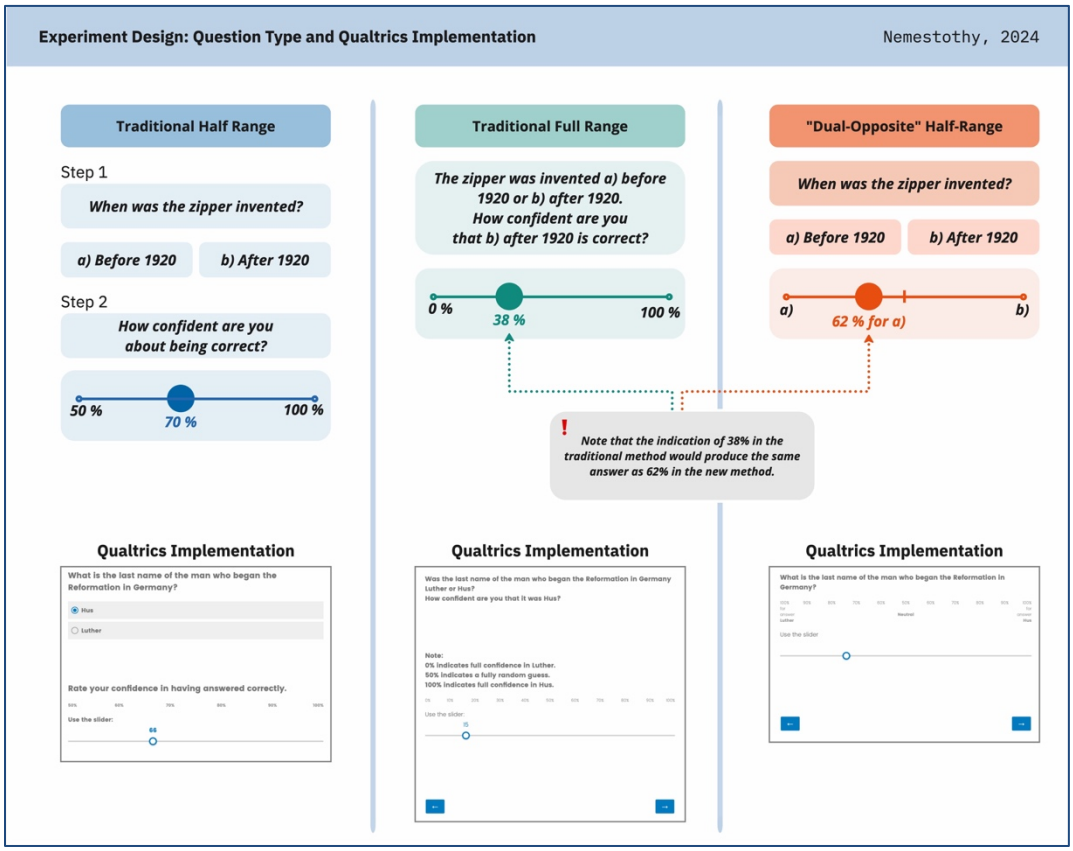
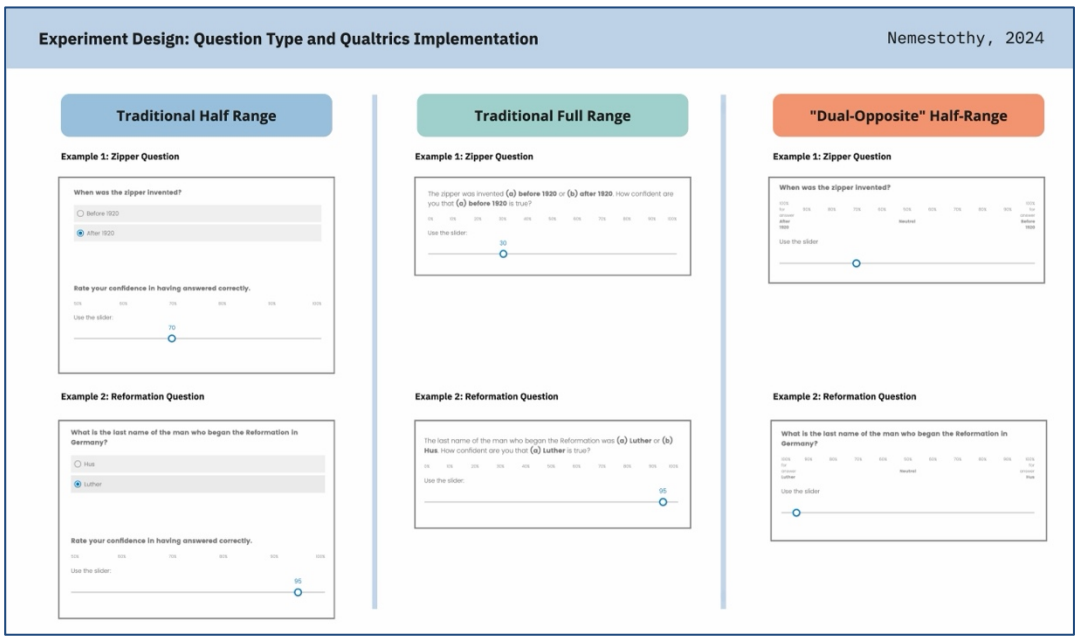


Figure 25 - Method Design for Qualtrics - Examples



8.5 Survey Instructions

Figure 26 - Survey Instructions - Group 1 Half-Range

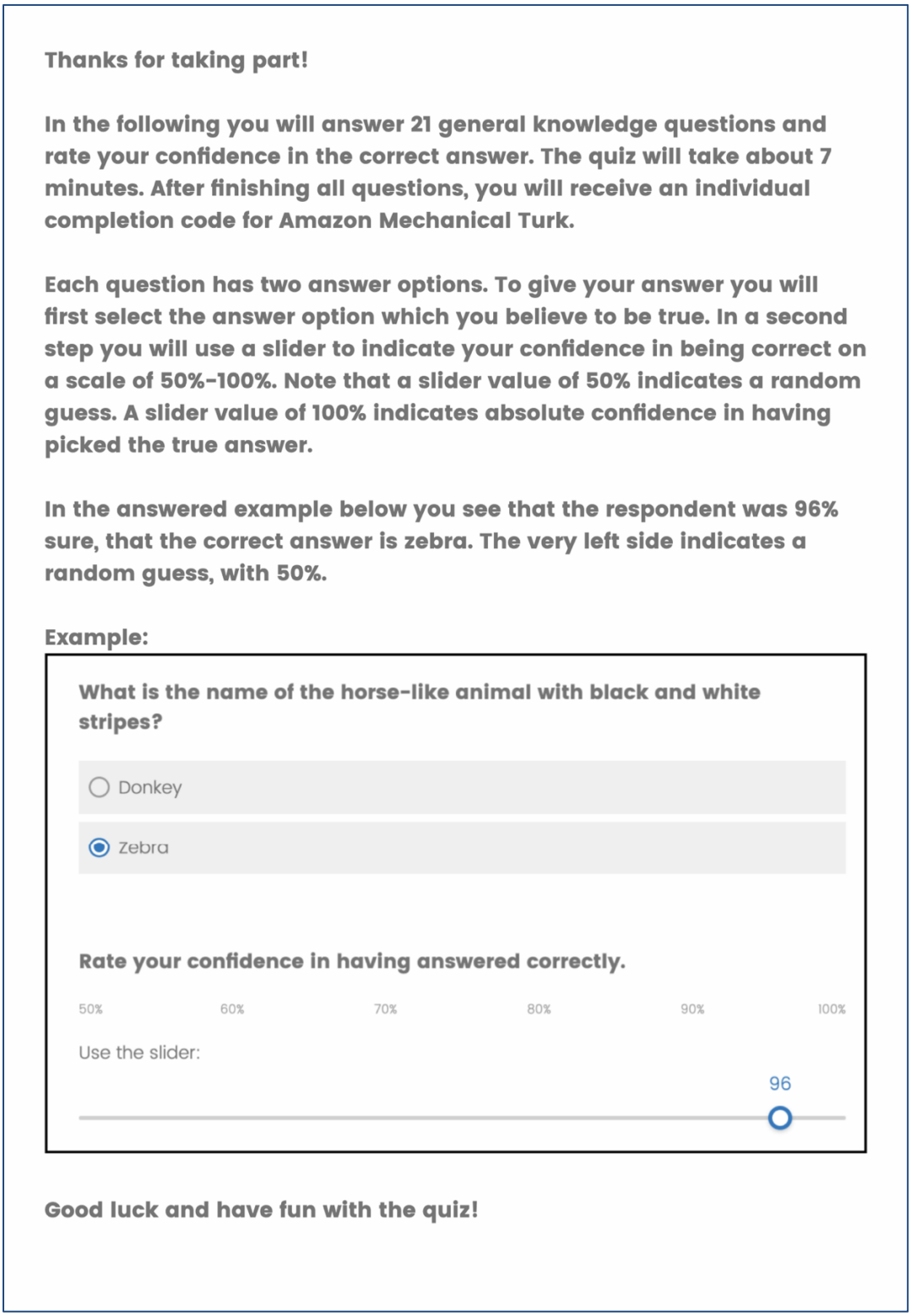


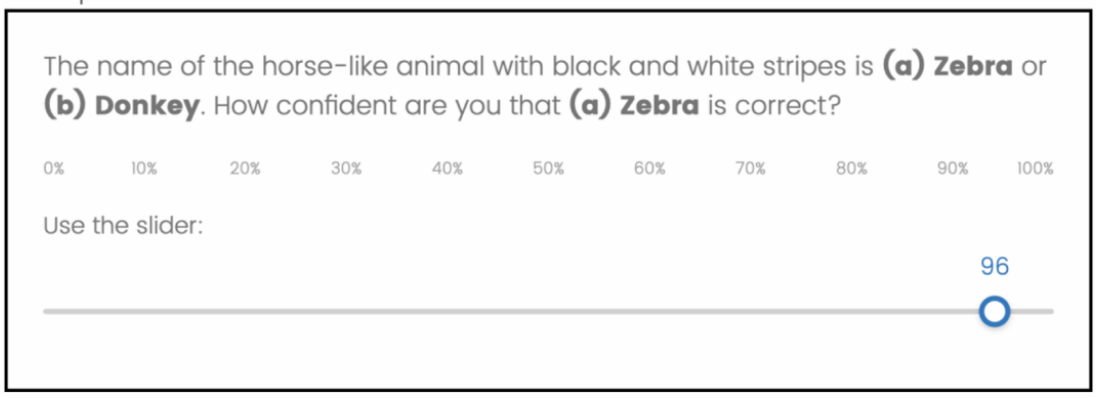
Figure 27 - Survey Instructions - Group 2 Full-Range

Thanks for taking part!

In the following you will answer 21 general knowledge questions and rate your confidence in the correct answer. The quiz will take about 7 minutes. After finishing all questions, you will receive an individual completion code for Amazon Mechanical Turk.

Each question has two answer options. For every question one answer option is preselected. Your task is to indicate how sure you feel about the preselected answer to be correct. You can do this by drawing a slider to the left or right.

Example:



Drawing the slider to the right end to 100% indicates that you are totally confident that the preselected answer is correct.
The middle at 50% means that you are indifferent between the two answer options, thus a random guess.
Drawing the slider to the left end at 0% means that you are totally confident that the preselected answer is wrong, thus the alternative answer option being correct.

In the answered example above you see that the respondent was about 96% sure, that the correct answer is (a) zebra.

Good luck and have fun with the quiz!

Figure 28 - Survey Instructions - Group 3 Dual-Opposite

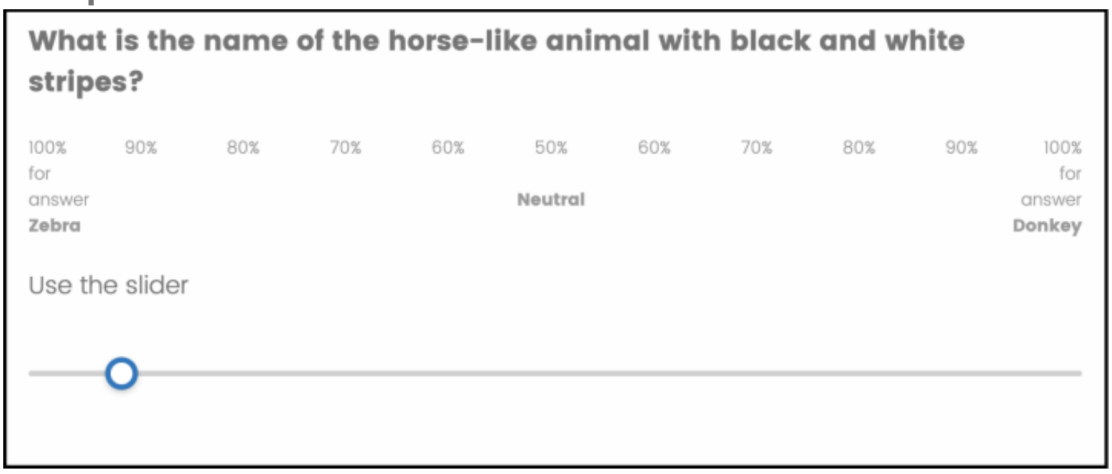
Thanks for taking part!

In the following you will answer 21 general knowledge questions and rate your confidence in the correct answer. The quiz will take about 7 minutes. After finishing all questions, you will receive an individual completion code for Amazon Mechanical Turk.

Each question has two answer options. To give your answer you will draw a slider to the side of answer option you believe to be true. The further you draw the slider, the more confident you indicate in being correct. Pay attention to your confidence level!

In the answered example below you see that the respondent was about 92% sure, that the correct answer is zebra. The middle would indicate a random guess, thus "Neutral". The very right shows the second answer option "Donkey".

Example:



Good luck and have fun with the quiz!

8.6 Nimble Links Tracking

Figure 29 - Nimble: MTurk Redirections

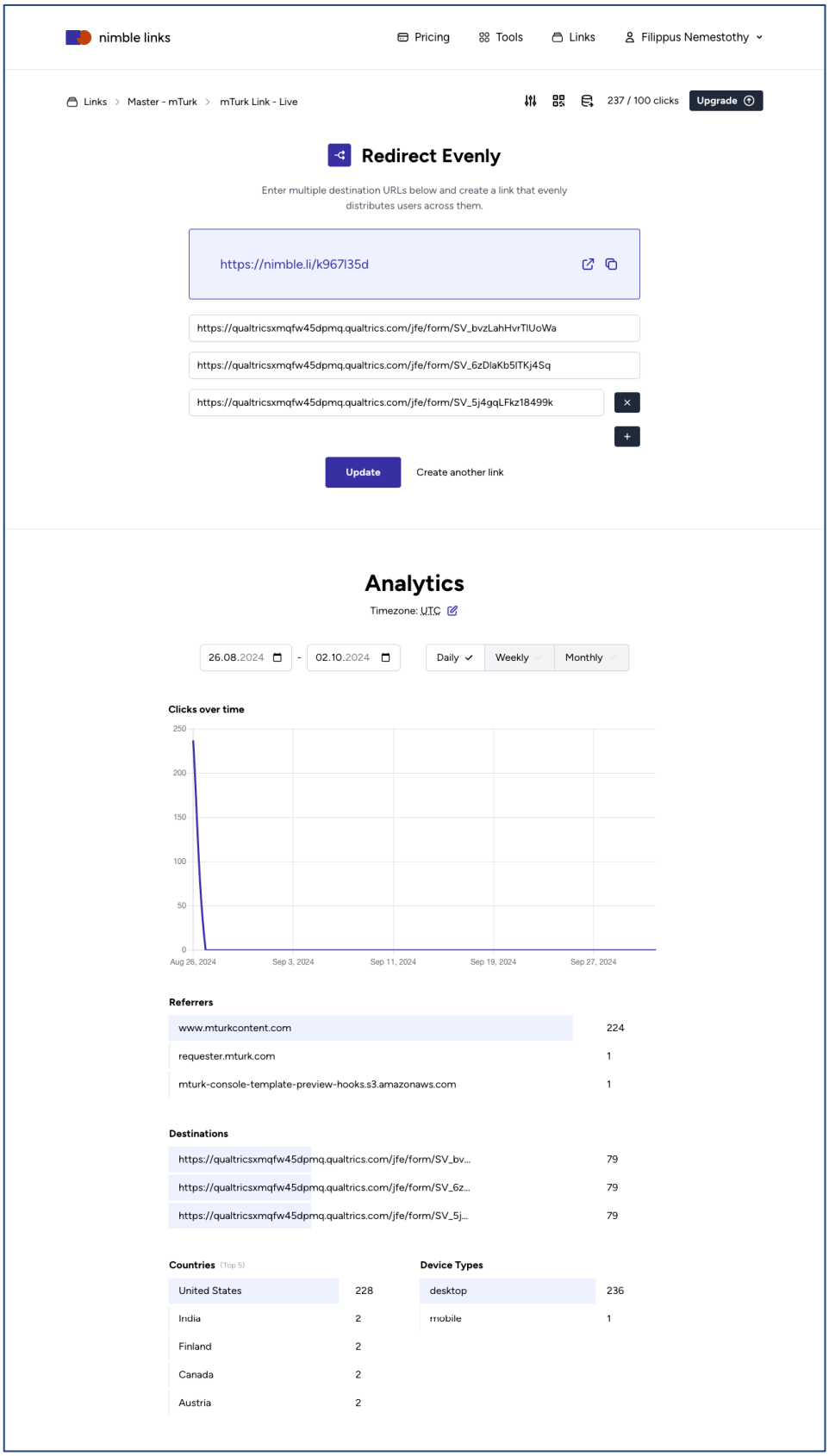
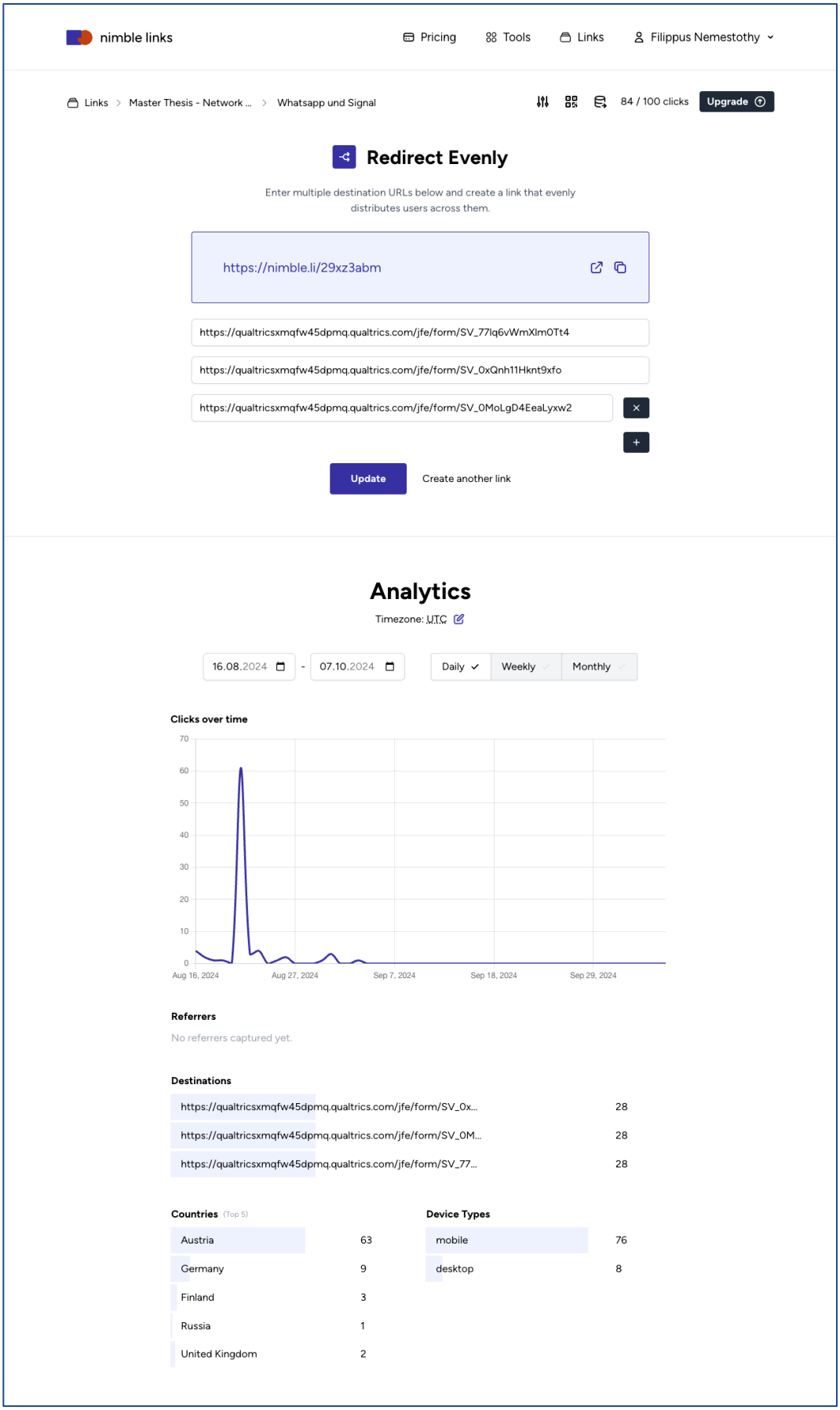


Figure 30 - Nimble: Direct Messaging Redirections



8.7 Data Treatment and Cleaning

Table 11 - Data treatment and cleaning

#	Sample	Group	Number	Reason	ID / *Country*
	Raw Data Import				
1.	MTurk	All	184 imported	Initial Import	-
2.	Private Network	All	135 imported	Initial Import	-
	Pre-Treatment of raw data in MS Excel				
3.	MTurk	All	2 deleted	Incorrect Completion Code	-
4.	MTurk	All	22 deleted	Unfinished Survey	-
5.	MTurk	All	*7 included*	Unfinished status but correctly completed	7ReKIV3rB0WSBdo; R_1GsDUzvYHpSVzAl; R_5dZqWnbx3SKsNdo; R_3R0G8ZHfoOw0Frz; R_1psPxoX0KwvQF85; R_5mh0Wef5LDnzR7r; R_1wHp8keZKPygrLh
6.	Private Network	All	36 deleted	Unfinished Survey	-
7.	Private Network	All	*5 included*	Unfinished status but correctly completed	R_8Qmtl5VvMyeaTA5; R_8NQd2KqQuGIQ3TT; R_2jGOUhCWxOq8Tqj; R_2jGOUhCWxOq8Tqj; R_83gbFfLBs95ZU6b; R_81ih8zVidnjBkmQ
	Subtotal	MTurk Priv. Network **Sum**	184 99 **283**		
	Treatment				
8.	All	All	29 deleted	Other country of residence than Austria or USA; Deletion for reasons of parsimony and risk of bias	*Australia (4); Albania (5); Algeria (1); Colombia (1); India (Canada (1); Finland (2); Germany (8); Jordan (1); Netherlands (1); Slovenia (1); Spain (1); UK (1)*
9.	All	All	7 deleted	Outlier (When duration = 2 Std. dev. from the mean)	R_1GsDUzvYHpSVzAl; R_7tBNEdZkcLLrC7f; R_1qpmEzoaSIkYRTb; R_1uHVMM8cgzLNN5t; R_2zr0yo2GWe1v1js; R_3GiGPPVqTUNVNdk; R_5V23DnSNTyWRO1N
10.	All	All	14 deleted	Too short duration for conscientious completion (Duration < 180 seconds)	-
11.	**Sum of Deleted Respondents**		50 deleted		
12.	**Final Data Set**		**233 included**		

8.8 Completion Time in Network Sample

Note that these reports were produced with Qualtrics prior to the proper data analysis, hence the format differs to common output tables from conventional statistics programs. These reports were used to estimate the completion time more accurately – again in order to define adequate payment for the MTurk sample.

Figure 31 - Network Sample - Duration - Half-Range

Group 1 - Traditional Half-Range V2 - Network

Duration (in seconds)

Field	Min	Max	Mean	Median	Standard Deviation	Variance	Responses	Sum
Duration (in seconds)	18.00	20775.00	1151.21	315.00	3531.65	12472516.73	39	44897.00

Figure 32 - Network Sample - Duration - Full-Range

Group 2 - Traditional Full-Range V2 - Network

Duration (in seconds)

Field	Min	Max	Mean	Median	Standard Deviation	Variance	Responses	Sum
Duration (in seconds)	9.00	1692.00	430.82	377.00	316.89	100420.01	45	19387.00

Figure 33 - Network Sample - Duration - Dual Opposite

Group 3 - Dual-opposite Half-Range V2 - Network

Duration (in seconds)

Field	Min	Max	Mean	Median	Standard Deviation	Variance	Responses	Sum
Duration (in seconds)	2.00	2460.00	520.30	360.50	560.19	313817.56	46	23934.00

8.9 Statistical analyses

8.9.1 Levene's Test

The p-value is less than 0.05 in all cases, thus the null hypothesis is rejected, meaning that variances are significantly different across all groups.

Table 12 - Levene's Test (Brier by Method)

Confidence Judgment Method	Summary of Mean Brier Score		
	Mean	Std. dev.	Freq.
HR	.12433684	.0830	76
FR	.1792229	.118	73
DO	.19218067	.140	83
Total	.16587874	.1200742	232

W0 = 11.6743736 df(2, 229) Pr > F = 0.00001486

W50 = 9.7304164 df(2, 229) Pr > F = 0.00008793

W10 = 10.8641754 df(2, 229) Pr > F = 0.00003107

8.9.2 Test for multicollinearity

Table 13 - First regression

VARIABLES	(1) Mean Brier Score
Duration (in seconds)	-5.22e-05***
	(1.57e-05)
Sample = 2, network	0.0595
	(0.124)
Confidence Judgement Method = 2, FR	0.0400**
	(0.0186)
Confidence Judgement Method = 3, DO	0.0623***
	(0.0178)
Device = 2, omitted	-
Age	0.00223*
	(0.00134)
Gender = 2, female	0.0159
	(0.0165)
Educational Level = 2, Some high school or less	0.0263
	(0.122)
Educational Level = 3, High school diploma or GED	0.0275
	(0.115)
Educational Level = 4, Some college, but no degree	-0.0579
	(0.119)
Educational Level = 5, Associates or technical degree	-0.0625
	(0.126)
Educational Level = 6, Bachelor	-0.0623
	(0.114)
Educational Level = 7, Graduate or professional degree (MA, ..	-0.0847
	(0.115)
Country of residence = 15, United States of America	0.0359
	(0.125)
Constant	0.102
	(0.172)
Observations	232
R-squared	0.194

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Variance Inflation Factors

Four variables were omitted from the model to avoid multicollinearity after an initial regression (see Table 13) and test for variance inflation factors (see Table 14). The dropped variables are Device (desktop/mobile), Sample (MTurk/Network), Education (7 categories) and Residence (USA/AT).

Table 14 - Test for Variance Inflation Factors

	VIF	1/VIF
2.method	1.413	.708
3.method	1.377	.726
duration	1.121	.892
2.sample	65.946	.015
age	1.347	.742
2.gender	1.126	.888
2.education	7.095	.141
3.education	26.651	.038
4.education	9.934	.101
5.education	5.099	.196
6.education	61.296	.016
7.education	54.892	.018
15.residence	67.527	.015
Mean VIF	23.448	.

8.9.3 Checks for Normality of Distribution

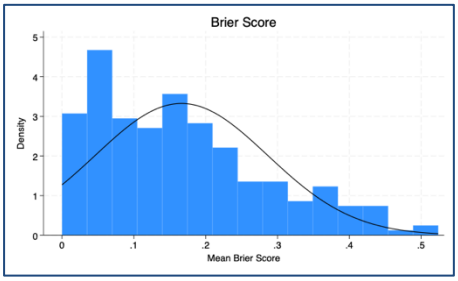
The normal distribution of the data of the key variables was checked. The conducted Shapiro-Wilk tests (see Table 15) show that all null-hypotheses can be rejected, hence confirming that almost all variables are skewed, as visible in respective histograms (see example with the Brier score in Figure 34 and further graphs for all variables in FIGURE XX).

The skewed distribution of the data requires further treatment for certain statistical analyses. Following the Central Limit Theorem, which states that the sampling distribution of the mean will be approximately normal for large sample sizes (>30), parametric tests are applied in qualified cases.

Table 15 - Shapiro Wilk Test for Normal Data

Variable	Obs	W	V	z	Prob>z
duration	233	0.797	34.561	8.214	0.000
age	233	0.936	10.895	5.537	0.000
education	233	0.865	22.954	7.265	0.000
brier	233	0.940	10.150	5.373	0.000
correct_answers	233	0.961	6.673	4.401	0.000
ratio_correct	233	0.961	6.673	4.401	0.000
confidence	233	0.918	14.006	6.120	0.000
overconfidence	233	0.942	9.904	5.316	0.000

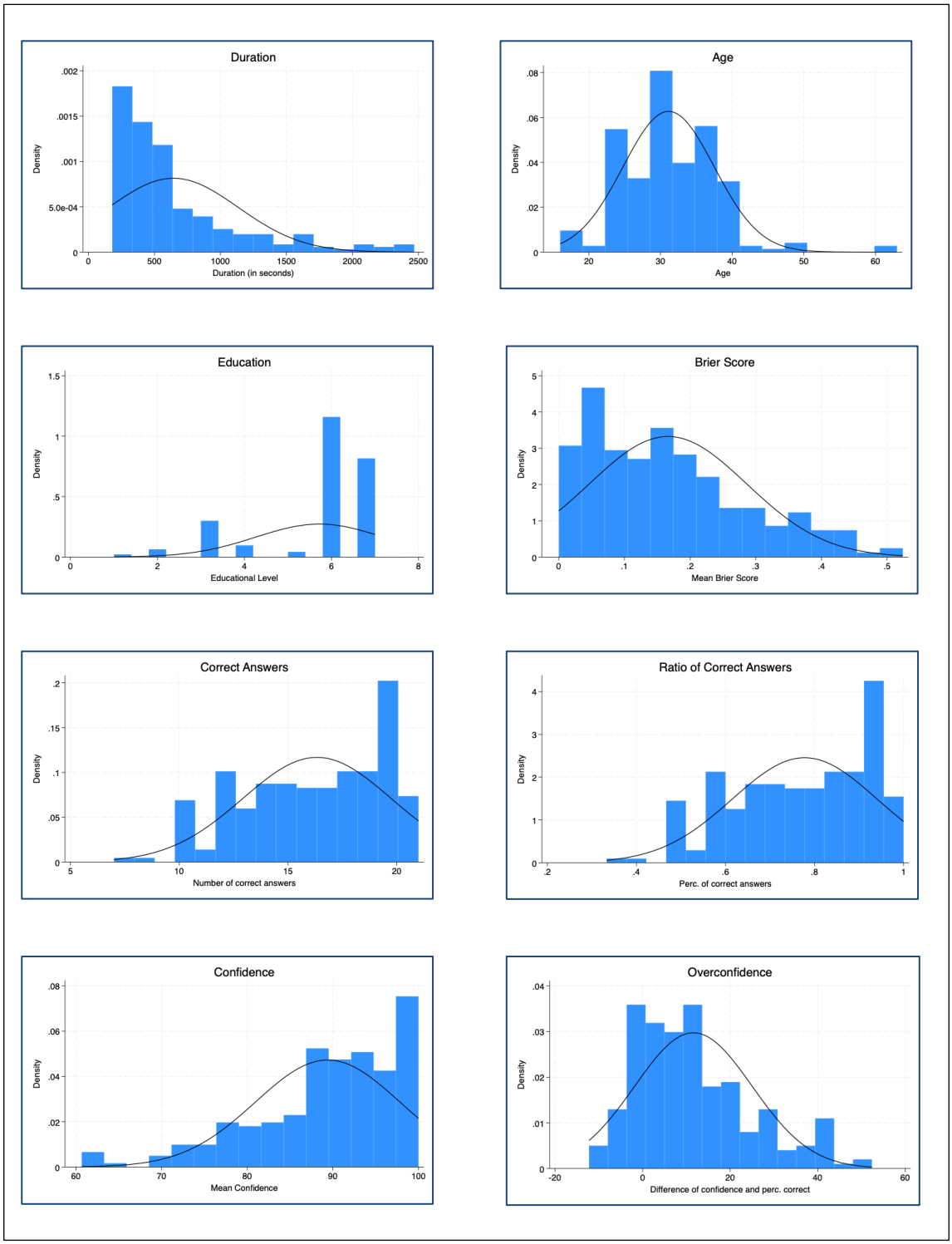
Figure 34 - Histogram - Brier Score



Histograms of relevant variables to check for skewness.

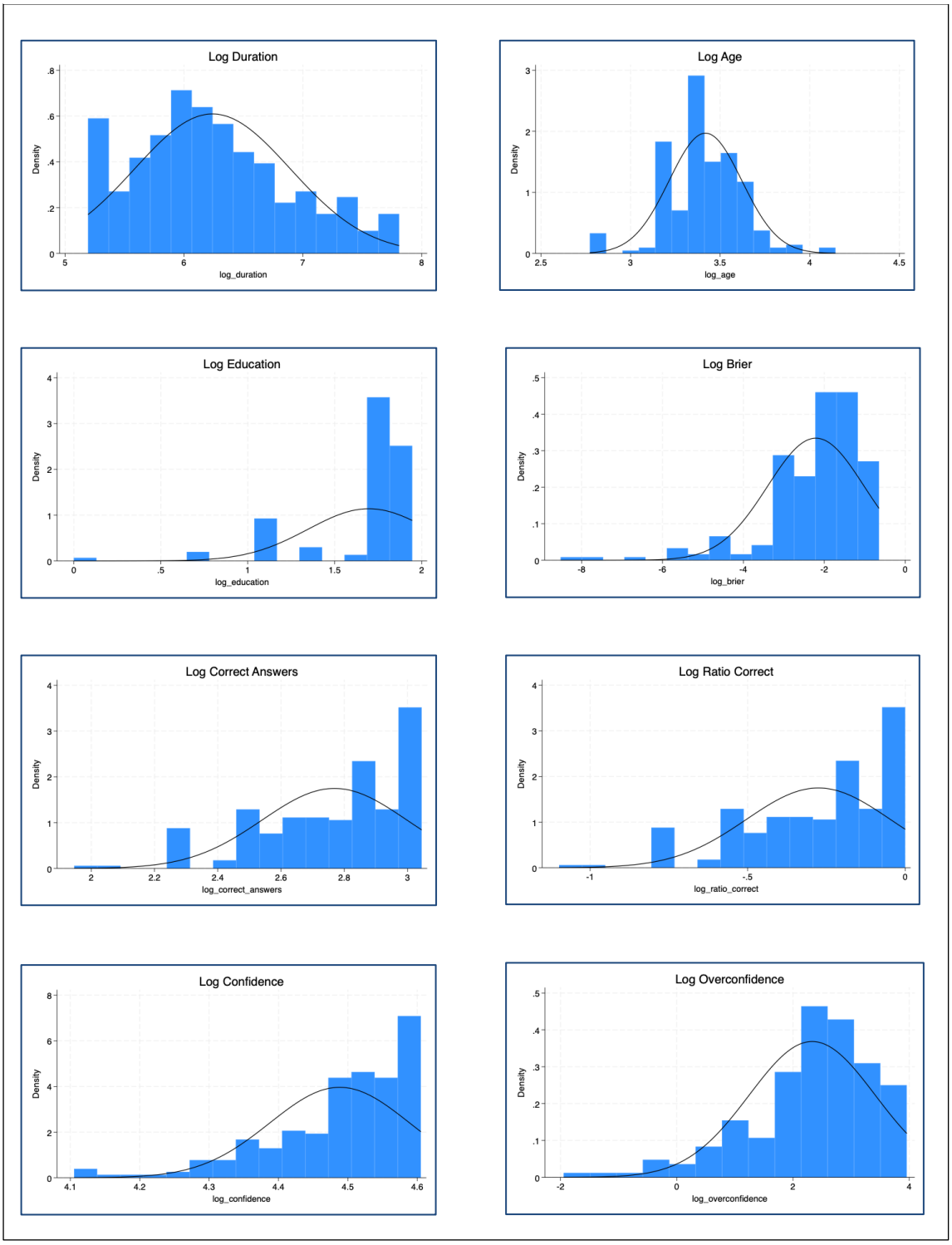
All variables except age appear skewed.

Table 16 - Histograms of all model variables



Histograms of the logarithmic variables.

Table 17 - Histograms of all logarithmised model variables



8.9.4 Regression

Table 18 - Multiple Linear Regressions

VARIABLES	(1) Model 1	(2) Model 2	(3) Combined	(4) Interaction
Full Range	0.052***		0.046**	0.101***
	(0.019)		(0.019)	(0.033)
Dual Opposite	0.064***		0.066***	0.069**
	(0.019)		(0.018)	(0.030)
Duration (in seconds)		-0.000***	-0.000***	-0.000
		(0.000)	(0.000)	(0.000)
FR * Duration				0.000
				(0.000)
HR * Duration				-0.000**
				(0.000)
DO * Duration				-0.000
				(0.000)
Age	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Female	0.002	0.008	0.004	0.008
	(0.017)	(0.017)	(0.017)	(0.017)
Network	0.032*	0.028*	0.022	0.021
	(0.017)	(0.017)	(0.017)	(0.017)
Constant	0.077*	0.162***	0.126***	0.109**
	(0.041)	(0.043)	(0.044)	(0.046)
Observations	232	232	232	232
R-squared	0.079	0.061	0.113	0.134

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Breusch-Pagan/Cook-Weisberg tests for heteroskedasticity

Model 1: Brier Method Age Gender
Assumption: Normal error terms
Variable: Fitted values of brier
H0: Constant variance
chi2(1) = 3.46
Prob > chi2 = 0.0627

Combined: Brier Method Duration Age Gender
Assumption: Normal error terms
Variable: Fitted values of brier
H0: Constant variance
chi2(1) = 2.82
Prob > chi2 = 0.0933

Model 2: Brier Duration Age Gender
Assumption: Normal error terms
Variable: Fitted values of brier
H0: Constant variance
chi2(1) = 6.14
Prob > chi2 = 0.0132

Interaction: Brier Method Duration Method*Duration Age Gender
Assumption: Normal error terms
Variable: Fitted values of brier
H0: Constant variance
chi2(1) = 4.22
Prob > chi2 = 0.0400

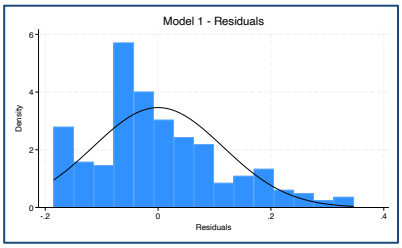
8.9.5 Normality of errors

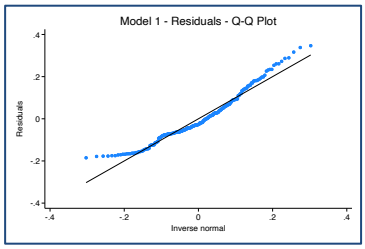
Model 1

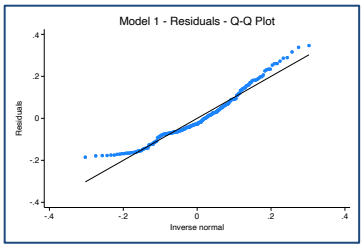
We reject the null hypothesis of the Shapiro-wilk test that the errors are normally distributed, due to the p-value of less than 0.05.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals_m1	232	0.953	7.899	4.791	0.000





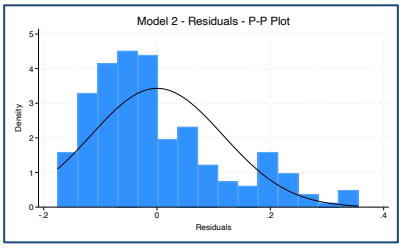


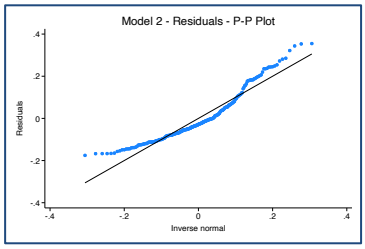
Model 2

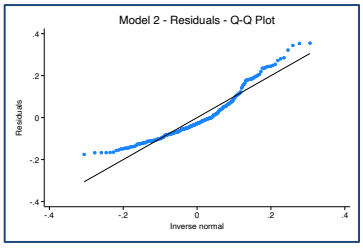
We reject the null hypothesis of the Shapiro-wilk test that the errors are normally distributed, due to the p-value of less than 0.05.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals_m2	232	0.918	13.881	6.098	0.000





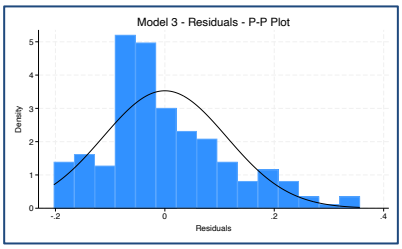


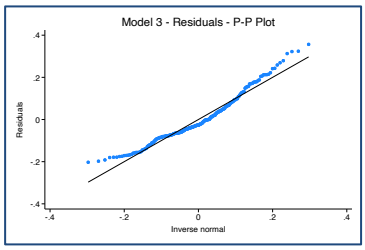
Model 3

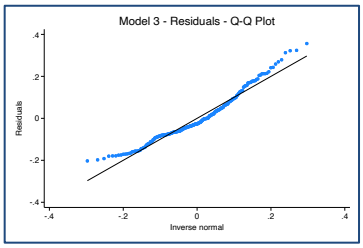
We reject the null hypothesis of the Shapiro-wilk test that the errors are normally distributed, due to the p-value of less than 0.05.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals_m3	232	0.956	7.469	4.661	0.000





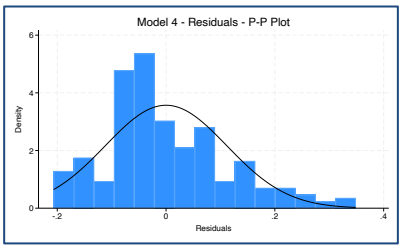


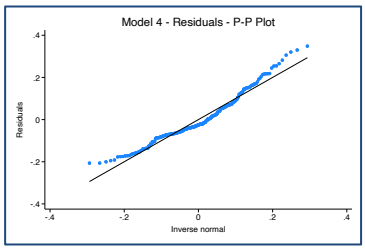
Model 4

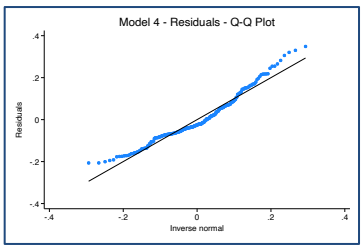
We reject the null hypothesis of the Shapiro-wilk test that the errors are normally distributed, due to the p-value of less than 0.05.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals_m4	232	0.956	7.391	4.637	0.000







Model 1:

Regression equations

Model 1: Brier Score explained by Confidence Judgement Method

$$\boldsymbol{BS = 0.077 + 0.052\,FR + 0.064\,DO + 0.001\,Age + 0.002\,Female + 0.032\,Network}$$

Model 2: Brier Score explained by Duration

$$\boldsymbol{BS = 0.162 + 0.000\,Duration + 0.001\,Age + 0.008\,Female + 0.028\,Network}$$

Model 3: Brier Score explained by Method & Duration

$$\boldsymbol{BS = 0.126 + 0.046\,FR + 0.066\,DO + 0.000\,Duration + 0.001\,Age + 0.002\,Female + 0.032\,Network}$$

Model 4: Brier Score explained by Method, Duration and their interaction

$$\boldsymbol{BS = 0.126 + 0.101\,FR + 0.069\,DO + 0.000\,Duration + 0.000\,HR * Duration + 0.000\,FR * Duration + 0.000\,DO * Duration + 0.001\,Age + 0.002\,Female + 0.032\,Network}$$

Quantile regression

Table 19 – 50% Quantile Regressions

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
Full Range	0.036*		0.031*	0.083***
	(0.020)		(0.018)	(0.024)
Dual Opposite	0.025		0.027	0.040
	(0.023)		(0.024)	(0.034)
Duration (in seconds)		-0.000***	-0.000***	-0.000***
		(0.000)	(0.000)	(0.000)
HR * Duration				0.000
				(0.000)
FR * Duration				-0.000***
				(0.000)
DO * Duration				-0.000
				(0.000)
Age	0.000	-0.000	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Female	-0.001	0.011	-0.002	0.001
	(0.011)	(0.015)	(0.009)	(0.010)
Network	0.072***	0.056***	0.056***	0.048***
	(0.014)	(0.015)	(0.013)	(0.015)
Constant	0.088*	0.149***	0.102**	0.085**
	(0.048)	(0.029)	(0.043)	(0.036)
Observations	232	232	232	232

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 20 - CLM Assumptions for Model 1

Gauss Markov Assumptions				
	Assumption	**Test**	**Treatment**	**Check**
Unbiased	Linearity and additivity in Parameters	Given as all models consist of non-squared equations.	-	✓
	Random sampling	All respondents were independent and identically distributed.	-	✓
	Zero conditional mean	Control variables are included.	-	✓
	No multicollinearity	Variance inflation factors test conducted.	Three variables (residence, device, education) deleted until no multicollinearity was found anymore.	✓
BLUE	Homoskedasticity	Breusch-Pagan test rejects assumption that the variance of the residuals is constant.	-	**X**
CLM	Normality of errors	Shapiro-Wilk test shows skewed residuals, hence NOT normally distributed errors.	More than 30 observations per group allows the assumption of Central Limit Theorem	**X**

Table 21 - ANOVA Stata Output

Num of obs = 232 R-squared =0.0605

Root MSE .116891 Adj R-squared = 0.0523

Source	Partial SS	df	MS	F	Prob>F
Model	.20157296	2	.10078648	7.38	0.0008
Method	.20157296	2	.10078648	7.38	0.0008
Residual	3.1289418	229	.0136635		
Total	3.3305147	231	.01441781		

8.9.6 Test for difference in Brier Score among Sample groups

Table 22 - Wilcoxon rank-sum test

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

sample	Obs	Ranksum	Expected
mTurk	150	15420	17475
network	82	11608	9553
Combined	232	27028	27028
Unadjusted variance	238825.00		
Adjustment for ties	-0.92		
Adjusted variance	238824.08		
H0: brier(sample==mTurk) = brier(sample==network)			
z = -4.205			
Prob >	z	=	0.0000

Note: Exact p-value is not computed by default for sample sizes > 200.

Table 23 - t-Test for Brier score by Sample Group

Two-sample t test with equal variances

	obs1	obs2	Mean1	Mean2	dif	St Err	t value	p value
brier by sample: 1 2	150	82	0.152	.19	-.037	.017	-2.3	.023

Correlation table

Table 24 - Correlation table

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) brier	1.000							
(2) duration	-0.216	1.000						
	(0.001)							
(3) sample	0.149	-0.182	1.000					
	(0.023)	(0.006)						
(4) method	0.232	0.017	0.115	1.000				
	(0.000)	(0.796)	(0.081)					
(5) device	0.149	-0.182	1.000	0.115	1.000			
	(0.023)	(0.006)	(0.000)	(0.081)				
(6) age	0.041	-0.100	-0.131	-0.019	-0.131	1.000		
	(0.530)	(0.127)	(0.046)	(0.778)	(0.046)			
(7) gender	0.060	-0.023	0.222	0.075	0.222	0.018	1.000	
	(0.361)	(0.730)	(0.001)	(0.255)	(0.001)	(0.782)		
(8) residence	-0.153	0.190	-0.991	-0.114	-0.991	0.133	-0.215	1.000
	(0.020)	(0.004)	(0.000)	(0.083)	(0.000)	(0.044)	(0.001)	

Overconfidence Model

A brief check on the possible influence of overconfidence in the model for future research.

Table 25 - Regression: Overconfidence Model

VARIABLES	(1) Overconfidence Model 5	(2) Overconfidence Model 6	(3) Overconfidence Model 7	(4) Overconfidence Model 8 (robust)
Overconfidence	0.008***	0.008***	0.008***	0.008***
	(0.000)	(0.000)	(0.000)	(0.000)
FR		-0.007	-0.003	-0.003
		(0.008)	(0.014)	(0.016)
DO		-0.014*	-0.015	-0.015
		(0.008)	(0.013)	(0.013)
Duration			-0.000	-0.000
			(0.000)	(0.000)
HR*Duration			0.000	0.000
			(0.000)	(0.000)
FR*Duration			-0.000	-0.000
			(0.000)	(0.000)
DO*Duration			0.000	0.000
			(0.000)	(0.000)
Age	-0.001**	-0.001**	-0.001**	-0.001**
	(0.001)	(0.001)	(0.001)	(0.000)
Female	0.004	0.004	0.005	0.005
	(0.007)	(0.007)	(0.007)	(0.007)
Network	0.003	0.004	0.000	0.000
	(0.007)	(0.007)	(0.007)	(0.007)
Constant	0.100***	0.107***	0.123***	0.123***
	(0.016)	(0.017)	(0.019)	(0.018)
Observations	232	232	232	232
R-squared	0.845	0.847	0.851	0.851

F(9, 222) = 195,93 | Prob>F = 0.0000 | R-squared = 0.8511
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity

Assumption: Normal error terms

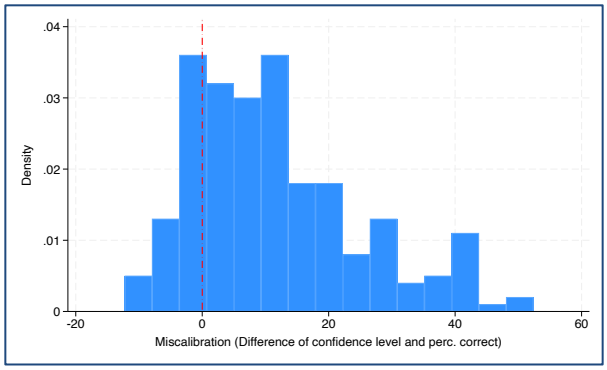
Variable: Fitted values of brier

H0: Constant variance

chi2(1) = 7.96

Prob > chi2 = 0.0048

Table 26 – Histogram: Miscalibration



Zusammenfassung

Ärzte benötigen es gleichermaßen wie Manager: Eine gute Einschätzung der Sicherheit, mit der sie Entscheidungen und Vorhersagen treffen. Das Maß dieser Sicherheit, geläufig als Confidence Calibration, hat in vielen Forschungsdomänen Bedeutung und eine lange Vergangenheit und trotzdem weist die Literatur Lücken auf. So werden zwei häufig verwendete Methoden zur Erfassung der subjektiven qualitativen Beurteilung einer Einschätzung in der Literatur unterschiedlich überlegen beschrieben. Die vorliegende Studie stellt die Hypothese auf, dass die Half-Range Methode (50-100) der Full-Range Methode (0-100) überlegen ist und stellt darüber hinaus eine neue verbesserte Methode vor. In einer experimentellen Studie mit 232 Teilnehmer*innen wurden alle drei Methoden mit Allgemeinwissensfragen (mit je zwei Antwortoptionen und einer erzwungenen Entscheidung) getestet und Brier Scores als Schätzgütemaße berechnet. Die Analyse der Daten bestätigt die Überlegenheit der Half-Range Methode, um genauere Kalibrierung von tatsächlichem Wissen und dessen Einschätzung zu erreichen. Für die neuentwickelte Messmethode konnte keine Verbesserung oder Überlegenheit in der Kalibrierungsgenauigkeit festgestellt werden. Die Studie schließt mit einer kritischen Diskussion der Ergebnisse ab und schlägt mögliche zukünftige Forschungsrichtungen vor.