CONFIDENCE JUDGMENTS AND RATIONAL DECISION MAKING: AN INVESTIGATION OF COGNITIVE PROCESSES AND POSSIBLE INTERVENTIONS

Author: Julia Felfeli

Supervisors:
Prof. Dr. Anja Achtziger
Prof. Dr. Carlos Alós-Ferrer

Disputation: June 25, 2020
Danksagung


## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danksagung</td>
<td>1</td>
</tr>
<tr>
<td>List of Figures</td>
<td>V</td>
</tr>
<tr>
<td>List of Tables</td>
<td>VI</td>
</tr>
<tr>
<td>Summary</td>
<td>1</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>4</td>
</tr>
<tr>
<td>2. Confidence Judgments in Own Skills</td>
<td>7</td>
</tr>
<tr>
<td>2.1. Types of Overconfidence</td>
<td>8</td>
</tr>
<tr>
<td>2.2. The Hard-Easy Effect</td>
<td>9</td>
</tr>
<tr>
<td>2.3. Gender Effects on Confidence Judgments</td>
<td>10</td>
</tr>
<tr>
<td>2.4. Processes of Confidence Judgments</td>
<td>11</td>
</tr>
<tr>
<td>2.5. The Effects of Incentives on Confidence Judgments</td>
<td>13</td>
</tr>
<tr>
<td>3. Present Research</td>
<td>17</td>
</tr>
<tr>
<td>4. Pretest on General Knowledge Questions</td>
<td>21</td>
</tr>
<tr>
<td>4.1. Methods</td>
<td>21</td>
</tr>
<tr>
<td>4.1.1. Participants</td>
<td>21</td>
</tr>
<tr>
<td>4.1.2. Material</td>
<td>21</td>
</tr>
<tr>
<td>4.1.3. Procedure</td>
<td>22</td>
</tr>
<tr>
<td>4.1.4. Results</td>
<td>22</td>
</tr>
<tr>
<td>4.1.5. Discussion</td>
<td>23</td>
</tr>
<tr>
<td>5. Study 1: Confidence Judgments and Their Processes Under Flat Rate Payment</td>
<td>24</td>
</tr>
<tr>
<td>5.1. Methods</td>
<td>26</td>
</tr>
<tr>
<td>5.1.1. Participants</td>
<td>26</td>
</tr>
<tr>
<td>5.1.2. Design</td>
<td>26</td>
</tr>
<tr>
<td>5.1.3. Materials</td>
<td>27</td>
</tr>
<tr>
<td>5.1.4. Procedure</td>
<td>27</td>
</tr>
<tr>
<td>5.2. Results</td>
<td>28</td>
</tr>
<tr>
<td>5.2.1. Number of Correctly Answered Knowledge Questions</td>
<td>28</td>
</tr>
<tr>
<td>5.2.2. Response Times on the Answers to the Knowledge Questions</td>
<td>28</td>
</tr>
<tr>
<td>5.2.3. Confidence Judgments</td>
<td>29</td>
</tr>
</tbody>
</table>
9.4. Research Questions and Hypotheses .............................................. 106
9.5. Methods ................................................................. 107
  9.5.1. Participants .......................................................... 107
  9.5.2. Design ............................................................... 107
  9.5.3. Materials ............................................................. 107
  9.5.4. Decision Task ....................................................... 108
  9.5.5. Decision Rules and Decision Errors ..................................... 109
  9.5.6. Experimental Manipulations .......................................... 110
  9.5.7. Procedure .......................................................... 111
  9.5.8. EEG Acquisition and Analysis ....................................... 113
9.6. Results ................................................................. 114
  9.6.1. Equivalence of Groups ............................................. 114
  9.6.2. Error Rates ........................................................ 116
  9.6.3. ERP Analysis – FRN ............................................... 121
9.7. Discussion .............................................................. 124

10. Conclusion ............................................................... 129

References ................................................................. 132

A. Appendix ................................................................. VII
  A.1. Pretest ................................................................. VII
  A.2. Material: Confidence Judgments ........................................ IX
  A.3. Material: Intentions & Rationality ..................................... XVI
List of Figures

5.1. Study 1: Response time on answers by correctness .......................... 29
5.2. Study 1: Confidence judgments by correctness .............................. 31
5.3. Study 1: Response time on confidence judgments by correctness ...... 32
5.4. Study 1: The hard-easy effect ............................................. 34
5.5. Study 1: Confidence bias by response time on confidence judgments ... 36

6.1. Study 2: Response time on answers by incentive condition .............. 46
6.2. Study 2: Response time on answers by correctness .......................... 46
6.3. Study 2: Confidence judgments by correctness and gender ............... 49
6.4. Study 2: Response time on confidence judgments by incentive condition . 51
6.5. Study 2: Response time on confidence judgments by correctness ....... 52
6.6. Study 2: The hard-easy effect ............................................. 53
6.7. Study 2: Confidence bias by gender and incentive condition ............ 54
6.8. Study 2: Confidence bias by response time on confidence judgments (for incentive for performance condition) .......................... 57

7.1. Study 3: Response time on answers by correctness .......................... 74
7.2. Study 3: Confidence judgments by incentive level and gender .......... 75
7.3. Study 3: Confidence judgments by correctness and gender ............... 77
7.4. Study 3: Response time on confidence judgments by correctness and incentive level .................................................. 80
7.5. Study 3: The hard-easy effect ............................................. 81
7.6. Study 3: Confidence bias by gender and incentive level .................... 82
7.7. Study 3: Confidence bias on questions of medium difficulty by incentive level .......................... 84

9.1. Distribution of the colored balls in both states of the world if blue balls are rewarded .......................................................... 109
9.2. Overview of the sequence of one trial ...................................... 113
9.3. Error frequency over time .................................................... 117
9.4. Understanding and reinforcement errors by condition ...................... 118
9.5. Grand-average of difference wave (all participants) ....................... 123
9.6. Grand-average of difference wave by condition ............................ 123
9.7. Topography of the FRN ..................................................... 124

V
List of Tables

5.1. Study 1: Linear regressions on confidence bias ........................................... 35
6.1. Study 2: Linear regressions on confidence bias ........................................... 56
7.1. Study 3: Linear regressions on confidence bias ........................................... 83
9.1. Descriptive statistics on valence and arousal ............................................. 115
9.2. Descriptive statistics of error rates by error type ....................................... 117
9.3. Random-effects panel probit regression on second-draw error (0 = correct choice, 1 = error) ................................................................. 120
9.4. Linear regressions on FRN amplitude (absolute values) ............................. 122
Summary

Confidence judgments and decision making are part of everyday life. In an ideal world, people would assess their skills and knowledge accurately and base their decisions only on rational deliberations. Yet, this is often not the case. Confidence judgments in own skills or performance are often biased (e.g., Dunning, 2011; Moore & Healy, 2008; Moore & Schatz, 2017; Sanchez & Dunning, 2018; Pikulina, Renneboog, & Tobler, 2017; Michailova & Schmidt, 2016). Also, people tend to deviate from rational decision strategies (e.g., Achtziger & Alós-Ferrer, 2014; Alós-Ferrer, Hügelschäfer, & Li, 2016, 2017; Charness, Karni, & Levin, 2010; Erev, Shimonovich, Schurr, & Hertwig, 2008; Fiedler, Brinkmann, Betsch, & Wild, 2000; Tschan et al., 2009). Therefore, the research aim of the present dissertation was twofold. In the first chapter of the present dissertation I investigated confidence judgments in own skills and the confidence bias, the processes underlying these confidence judgments, and the influences of gender and monetary incentive on confidence judgments. The second aim was to investigate the influence of goal and implementation intentions on rational decision making and how this influence is reflected in the neural correlate of reinforcement learning.

A common finding in research on confidence judgments is the confidence bias (e.g. Moore & Schatz, 2017; Moore & Healy, 2008; Pikulina et al., 2017; Sanchez & Dunning, 2018; Lebreton et al., 2018). In most cases, the confidence bias reflects overconfidence, which means that people’s subjective confidence exceeds their actual ability or performance (Fischhoff, Slovic, & Lichtenstein, 1977). In some cases, there is also evidence for underconfidence, suggesting that people underestimate their abilities (Kruger & Dunning, 1999; Kruger & Burrus, 2004). Gender is an important predictor of the confidence bias. Underconfidence is more prevalent in females, whereas males often display overconfidence (e.g., Barber & Odean, 2001; Hügelschäfer & Achtziger, 2014; Niederle & Vesterlund, 2007). In Study 1, I investigated the processes underlying confidence judgments and the confidence bias by means of response times, and I examined potential gender differences. Participants answered general knowledge questions and judged their confidence on the correctness of each answer. Participants had overall a good sense of whether their answer was correct or incorrect. This was reflected by higher confidence judgments on correct
answers compared to incorrect ones. The analysis of response times on the confidence judgments revealed that male participants who took longer to judge their confidence made more accurate judgments than males who responded quickly. This relationship was not found for females. In Study 2, half of the participants received a monetary incentive for good performance in the general knowledge test. The monetary incentive for performance increased the time invested in both tasks (the knowledge questions and the confidence judgments). However, this increased effort did not lead to better performance on the knowledge questions, nor did it yield more accurate confidence judgments. The response times suggested that males who invested more time in the confidence judgments were more accurate (as in Study 1). However, the opposite was true for females. The more time females invested in their judgment the more underconfident they were. This influence of the response time on the confidence bias was only found for incentivized participants. In Study 3, the accuracy of the confidence judgment was incentivized. Contrary to the expectations, the monetary incentive did not reduce the confidence bias but led both males and females to be overconfident. In this study, the response time on the confidence judgment did not predict the confidence bias. On the whole, the results demonstrate that (a) the processes of confidence judgments differ between females and males, and (b) the effectiveness of monetary incentives for improving the accuracy of confidence judgments depends strongly on the incentive being contingent on the performance in the task at hand.

The second chapter of the present dissertation investigated the influence of goal and implementation intentions (P. M. Gollwitzer, 1999) on rational decision making (see also Hügelschäfer & Achtziger, 2017). The impact of intentions was examined by the neural correlate of reinforcement learning, i.e. the feedback-related negativity (FRN; Holroyd & Coles, 2002). Participants worked on a probability updating task in which the optimal strategy to maximize the expected payoff was to follow Bayes’ rule by integrating new information with prior probabilities (Bayes & Price, 1763). The optimal decision rule conflicted with a simpler suboptimal decision strategy, i.e. the reinforcement heuristic (see Achtziger & Alós-Ferrer, 2014; Charness & Levin, 2005). The goal and implementation intention manipulation was proposed to control the automatic process of the reinforcement heuristic and hence foster rational decision making. The results showed that the goal intention and the implementation intention had no influence on the number of reinforcement errors (in contrast to the findings of Hügelschäfer & Achtziger, 2017). However, both, the goal and implementation intentions increased the amplitude of the FRN which, on the neural level, indicated a stronger reliance on the reinforcement
heuristic than in the control group. The findings shed some light on the impact of goal and implementation intentions on rational decision making. They demonstrate that careful consideration of the use of intentions as an intervention for improved decision making is required to avoid undesired side effects. Taken together, the present dissertation provided new insights into the processes underlying confidence judgments, the confidence bias, rational decision making, and its neural correlates.
1. Introduction

Confidence plays a fundamental role in everyday life. On a daily basis, people face multiple situations in which it is necessary to make confidence judgments concerning their own skills. This ranges from small trivial judgments as for example ‘Can I carry all the grocery bags at once?’ to important confidence judgments with long-term consequences as for instance ‘Do I have the skills to apply for this job?’. The rigorous study of confidence judgments is driven by the crucial role confidence plays in goal striving. Confidence serves as a form of internal, subjective feedback that helps to assess one’s own ability in pursuing a desired goal. The quality of this feedback depends strongly on the accuracy of the confidence judgment.

However, research in psychology and economics has shown that people tend to have biased confidence judgments (e.g., Dunning, 2011; Lebreton et al., 2018; Michailova & Schmidt, 2016; Moore & Healy, 2008; Moore & Schatz, 2017; Pikulina et al., 2017; Sanchez & Dunning, 2018). A common finding is the overconfidence bias whereby people’s subjective confidence exceeds their actual ability or performance (Fischhoff et al., 1977). Although overconfidence is the predominant finding in confidence research, there is also evidence showing that people are underconfident in some situations meaning they underestimate their abilities (Kruger & Dunning, 1999; Kruger & Burrs, 2004). This phenomenon is particularly prevalent in females (e.g., Barber & Odean, 2001; Hügelschäfer & Achtziger, 2014; Niederle & Vesterlund, 2007). Since biased confidence judgments have many negative consequences (e.g., Barber & Odean, 2001; Camerer & Lovallo, 1999; Goel & Thakor, 2008; Malmendier & Tate, 2005; Pikulina et al., 2017; Zant & Moore, 2013), it is crucial to investigate what motivates people to be accurate in their confidence judgments and what could be possible interventions to improve the accuracy of confidence judgments in own skills. In order to test for the effectiveness of such interventions, it is necessary to first understand the processes behind confidence judgments.

The first chapter of the present dissertation aimed to investigate the processes of confidence judgments in own skills and the confidence bias. Based on a dual stage
account of confidence (see Moran, Teodorescu, & Usher, 2015), I examined how post-decisional information accumulation influences confidence judgments. Moreover, the research focused on gender differences in confidence judgments and the underlying processes, as well as the potentially beneficial effect of monetary incentives on reducing the confidence bias. I investigated whether incentivizing performance generally reduces the confidence bias and in turn increased accuracy, or whether it is necessary to directly incentivize the confidence judgment instead of the performance.

Biased confidence judgments are not the only pitfall people face in their daily lives. When people encounter situations in which a decision needs to be made under uncertainty, decision-makers are prone to rely on suboptimal decision strategies. In the field of economics, humans are traditionally considered as rational decision makers (G. T. Huang, 2005) who optimize the expected payoff of a decision by integrating new information with previous beliefs (Bayes & Price, 1763). However, decision research consistently demonstrated that people tend to deviate from this rational decision strategy (e.g. Achtziger & Alós-Ferrer, 2014; Alós-Ferrer et al., 2016, 2017; Charness et al., 2010; Erev et al., 2008; Fiedler et al., 2000; Tschan et al., 2009). Rationality often competes with heuristics i.e. simple experience-based decision rules. One particular decision strategy that conflicts with rational decision making is the reinforcement heuristic (Charness & Levin, 2005). This heuristic is based on reinforcement learning (see Thorndike, 1911). It simply prescribes to repeat successful behavior and avoid unsuccessful behavior. The reliance on the reinforcement heuristic however may lead to detrimental decisions (see Achtziger & Alós-Ferrer, 2014; Charness & Levin, 2005). Contrary to the traditional view in economics, an intervention to increase rational decision making through monetary incentives was not effective (see Achtziger, Alós-Ferrer, Hügelschäfer, & Steinhauser, 2015) but rather increased reinforcement-based decision errors. Yet, work by Hügelschäfer and Achtziger (2017) demonstrated the effectiveness of goal and implementation intentions in controlling the automaticity of the reinforcement heuristic. In the interest of understanding the effect of intentions on rational decision making more thoroughly, the second chapter of the present dissertation investigated how the impact of intentions is reflected in the neural correlate of reinforcement learning, more specifically the feedback-related negativity (FRN). This event-related potential reflects a reward prediction error (Holroyd & Coles, 2002) and occurs approximately 200-300 ms after the presentation of the decision feedback. Hence it constitutes a neural indicator of automatic reinforcement learning (Chase, Swainson, Durham, Benham, & Cools, 2011; Holroyd & Coles, 2002; Holroyd, Coles, & Nieuwenhuis, 2002; Holroyd, Yeung, Coles, & Cohen, 2005). The aim of the
present research in the second chapter was to investigate how the manipulation of goal and implementation intentions affects rational decision making and how this is reflected by electrophysiological measurements.

In conclusion, the present dissertation aimed to investigate confidence judgments and rational decision making. Furthermore, I tested the effects of different interventions that were intended to improve confidence judgments and rational decision making. The findings of the present dissertation shed light on the processes underlying confidence judgments, contribute to the clarification of the effect of monetary incentives on the confidence bias and advances the debate about the effectiveness of intentions manipulations on rational decision making and neural correlates of reinforcement learning.
2. Confidence Judgments in Own Skills

How accurate do people evaluate their own abilities? According to psychological and economic research they do it rather poorly, not to say people often overestimate their abilities (Dunning, 2011; Lebreton et al., 2018; Moore & Healy, 2008; Moore & Schatz, 2017). The bias whereby people’s subjective confidence in their own judgments is often greater than their objective accuracy is referred to as overconfidence (Fischhoff et al., 1977). There are many cases in which overconfidence has some benefits for an individual. It significantly strengthens the motivation to overcome obstacles and supports goal attainment (Bi, Liu, Li, & Zhang, 2017; Hirshleifer, Low, & Teoh, 2012) and potentially leads to a higher social status (Anderson, Brion, Moore, & Kennedy, 2012; Kennedy, Anderson, & Moore, 2013). Notably, these are all benefits on the level of the individual (not the societal or company level).

Yet, if you ask people whether overconfidence in own skills is valuable or damaging, most of them might say that people are better off if they were accurate in their confidence judgments. In many cases this is true, because overconfidence per definition means that individuals are inaccurate in the self-evaluation of their skills and performance and that this might lead to problems. In fact, the findings of many studies are in favor of this concern. Overconfidence can damage an individual’s performance (Barber & Odean, 2001), harm a company’s performance and value (Goel & Thakor, 2008; Malmendier & Tate, 2005; Zant & Moore, 2013), and leads to reference group neglect (Camerer & Lovallo, 1999). As an example, Pikulinaa, Renneboog, and Tobler (2017) report that strong overconfidence in financial knowledge results in excessive investment, whereas underconfidence is linked to underinvestment.

The studies reported in the present dissertation try to contribute to the investigation of processes of confidence judgments on knowledge questions. There is an ongoing debate in confidence research between two classes of confidence theories. Single stage models assume that confidence judgments are exclusively based on the information provided by the initial decision. One such model is the self-consistency model of subjective confidence (Koriat, 2012a). The model assumes that the response to a two alternative forced choice
questions is based on a random sample of cues and representations of that question. The subjective confidence relies on the consistency with which the answer is favored across the sampled representations. Several studies show that response speed is a mnemonic cue for self-consistency (Koriat, Ma'ayan, & Nussinson, 2006; Koriat, 2008, 2012a). In contrast, dual stage models postulate that confidence judgments are influenced by information acquired after the initial decision. One example is the Collapsing Confidence Boundary Model (Moran et al., 2015). The model assumes that evidence accumulation continues after the initial choice and that with the passage of time, the confidence level decreases due to the decreasing level of choice supportive evidence.

The assumptions of Koriat (2012a) as well as Moran et al. (2015) refer to two alternative forced choice tasks. I expand to multiple-choice questions and concentrate on processes of confidence judgments in performance in tasks that do not allow any kind of learning, and that do not provide feedback on performance. By analyzing process data (response times), I examine the processes of confidence judgments and the information the judgments are based on (single stage vs. dual stage). Specifically, I explore processes underlying confidence judgments when performance in a knowledge test or the accuracy of confidence judgments on own performance is incentivized and whether these incentives influence the type of information the confidence judgments rely on.

The literature on confidence judgments often shows that females and males differ in their confidence judgments in own skills and performance. Males are often reported as being overconfident on their performance, while females are often underconfident (e.g., Barber & Odean, 2001; Hügelschäfer & Achtziger, 2014; Niederle & Vesterlund, 2007) or less overconfident than males (Bench, Lench, Liew, Miner, & Flores, 2015; Magnus & Peresetsky, 2018). In order to test whether processes of confidence judgments are different for females and males, I consider gender as a predictor of over- and underconfidence and its potential interaction with incentives for performance and for accurate judgments in the present line of research.

2.1. Types of Overconfidence

Most research on confidence judgments is interested in the correspondence between confidence judgments and actual performance. There are two aspects of this correspondence to be distinguished. Monitoring resolution (also called discrimination accuracy) refers to the within-person confidence accuracy correlation, which reflects the ability to discriminate between correct and incorrect answers (Koriat, 2012a; Liberman & Tversky,
The other aspect is calibration, meaning the absolute discrepancy between confidence judgments and performance (Lichtenstein, Fischhoff, & Phillips, 1982). Calibration reflects the degree to which confidence judgments are realistic or biased (over- or underconfidence). In the present work, I investigate processes of resolution and calibration under incentivization.

Since overconfidence is a broadly used term in psychology and economics, Moore and Healy (2008; see also Moore & Schatz, 2017) distinguish three ways to define it. First, people can overestimate their actual performance. As an example, people overestimate how fast they can finish a marathon (Krawczyk & Wilamowski, 2017). Second, people overestimate their abilities relative to others. For instance, 63% of Swedish drivers report that they are more skilled than the median driver in Sweden (Svenson, 1981). Moore and Healy (2008) call this type of overconfidence overplacement. Third, people can be excessively certain on the accuracy of their beliefs. This phenomenon, called overprecision, is observed when participants make numerical estimates and then provide confidence intervals for these approximates. Confidence intervals often turned out to be too narrow (Soll & Klayman, 2004). The present research explores processes of overestimation because this type of overconfidence has been investigated in approximately 60% of empirical publications on confidence judgments (Moore & Schatz, 2017).

Even though the discussion on the confidence bias mainly concentrate on overconfidence and its consequences, people can also be underconfident in their performance. For example, people believe that they perform below average in unicycle riding, computer programming, and their chances of living past 100 (Kruger & Dunning, 1999; Kruger & Burrus, 2004). Apparently, many of us are underconfident in their performance in tasks in which we do not have much, or even no experience at all. The fact that the confidence bias can go both ways only underlines the importance of investigating the processes of confidence judgments.

2.2. The Hard-Easy Effect

A common finding in the research on confidence judgments is the hard-easy effect. It refers to the occurrence of over- and underconfidence affiliated to varying degrees of task difficulty (Brenner, 2003; Grieco & Hogarth, 2009; Griffin & Tversky, 1992; Larrick, Burson, & Soll, 2007; Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982; Merkle, 2009; Moore & Healy, 2008). The hard-easy effect stands for the observation that people are often overconfident in the correctness of their answers on difficult questions while
being underconfident on their performance on easy questions.

Note that a lot of research on the hard-easy effect examined confidence judgments on the performance on easy and difficult questions and neglected the investigation of questions of moderate difficulty (see Grieco & Hogarth, 2009; Hoelzl & Rustichini, 2005; Larrick et al., 2007; Lichtenstein & Fischhoff, 1977). Studies that included tasks of medium difficulty in exploring the hard-easy effect observe fairly accurate confidence judgments on the performance on these items (Ilieva, Brudermann, & Drakulevski, 2018; Kriat, 2018; Moore & Healy, 2008).

I claim that the investigation of confidence judgments on questions of medium difficulty is important because it can answer the question of whether people are just overconfident on difficult questions due to the hard-easy effect or whether they are already overconfident on questions of medium difficulty. If incentivizing accurate judgments has an influence on people’s accuracy, this might mostly be observed on questions of medium difficulty. Further, the research aim of shedding light on processes of confidence judgments and specifically the confidence bias raises questions on whether processes of confidence judgments are the same or distinctive for tasks of differing difficulty. And also, questions as to whether the incentivization for performance or for accuracy of confidence judgments interacts with task difficulty on processes of confidence judgments.

### 2.3. Gender Effects on Confidence Judgments

As mentioned above, males are often rather overconfident in their own performance, while females tend to be, or even are frequently underconfident (see Barber & Odean, 2001; Hügelschäfer & Achtziger, 2014; Pulford & Colman, 1997). The seminal paper on this research was an article by Barber and Odean (2001). The authors compared the common stock investments of females and males and reported that males traded 45 percent more than females, although it reduced male traders’ net returns. Barber and Odean (2001) concluded that this behavior could be explained by overconfidence in own skills in males.

Huang and Kisgen (2013) compared financial and investment decisions made by female and male executives. Males made more acquisitions and issued debt more often than females. In addition, male executives’ returns were lower than those of females and debt issues of firms with male executives had lower announcement returns than those with female executives. This behavior reflects males’ relative overconfidence in corporate decision making compared to females. Males’ overconfidence is also used as an explanation of why females are chosen less often as group leaders than males (Reuben, Rey-Biel,
Balafoutas, Kerschbamer, and Sutter (2012) showed that males were consistently overconfident in their self-ranking of performance in a math task. In contrast, very well performing females (with a rank of 1 or 2 out of 6) were strongly underconfident in their task performance as indicated by their low self-rankings. In an experiment by Reuben, Rey-Biel, Sapienza, and Zingales (2012), groups selected a leader to compete against the leaders of other groups. Females were chosen less often which could not be explained by past performance. The authors found that the gender difference in overconfidence concerning past performance was the reason behind the prevalence of male representatives. Note that besides the vast literature on gender differences in confidence judgments and confidence biases, there are also studies that did not observe a gender effect in this domain (e.g., Acker & Duck, 2008; Deaves, Lüders, & Schröder, 2010; Mannes & Moore, 2013; Moore & Swift, 2011). Moreover, there is not much knowledge about the processes underlying gender differences in confidence judgments and biases, in particular not under the incentivization of performance or of accurate confidence judgments. In the present work, an almost equal number of females and males was recruited for each between condition of the experiments in order to shed some light on these questions.

2.4. Processes of Confidence Judgments

In psychology and economics, the investigation of judgment and decision processes often builds on dual process theories (Alós-Ferrer & Strack, 2014; Evans & Stanovich, 2013; Rustichini, 2008; Weber & Johnson, 2009). These theories assume that there are two broad types of processes. Controlled processes that are relatively slow, conscious, and effortful, meaning that they consume cognitive resources (Schneider & Shiffrin, 1977). Automatic processes are typically defined by opposing properties. They are characterized as fast, unconscious, and effortless, and they are often instigated immediately (see Schneider & Shiffrin, 1977). The simplest method to distinguish these two types of processes is the measurement of response times (Achtziger & Alós-Ferrer, 2014). Automatic processes should result in quicker responses than controlled processes (Schneider & Shiffrin, 1977).

Previous research on confidence judgments assumes that metacognitive judgments are inferential and rely on different beliefs and heuristics (Koriat et al., 2006). There are two types of judgments to be distinguished: information-based judgments and experience-based judgments (see Koriat, 1997; Koriat et al., 2006). Information-based judgments rely on an analytic inference in which several considerations retrieved from memory are
taken into account to make a deliberated judgment (Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980). On the other hand, experience-based judgments build on the information from mnemonic cues (for example processing fluency, accessibility, ease of retrieval) that derive online from task performance (see Koriat, 2000). Koriat et al. (2006) demonstrated that response latency is a mnemonic cue for confidence in an answer meaning that quickly answered questions were rated with higher confidence. Some researchers integrated the processes underlying the two types of judgments within a dual-process model (see Epstein & Pacini, 1999; Evans, 2008; Koriat & Levy-Sadot, 1999; Strack, 1992).

The question of what type of information confidence judgments are based on is highly discussed in the field of confidence models. Single stage models assume that confidence judgments are based on the same evidence as the response to the task itself (Koriat, 2012a; Galvin, Podd, Drga, & Whitmore, 2003; Higham, Perfect, & Bruno, 2009; Vickers & Lee, 1998). In contrast, dual stage models propose that confidence judgments result from a separate, post-decisional stage of evidence accumulation (for an overview see Moran et al., 2015). The claim of dual stage models is supported by findings that the response time for the confidence judgments predicts the confidence level (Hilgenstock, Weiss, & Witte, 2014; Petrusic & Baranski, 2003; Siedlecka et al., 2019; Yu, Pleskac, & Zeigenfuse, 2015). This result indicates that confidence is further processed after the initial decision on the task. Charles and Yeung (2019) found that confidence judgments are informed by evidence before and after the decision on the task which suggests that there is a second stage of evidence accumulation after the initial decision. Siedlecka et al. (2019) also found that if evidence is accumulated before and after the initial choice on a task, it improves metacognitive accuracy.

In all three studies of the present research, I measured the response times on the answers in a general knowledge test and response times on confidence judgments. I was interested whether the processes indicated by these response times and their relationship with confidence judgments (see Koriat et al., 2006) would be different for females and males, and if they were affected by incentives for performance and for the accuracy of confidence judgments. I expected that the investigation of response times would provide new insights into the nature of processes of confidence judgments, their suggestibility by incentives and their dependence on gender.
2.5. The Effects of Incentives on Confidence Judgments

The impact of incentives on performance and on the accuracy of confidence judgments is a research topic both in psychology and economics. The main question asked in both fields is whether incentives on the accuracy of confidence judgments promote accurate confidence judgments (e.g. Caplan, Mortenson, & Lester, 2018; Coutts, 2019; Lebreton et al., 2018).

Camerer and Lovallo (1999) observed that participants in an experimental market entry game overestimated their chances of success when the payoff depended on their abilities. In a study by Kirchler and Maciejovsky (2002) participants played an experimental asset market game. It turned out that they did not display overconfident judgments in predicting the average trading prices, but they demonstrated overprecision in their confidence intervals of those trading prices, suggesting that the participants were overly sure. Participants in a study by Michailova and Schmidt (2016) also traded in an experimental asset market. The markets were constructed based on the participant’s overconfidence in a general knowledge test. The prices in low overconfidence markets tended to track the fundamental asset value more accurately, were significantly lower and less volatile than in high overconfidence markets. In addition, the bubble measures and trading volume were significantly higher in high overconfidence markets. Pikulinaa, Renneboog and Tobler (2017) found that strong overconfidence in financial knowledge resulted in excess investment, whereas underconfidence was linked to underinvestment. Participants rewards depended on their investment level and their skill level, meaning that reward was maximized when investment decisions were congruent with their actual skill level.

Krawczyk (2012) discovered that incentives on the prediction accuracy of performance in an exam increased the confidence judgments’ accuracy. Participants were asked to forecast whether they would do better or worse than an average student on an upcoming exam. Participants reported higher levels of confidence when being incentivized on the accuracy of their judgment compared to no incentive. Males were more confident in their performance in the exam than females when no incentive was offered but not when being incentivized for accurate predictions. This indicated that incentives could reduce gender differences in over-/underconfidence in one’s performance.

Yet, there is also evidence for the opposite effect of incentives. In research by Hoelzl and Rustichini (2005) participants were less confident in their performance in a knowledge test if money was at stake. Participants in an incentive condition could win money
depending on their confidence judgment whereas control participants were only asked to imagine that they could win the same amount of money but were paid flat rate (i.e., their judgment did not determine their payment). In addition, Hoelzl and Rusticchini (2005) manipulated task difficulty by presenting easy or difficult questions in a knowledge of vocabulary test. Prior to the test, the participants voted if winning money should depend on their performance (e.g., performing like the upper half of all participants would be awarded with 150 Austrian Schilling worth approximately 10 USD or on the toss of a die, with a 50% chance of winning the same amount of money). Findings implied that the vote shifted from the incentive on test performance to the toss of a die when the test consisted of difficult questions compared to a test including only easy questions. This observation indicated lower confidence levels in the performance on difficult questions in participants who were incentivized compared to participants who only imagined being incentivized. Thus, people were less confident in their performance on difficult questions when the incentive depended on this performance. The incentive had no influence on the confidence in the performance on easy questions.

The idea that monetary incentives on the accuracy of confidence judgments can reduce overconfidence was also confirmed by Blavatskyy (2009). Participants even displayed underconfidence in their performance when their experimental payoff depended on their individual performance in a general knowledge test. Similarly, Clark, and Friesen (2008) observed underconfidence in forecasts of performance when participants were incentivized both on performance and on the accuracy of their predictions. In this study participants were asked to complete two unfamiliar tasks (e.g. maximizing a function, decode five-letter words). Underconfidence prevailed in this study.

Caplan, Mortenson, and Lester (2018) observed that incentives on accurate forecasts mitigated the confidence bias in forecasts of exam performance. The incentive reduced overconfidence in poorly performing students and underconfidence in top-performing students. In a study by Koriat (2006), the performance on two different tasks was incentivized. Participants additionally indicated their confidence, but the incentive depended only on performance. The results showed that a high incentive (compared to a low incentive) increased confidence in a figural matrices and figural series task (Experiment 7). It is important to mention that the incentive had no influence on performance. The participants displayed overconfidence (higher confidence levels than accuracy) in both tasks. Comparing this result to the previously mentioned findings on incentives for accuracy indicates that paying an incentive for performance has a different effect on confidence judgments than paying for accuracy.
Sanchez and Dunning (2018) found similar results on the influence of incentives on confidence judgments. The participants had to diagnose patients on the basis of patient profiles (set of symptoms). The patients had either one of two illnesses or were not sick at all. After the diagnosis, participants indicated their confidence. In Study 1, participants received an incentive for accuracy by achieving an accuracy level of at least 80%. The results showed that participants were overconfident in the accuracy of their diagnosis and this overconfidence increased with experience (trial number). In Study 3, an incentive was provided for accuracy as well as a valid expression of confidence. Participants indicated their confidence in their diagnosis by betting on its accuracy instead of a lottery to win the incentive. The results showed that the participants were still overconfident in the accuracy of their diagnosis. The incentivized confidence judgment seemed to have no influence on the participants’ overconfidence bias. Lebreton et al. (2018) investigated the effects of incentives on confidence judgments on binary decisions in a visual task. Participants judged the probability that their answer (which of two Gabor patches has the highest contrast) was correct (ranging from 50 to 100%). Note that the difficulty of the task was adapted according to the performance of a calibration session. Hence there are no results on item difficulty. After each decision on the contrast of the Gabor patches, a lottery number between 50 and 100% was drawn randomly by the computer. If the confidence in the correctness of the answer was rated higher than the lottery number, participants were awarded an incentive if their decision in the perceptual task was correct. If the confidence was smaller than the lottery number, then the lottery was implemented, and participants received the incentive with the probability of the lottery number. The authors varied the magnitude (nothing at stake, 10 Cent, 1€, 2€) and the valence of the incentives (winning or losing). In experiment 1, they varied the valence of the incentive (gain or no gain of 1€ vs. loss or no loss of 1€) and compared it to a neutral outcome (nothing at stake). The results showed that the incentive (gain and loss) increased the sensitivity meaning the efficacy of discriminating between correct and incorrect answers. The valence of the incentive had an influence on the cognitive bias. The gain of an incentive increased overconfidence compared to the neutral outcome whereas the loss of an incentive decreased overconfidence compared to the neutral outcome. In experiment 2, the authors demonstrated that an incentive for performance (not accuracy of confidence judgments) did not influence the confidence bias. Experiment 3 (10 Cents vs. 1€) and 4 (10 Cents vs. 1€ vs. 2€), confirmed that not only the valence of the incentive influenced the confidence bias but also the magnitude of the incentive. Higher incentives (10 Cent vs. 1€ vs. 2€) increased the confidence bias in a gain frame and decreased the bias in a
loss frame.

Lebreton et al. (2018) also reported response times on the choice and the confidence judgments (supplementary materials). The results showed that the incentives in all 4 experiments increased the response time on the confidence judgments. The authors explained it by the participants’ increased care when money was at stake. The incentives had no influence on the response times of the choice in the task. This is somewhat surprising since an incentive was explicitly paid only for performance in experiment 2B. Note that Lebreton et al. (2018) did not examine gender differences and interactions between incentives and gender on confidence judgments and their underlying processes in their research. They also did not explore the potential relation between response times and the confidence bias. Further, the authors could not comment on the influence of task difficulty on confidence judgments due to their task design.

Note that most of these studies (except for Krawczyk, 2012) did not test for gender effects or interactions between gender and incentives on confidence judgments. I addressed this gap in the literature and tested for gender differences and effects of incentives on performance and confidence judgments and their processes. Further, some of the studies mentioned above had an overrepresentation of male (e.g. Sanchez & Dunning, 2018; 60% males) or female participants (Hoelzl & Rustichini, 2005: 66% females; Lebreton et al., 2018: 74% females) in their samples. Therefore, their findings might not be generalizable to all genders.
3. Present Research

The research program was the following. First, I pretested knowledge questions taken from an intelligence test (Intelligenz Struktur Test; I-S-T 2000 R; Amthauer, Brocke, Liepmann, & Beauducel, 2001; standardized for the German population) regarding their difficulty (i.e., their likelihood of being correctly answered) and potential gender biases. The questions belonged to the general knowledge subsample of items of the IQ test. Based on this pretest, I selected questions that did not have a gender bias (i.e., females and males performed alike on these items) and categorized them into easy, medium, and difficult questions (see below). All questions that were selected based on this pretest were used as experimental material in the Studies 1 to 3 of the present research.

Across three experiments, I explored processes of confidence judgments on performance and the respective confidence bias. Participants were asked to answer general knowledge questions by choosing one out of five options (see below). I investigated whether the confidence judgments and the confidence bias (that could result in over- and underconfidence; see Koriat, 2012a; Pikulina et al., 2017; Lebreton et al., 2018; Koriat, 2018) relied on rather automatic or on controlled processes by measuring response times on the answers to the knowledge questions and on the confidence judgments. Furthermore, I examined whether the confidence judgments were influenced by information that was accumulated after the initial choice (based on dual stage models of confidence, see Moran et al., 2015). I also examined whether these processes of confidence judgments varied with the difficulty of the knowledge questions (Studies 1 to 3), incentives for performance (Study 2), incentives for the accuracy of the confidence judgment (Study 3), and if they differed between females and males.

As dependent variables in all three studies, performance in the knowledge test (indicated by the percentage of correctly answered questions), the confidence judgments on these answers (measured by a continuous scale from 0 to 100 in Studies 1 and 2, and by the random binary choice elicitation mechanism (Healy, 2016) in Study 3), the confidence bias (that relates performance to the confidence judgments and indicates under- and overconfidence, and accurate confidence judgments, respectively; see for example Koriat,
response times on the answers to the knowledge questions, and response times on confidence judgments were measured.

I expected the hard-easy effect (Brenner, 2003; Fischhoff et al., 1977; Grieco & Hogarth, 2009; Griffin & Tversky, 1992; Larrick et al., 2007; Lichtenstein et al., 1982; Merkle, 2009; Moore & Healy, 2008) in all three studies, i.e., independent of incentives and gender. Participants would be underconfident on easy, rather accurate on questions of medium difficulty, and overconfident on difficult questions. I claimed that response times on the answers on the knowledge questions would reflect the item difficulty (see Wixted & Mickes, 2010). More specifically, I predicted for all three studies that the answers would be quicker on easy than on medium difficult, and on difficult questions. Consequently, response times on answers should be quicker on medium difficult as on difficult questions. The pattern of response times on the confidence judgments is more complex. Based on Koriat’s (2012a) assumption that confidence judgments rely only on the online feedback of the process of choosing an answer to the knowledge question (response speed), then the response time on the confidence judgments should not vary depending on the item difficulty (in agreement with single stage models of confidence). The response speed would always inform the confidence judgment in the same way independent of its duration. However, there is also evidence for the assumption of dual stage models of confidence judgments (see Moran et al., 2015) that confidence judgments result from a pre- and postdecisional stage. Research findings showing that the response time on the confidence judgments predicts the confidence level (Hilgenstock et al., 2014; Petrusic & Baranski, 2003; Siedlecka et al., 2019; Yu et al., 2015) indicate that confidence judgments are processed after the initial decision and go beyond simple mnemonic cues. Hence, the pattern of response times on the confidence judgments remained an open question.

I expected participants to have a good monitoring resolution meaning that they are capable to discriminate between correct and incorrect answers (Koriat, 2012a, 2018; Lebreton et al., 2018). I predicted that the confidence judgments should be higher for correct than for incorrect answers. Consequently, the response time for the answers to the knowledge questions should be quicker for correct than for incorrect answers. Based on the findings of Lebreton et al. (2018), I expected an increase in monitoring resolution when an incentive was payed for for accuracy of the confidence judgments (Study 3) but not when an incentive was payed for performance (Study 2).

Due to studies that reported gender differences on confidence judgments (see above), I also explored whether the default response in confidence judgments was different
for females and males. This question was investigated in Study 1 under conditions of flat rate payment, that is no incentives were offered for performance or for the accuracy of the confidence judgment. I conducted this first experiment without providing incentives in order to explore whether a gender bias would emerge when no money was at stake for the participants. Hence, all participants earned the same amount of money in exchange for participation. From a process perspective, I was interested whether potential overconfidence in males relied on rather quick (i.e., automatic) or rather slow (i.e., controlled) processes. Likewise, I explored whether females in case of being underconfident would rely on rather automatic processes in their confidence judgment and hence would make confidence judgments rather quickly. In case that males’ overconfidence and females’ underconfidence would be observed and would rely on quicker processes than accurate confidence judgments in both genders, one could argue that rather automatic processes (experience-based judgments; see (Koriat, 2012b)) are the default in gender effects on confidence biases.

Another research aim was the investigation of whether incentivizing performance in the knowledge test (Study 2) or incentivizing the accuracy of confidence judgments on performance in the knowledge test (Study 3) impacts (e.g., reduces) the confidence bias. Most important, I was interested in the influence of incentives in Study 2 and 3 on processes of answering the knowledge questions (i.e., response times on answers) and on processes of confidence judgments (response times on confidence judgments). Potential gender effects were also considered in these analyses. The question was whether incentives on performance and on confidence judgments would have the same impact on females and males.

For Study 2, I claimed that an incentive on performance improves the performance in the general knowledge test compared to a flat rate payment condition based on the natural economic hypothesis that incentives increase the attention in case of a decision conflict (note that the findings of Koriat et al. (2006) and Lebretion et al. (2018) did not support this hypothesis, but see Camerer and Hogarth (1999) for a more detailed discussion on the relationship between incentives and performance). I predicted prolonged response times on the answers to the knowledge questions in the incentive condition compared to the flat rate payment condition. I claimed that the incentive on performance supports controlled processes, i.e., thinking carefully about the answers on knowledge questions, and therefore increases response times on the answers. For confidence judgments and their processes, I did not expect prolonged response times in the incentive condition compared to the flat rate payment condition because the incentive depended only on
the performance on the knowledge questions and was independent of the confidence judgments. Regarding gender, the incentive on performance was expected to have the same effect on performance in both females and males, namely an increase of performance in the knowledge test. However, whether the incentive for performance would affect the (potential) gender difference on confidence judgments and the confidence bias (i.e., males’ over-, and females’ underconfidence) was an open question.

In Study 3, I offered a low vs. a high monetary incentive for accurate confidence judgments. I claimed more accurate confidence judgments in the high incentive condition compared to the low incentive condition as the incentive was linked to the confidence judgments directly and therefore should increase their accuracy. I expected longer response times on confidence judgments in the high incentive compared to the low incentive condition. The high incentive would spark off more time and higher effort investments in the confidence judgments as the low incentive and whereby decelerate confidence judgments. Charles and Yeung (2019) and Siedlecka et al. (2019) found that if information is accumulation before and after the initial choice metacognitive accuracy improves (see also Moran et al., 2015). The incentive was expected to prolong the response time on the confidence judgments and hence the information accumulation after the initial choice. Contrary to Study 2, I did not expect participants to take more time for answering the knowledge questions in the high compared to the low incentive condition in Study 3. The reason was that in both conditions (high vs. low) the incentive was offered for accurate confidence judgment and not for the performance in the knowledge test, hence processes of answering the knowledge questions should not be affected by the incentive. Regarding gender differences on confidence judgments and the confidence bias, the high incentive was expected to increase the accuracy of confidence judgments compared to the low incentive in both genders and hence was claimed to prevent a difference between genders in confidence judgments and the confidence bias.
4. Pretest on General Knowledge Questions

The aim of the pretest was to generate suitable materials to investigate confidence judgments and more specifically possible gender differences. General knowledge questions are commonly used in research on confidence judgments or the confidence bias. However, general knowledge questions are sensitive to gender effects (Larrick et al., 2007; Soll & Klayman, 2004) and item difficulty (see Section 2.2 on the hard-easy effect). Consequently, it has been criticized that overconfidence has only been found because the research materials were not tested in advance (Moore & Healy, 2008; Juslin, Winman, & Olsson, 2000). In order to clearly attribute any differences in confidence judgments to the effect of incentives or gender, I conducted a pretest to select gender-neutral knowledge questions which were categorized in three levels of difficulty (easy, medium, difficult).

4.1. Methods

4.1.1. Participants

71 participants (40 female, age range: 18-39; $M = 24.71; SD = 3.75$) were recruited via student groups in social networks for an online study. Ten 20 € amazon gift cards were raffled among all participants in exchange for participation.

4.1.2. Material

84 general knowledge questions were taken from the I-S-T 2000 R (Amthauer et al., 2001), a standardized German intelligence test. All questions were multiple choice questions with five options to choose from. One example was: ‘Which country does Greenland belong to? (A)Canada, (B)Iceland, (C)Russia, (D)USA, (E)Denmark.’ An internet-based questionnaire was created using Sosci-Survey (Leiner, 2019). To detect deceiving while answering the questions (i.e., searching in the internet for the correct answer), switches
between browser tabs were registered. This was necessary because the pretest was run online and hence I could not control if participants tried to deceive in order to improve their performance. The 84 questions were randomized across participants to avoid fatigue and sequence effects. In addition, demographic data (age, gender, academic year, and field of study) were collected.

4.1.3. Procedure

The link to the online questionnaire was posted in several groups in social networks frequently visited by students. Participants were instructed to answer the 84 questions to the best of their knowledge. They were told that in case of not knowing the answer to a question, they should guess the correct answer, since a response was required to continue with the questionnaire; otherwise, they would drop out. Participants were informed that only one out of five options was correct. It was emphasized that they should refrain from looking up answers on the Internet.

4.1.4. Results

To ensure an unbiased test of item difficulty, answers given after a tab switch were excluded from the analysis. This was the case in 164 out of 5964 answers, meaning that participants were quite honest. The answers were dummy coded, correct answers were coded 1, whereas false answers were coded 0. The mean score for each question over all participants could range from 0.0 to 1.0. Accordingly, the mean score of each question (solving rate of an question) represented the average likelihood of the correct answer, indicating the difficulty of an question. The aim of this pretest was to identify general knowledge questions on which the performance of females and males did not differ. In order to determine which questions met this criterion of gender equality, I ran a two-sample \( t \)-test for females and males for each of the 84 questions. 21 questions were excluded due to a significant difference between mean scores of correct answers (and hence item difficulty) between females and males, \(|t|s \leq 0.81, ps \geq .040\).\(^1\)

The remaining 63 questions were categorized as easy, medium, and difficult questions. Note that the likelihood of a correct answer based on mere guessing was 0.2 for each question. Questions with a mean solution rate between 0.0 and 0.5 were categorized as difficult questions. Questions with a mean solution rate between 0.5 and 0.75 were categorized as questions of medium difficulty and questions with a mean solution rate

\(^1\)the selection of questions was additionally confirmed by a test of proportions (\(z_s \leq 1.55, ps \geq .121\)
between 0.75 and 1.0 were categorized as easy questions since nearly everybody could solve them correctly (see Appendix A.1).

4.1.5. Discussion

Most questions of the pretest were not gender biased (see Appendix A.1). I excluded those questions that were gender biased from the item pool for Study 1 to 3. Three categories of item difficulty were defined based on the likelihood that an question of the general knowledge test was correctly solved. The sub-sample of easy questions (e.g., “What is the meaning of the following symbol? $\Sigma$”), questions of medium difficulty (e.g., “In which year did the Titanic sink?”), and questions of high difficulty (e.g., “How many hydrogen atoms contains alcohol (ethanol)?”) included 10 questions each. The selected questions for Study 1 to 3 are listed in Appendix A.2.
5. Study 1: Confidence Judgments and Their Processes Under Flat Rate Payment

The aim of this first study was to examine processes of confidence judgments on performance without potential interferences with monetary incentives. Participants answered general knowledge questions (selected based on the pretest; see above) and judged their confidence on the correctness of an answer on each of these questions on a rating scale (see Lichtenstein & Fischhoff, 1977; Koriat, 2018; Koriat et al., 2006; Pikulina et al., 2017). I measured the response times on the answers to the knowledge questions as well as the response times on the confidence judgments. All participants received a flat rate payment, meaning I did neither offer an incentive for performance in the knowledge questions or for the accuracy of the confidence judgments. First, I expected that in accordance with the categorization of knowledge questions as easy, medium difficult, and difficult questions (based on the pretest), participants would perform best (i.e., answer most of these questions correctly) on easy questions, followed by their performance on questions categorized as medium difficult, and finally would perform worst on difficult questions. Performance was operationalized as the percentage of correctly answered questions.

Second, I suggested that the item difficulty should also be reflected in the response times on the answers. Answers should be given quicker on easy questions than on questions of medium difficulty and the latter should be answered more quickly than difficult questions. This pattern of response times on answers in the knowledge test validates the categorization of the questions into easy, medium, and difficult questions in addition to the performance on a question (see above). Further, these response times on the questions should be used as a mnemonic cue for the confidence judgments (see Koriat, 2012b, 2012a). The longer the response time on the answer to the question, the lower the confidence in the correctness of that answer (see also Siedlecka et al., 2017).
This means, that the confidence judgments should be highest on easy questions, followed by medium questions and lowest on difficult questions. As discussed above (see Section 3 on the present research), the pattern of the response times on the confidence judgments remained an open question to be answered. Next, I claimed the hard-easy effect (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982). All participants were proposed to be underconfident on their performance on easy questions, rather accurate in their confidence judgment on performance on medium difficult questions, and overconfident regarding their performance on difficult questions.

I also expected that confidence judgments would reflect good monitoring resolution (i.e., good discrimination between correct and incorrect answers) over all difficulty levels and independent of gender (see Koriat et al., 2006; Koriat, 2012a; Lebreton et al., 2018). Further, individuals’ sensitivity or meta-cognition regarding the correctness of their answer (i.e., they consider ‘evidence’ before they judge their confidence; see Lebreton et al., 2018) should be reflected in the response times on the answers, meaning they should give correct answers quicker than incorrect answers. The question whether the response time on the confidence judgments should reflect the same pattern remained open (see discussion on single stage vs. dual stage models of confidence).

Regarding the confidence bias, males were expected to be overconfident over all questions, while females were predicted to be underconfident over all questions. In terms of response times on confidence judgments, I expected that biased participants (overconfident males and underconfident females) judge their confidence quicker than unbiased individuals (see Yu et al., 2015). The latter were defined as participants of both genders who judged their confidence rather accurately. These predictions build on dual process theories that suggest the dominance of automatic processes (see Alós-Ferrer & Strack, 2014) in case of a default mode of a confidence judgment (e.g., overconfidence). Note that Siedlecka et al. (2019) found a negative relationship between the response time on the confidence judgments and the metacognitive accuracy (the longer participants took for their confidence judgment, the lower the metacognitive accuracy). However, their sample consisted of significantly more females than males and hence is not a representative sample.

To conclude, the hypotheses of Study 1 were the following:

**Hypothesis 1a:** The item difficulty is reflected in the performance on the knowledge questions and the confidence judgments (easy > medium > difficult).
Hypothesis 1b: The item difficulty is reflected in the response time on the answers (easy < medium < difficult).

Hypothesis 2: The response time on the answers serves as mnemonic cue for the confidence judgment which is shown by a negative correlation between the two variables.

Hypothesis 3: The hard-easy effect is expected.

Hypothesis 4a: Participants are expected to have a good monitoring resolution. Confidence judgments on correct answers should be higher than on incorrect answers.

Hypothesis 4b: The response time on correct answers should be quicker than on incorrect answers.

Hypothesis 5: Males are expected to display overconfidence whereas females are expected to be underconfident.

Hypothesis 6: The response time on the confidence judgments is negatively related to the confidence bias.

5.1. Methods

5.1.1. Participants

123 participants (56 females, 65 males, 2 without specification) were recruited among the student population of a university (Zeppelin University, Friedrichshafen). They were between 18 and 33 years old, \( M = 22.15 \) (\( SD = 2.54 \)). In exchange for participation, they received 10 € (i.e., flat rate payment). All data were included in the analysis. The data of two participants without gender specification were excluded from analysis of gender effects but were included in all analyses that did not test for gender effects.

5.1.2. Design

The study followed a 2 (gender, between: female vs. male) \( \times \) 3 (item difficulty, within: easy vs. medium vs. difficult) mixed model design. Dependent variables were percentage
of correct answers (performance), confidence judgments, the confidence bias, response
times on the confidence judgments, and response times on answers on the knowledge
questions.

5.1.3. Materials

Knowledge Items & Confidence Judgments

30 questions of the pretest were presented (see Appendix A.2), 10 questions per each
difficulty level in random order to avoid sequence effects (randomized for each participant).
The criteria for assigning each questions to one of the three categories of difficulty were
the same as in the pretest (see above). Three more questions were chosen as example
questions (i.e., they did not show up in the main study), presented at the very beginning
of the study as practice trials, for each item difficulty. The participants rated their
confidence on the correctness of their answer on a continuous scale ranging from 0 to
100% using a scroll bar. The rating followed right after an answer was given by choosing
an option of a multiple-choice test through a mouse click.

Response Times on Knowledge Questions

Each knowledge question and the corresponding confidence rating scales were presented
on the same screen. The response time on the answer to the knowledge question was
measured starting from the presentation of the screen until the participant chose an
answer to the knowledge question by mouse click.

Response Times on Confidence Judgment

The response time on the confidence judgment was measured beginning when the
participant chose an answer to the knowledge question by a mouse click until the
participant indicated their confidence judgment on the correctness of this answer by
mouse click.

5.1.4. Procedure

The study was conducted in group sessions of maximum 10 students using SoSci Survey
(Leiner, 2019) in a laboratory. A session lasted about 45 minutes. At the beginning
of each session, participants solved 3 example questions on the computer. Then, they
began answering the knowledge questions and judging their confidence on the correctness
of these answers in the computer. The response times on the knowledge questions and
the response times on the confidence judgments were measured separately (see above).
Subsequently, participants filled in a computerized questionnaire\textsuperscript{2}. Afterwards they were
thanked, debriefed, and paid for participation.

5.2. Results

5.2.1. Number of Correctly Answered Knowledge Questions

The percentage of correctly answered knowledge questions was calculated for each
participant and then compared across gender and item difficulty. A mixed model
ANOVA with the between-subjects factor gender and the within-subjects factor item
difficulty (easy vs. medium vs. difficult). It showed the expected effect of item difficulty,
\( F(2, 238) = 436.01, p < .001, \eta^2_p = .79 \), and replicated the pretest. On average,
participants answered \( M = 87.48\% \) of the easy questions (\( SD = 11.78\% \)), \( M = 58.70\% \)
of the medium questions (\( SD = 17.74\% \)) and \( M = 36.18\% \) of the difficult questions
(\( SD = 16.67\% \)) that I assigned to these difficulty levels based on the pretest. Subsequent
t-tests showed that participants answered significantly more easy than medium questions
correctly, \( t(122) = 16.64, p < .001 \), and significantly more medium than difficult questions
correctly, \( t(122) = 12.60, p < .001 \). There was no gender effect , \( F(1, 119) = .36, p = .551 \),
and no significant interaction between gender and item difficulty, \( F(238, 2) = 1.30, p = .275 \).

5.2.2. Response Times on the Answers to the Knowledge Questions

I logarithmized the response times since they were skewed (which is normal for response
times). I analyzed the response time on the answers in the knowledge question across
gender and item difficulty. A mixed model ANOVA with the between-subjects factor
gender and the within-subjects factor item difficulty yielded a significant effect of item
difficulty, \( F(2, 238) = 172.44, p < .001, \eta^2_p = .59 \). All participants answered easy
questions in \( M = 7255.64 \text{ ms} (SD = 2109.66 \text{ ms}) \), medium questions in \( M = 10091.09 \text{ ms} (SD = 3100.18 \text{ ms}) \), and difficult questions in \( M = 11540.04 \text{ ms} (SD = 3768.87 \text{ ms}) \).

\textsuperscript{2}The questionnaire included the State Self-Control Scale (Ciarocco, Twenge, Muraven, & Tice, 2004),
the Strenuous Mental Activity Scale (Job, Dweck, & Walton, 2010), the Sex-Role Inventory (Berger
& Krahé, 2013), Faith in Intuition (Keller, Bohner, & Erb, 2000), Self-Esteem (Rosenberg, 1965)
and the Situational Motivation Scale (Guay, Vallerand, & Blanchard, 2000). These scales are not
further mentioned due to their null effects on all dependent variables.
Subsequent t-tests showed that easy questions were answered significantly quicker than medium questions, \( t(122) = 13.67, p < .001 \), and medium questions were responded to significantly quicker than difficult questions, \( t(122) = 5.30, p < .001 \). The analysis showed no gender effect, \( F(1, 119) = 1.26, p = .263 \). The interaction between gender and item difficulty did not reach significance, \( F(2, 238) = 2.26, p = .106 \). Hence, the experimental material did not have a gender bias on performance.

Next, I compared response times on correct answers with those on incorrect answers. A mixed model ANOVA with the between-subjects factor gender and the within-subjects factor correctness (correct vs. incorrect answers) yielded a significant effect of correctness, \( F(1, 119) = 205.72, p < .001, \eta_p^2 = .63 \) (see Figure 5.1). Participants gave correct answers quicker, \( M = 8762.17 \text{ ms} \) (\( SD = 2331.71 \text{ ms} \)) than incorrect answers, \( M = 12408.34 \text{ ms} \) (\( SD = 4045.75 \text{ ms} \)). The analysis revealed no gender effect, \( F(1, 119) = 2.08, p = .152 \), and no interaction between gender and correctness, \( F(1, 119) = .03, p = .863 \).

5.2.3. Confidence Judgments

The average confidence judgment on the confidence scale was calculated for each participant and compared across gender and item difficulty. In case of a violation of sphericity,
the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factor gender and the within-subjects factor item difficulty yielded a significant effect of item difficulty, $F(1.91, 227.18) = 486.99, p < .001, \eta^2_p = .80$. Participants judged their confidence on the correctness of their answers on easy questions with $M = 81.46$ ($SD = 12.28$), on medium questions with $M = 56.83$ ($SD = 15.28$), and on difficult questions with $M = 44.27$ ($SD = 14.87$). Subsequent $t$-tests showed that the confidence judgment on easy questions was significantly higher than on medium questions, $t(122) = 19.59, p < .001$, and the confidence judgments on medium questions were significantly higher than on difficult questions, $t(122) = 9.88, p < .001$. There were no gender effect, $F(1, 119) = 2.14, p = .146$, and no significant interaction between gender and item difficulty on confidence judgments, $F(1.91, 227.18) = 1.28, p = .253$. Note that this variable indicated absolute confidence judgments on the confidence scales, not the confidence bias (i.e., under- or overconfidence) because the latter requires accounting for both, performance in the knowledge test and the confidence judgments (see Hügelschäfer & Achtziger, 2014; Koriat et al., 2006; Lebreton et al., 2018; Pikulina et al., 2017).

Following Koriat (2012a) I calculated the correlation between confidence judgments and the response time on the answer to the knowledge questions. The assumption was that if this correlation was negative, the confidence judgments were experience-based and relied on mnemonic cues (in this case response time on the answers). The correlation was negative but only weak, $r = -.15, p = .089$. Due to the gender perspective, I calculated the correlation separately for females and males. The correlation for females was negative and marginally significant, $r = -.24, p = .076$. The correlation for males was also negative but missed significance, $r = -.12, p = .336$. Hence, only for females there was a tendency that longer response times on the answers to the knowledge questions were related to lower confidence levels, but not for males. This suggests that females rely more on their experience with the task when judging their confidence in their performance than males.

Following Lebreton et al. (2018) with the argument that confidence is built on (noisy) evidence, I scrutinized the confidence judgments for correctly and incorrectly answered questions separately. A mixed model ANOVA with the between-subjects factor gender and the within-subjects factor answers (correct vs. incorrect) showed a significant main effect of the correctness of the answers on the confidence judgments, $F(1, 119) = 753.78, p < .001, \eta^2_p = .86$ (see Figure 5.2). Not surprisingly, the confidence judgments for correctly answered questions $M = 87.50$ ($SD = 14.99$) were significantly higher than for incorrect answers $M = 73.45$ ($SD = 11.90$). This result validated the experimental
material. No gender effect emerged, \( F(1, 119) = 1.65, p = .201 \), and no interaction between gender and the correctness of an answer, \( F(1, 119) = 1.78, p = .185 \).

Figure 5.2.: Study 1: Confidence judgments by correctness

![Confidence judgments by correctness](image.png)

### 5.2.4. Response Times on the Confidence Judgments

I also analyzed the response time on the confidence judgment (also logarithmized). In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subject factor gender, and the within-subjects factor item difficulty yielded a significant effect of item difficulty, \( F(1.88, 222.24) = 163.64, p < .001, \eta^2_p = .58 \). Participants judged their confidence on easy questions in \( M = 9358.91 \) ms \((SD = 2454.19\) ms), on medium questions in \( M = 12036.27 \) ms \((SD = 3466.21\) ms), and on difficult questions in \( M = 13745.30 \) ms \((SD = 4263.29\) ms). Participants judged their confidence on easy questions significantly quicker than on medium questions, \( t(122) = 13.20, p < .001 \), and on medium questions significantly quicker than on difficult questions, \( t(122) = 6.23, p < .001 \). These findings additionally supported the validity of the experimental material. The analysis showed no gender effect, \( F(1, 119) = 1.59, p = .210 \), and no interaction between gender and item difficulty, \( F(1.87, 222.24) = 2.28, p = .108 \).
Next, I compared response times on confidence judgments for correct and incorrect answers. A mixed model ANOVA with the between-subjects factor gender and the within-subjects factor correctness (correct vs. incorrect answers) yielded a significant effect of correctness, $F(1, 119) = 75.87, p < .001, \eta^2_p = .39$ (see Figure 5.3). Participants judged their confidence level for correct answers quicker, $M = 10899.48$ ms ($SD = 2736.94$ ms) than incorrect answers, $M = 13221.03$ ms ($SD = 4308.32$ ms). No gender effect, $F(1, 119) = 2.22, p = .139$, and no interaction between gender and correctness appeared, $F(1, 119) = .12, p = .736$.

Then I calculated the correlation between the response time on the answers to the knowledge questions and the response times on the confidence judgments. The correlation was highly significant and positive, $r = .94, p < .001$. This illustrates the strong link between the response time on the answers and the response time on the confidence judgments.

![Figure 5.3.: Study 1: Response time on confidence judgments by correctness](image)

**5.2.5. Confidence Bias**

The confidence bias was computed as follows: The mean of confidence judgments was calculated to obtain an average measure of confidence in the correctness of their
answers in the knowledge test for each participant. Then, the percentage of correctly answered questions was subtracted from this average of confidence judgments to get a measure of the confidence bias (i.e., of over-/underconfidence; Fischhoff et al., 1977; Hügelschäfer & Achtziger, 2014; Lebreton et al., 2018; Pikulina et al., 2017). Positive scores of the confidence bias reflect overconfidence (i.e., overestimating the number of correctly answered questions), whereas negative scores indicate underconfidence (i.e., underestimating the number of correctly answered questions). A score of zero indicates an accurate, unbiased estimation of one’s performance in the knowledge test.

First, I tested the predictions on the hard-easy effect (Lichtenstein & Fischhoff, 1977). A mixed model ANOVA with the between-subjects factors gender and the within-subjects factor item difficulty (easy vs. medium vs. difficult) showed the expected main effect of item difficulty, \( F(2, 238) = 33.20, p < .001, \eta^2_p = .22 \) (see Figure 5.4). Participants were underconfident on easy questions, \( M = -6.02 (SD = 11.90) \), rather accurate in their confidence judgment on questions of medium difficulty, \( M = -1.87 (SD = 18.11) \) and clearly overconfident on difficult questions, \( M = 8.09 (SD = 19.41) \). T-tests confirmed that the mean confidence bias for easy, \( t(122) = 5.61, p < .001 \), and difficult questions, \( t(122) = 4.62, p < .001 \), differed significantly from zero, whereas the mean confidence bias for medium questions did not differ significantly from zero, \( t(122) = 1.15, p = .254 \). Subsequent t-tests confirmed that the confidence bias on easy questions differed significantly from the confidence bias on medium questions, \( t(122) = 2.44, p = .016 \), and the confidence bias on medium questions differed significantly from the confidence bias on difficult questions, \( t(122) = 5.27, p < .001 \). That perfectly underlined the predicted hard-easy effect (see Brenner, 2003; Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982).

There was no gender effect, \( F(1, 119) = .73, p = .394 \), and no interaction between gender and item difficulty, \( F(2, 238) = .14, p = .869 \). T-tests confirmed that males’ confidence bias \( M = .85 (SD = 13.32) \) did not differ from zero, \( t(64) = .52, p = .607 \). The same was true for females’ confidence bias \( M = -1.08 (SD = 11.30) \), \( t(56) = -.72, p = .476 \). These findings suggested that males were not overconfident in their performance and females were not underconfident. Instead, over- and underconfidence exclusively depended on the difficulty of the knowledge questions in both genders.
I computed some linear regressions in order to explore the confidence bias more thoroughly. To test whether gender affected the confidence bias while controlling for response times on confidence judgments (because I assumed that longer response times reflect more deliberation and attention and result in less biased judgments), I ran a linear regression on the confidence bias. Prior to the regression, all variables were z-standardized, except for the response times which were logarithmized since they were skewed. Gender (male = 1, female = 0) was dummy-coded. Regressions on the confidence bias are shown in Table 5.1. Model 1 accounted for gender. It showed that females and males did not differ on the confidence bias. Model 2 added response times on confidence judgments as a predictor. Response time on the confidence judgments negatively predicted the confidence bias (on a 10% significance level). Participants, who invested more time to judge their confidence on the correctness of their answers were less overconfident than individuals who took less time for their confidence judgments. Model 3 added the interaction between gender and response time on the confidence judgments as a predictor of the confidence bias. It did not reach significance. Interestingly, adding this interaction reduced the significance of the response times on confidence judgments as a predictor. Thus, the latter was not a very robust predictor of the confidence bias.
Table 5.1.: Study 1: Linear regressions on confidence bias

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male = 1)</td>
<td>.16</td>
<td>.20</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>(.18)</td>
<td>(.18)</td>
<td>(6.87)</td>
</tr>
<tr>
<td>Response time on confidence judgments</td>
<td>-.71*</td>
<td>-.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.37)</td>
<td>(.53)</td>
<td></td>
</tr>
<tr>
<td>Gender × Response time on confidence judgments</td>
<td>-.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-.09</td>
<td>6.53*</td>
<td>4.86</td>
</tr>
<tr>
<td></td>
<td>(.13)</td>
<td>(3.41)</td>
<td>(4.89)</td>
</tr>
<tr>
<td>N</td>
<td>121</td>
<td>121</td>
<td>121</td>
</tr>
<tr>
<td>R² adj.</td>
<td>&gt;.01</td>
<td>.02</td>
<td>.01</td>
</tr>
<tr>
<td>F</td>
<td>.73</td>
<td>2.26</td>
<td>1.58</td>
</tr>
</tbody>
</table>

Notes. Standard errors in brackets. *p < .10, **p < .05, ***p < .01.

Due to my interest in gender differences on processes of confidence, I computed correlations between response times on the confidence judgments and the confidence bias separately for females and males. This correlation was only significant for males $r = -.26$, $p = .037$ (females: $r = -.15$, $p = .289$). To interpret this finding, participants were categorized into those with short response times on confidence judgments and participants with long response times on confidence judgments by a median split separately for females and males. The confidence bias of these two groups is plotted in Figure 5.5. It shows that the more time males took for their confidence judgments, the more accurately they judged their confidence. Put differently, the less time males invested in their confidence judgments the more overconfident they were in their performance. Following Koriat (2012b), this finding suggested that more information-based judgments are more accurate for males. I did not find such a distinction of processes for females. These results imply that overconfidence in males could be the default mode and might rely on automatic processes; the more overconfident they were, the quicker they gave their confidence judgment. Interestingly, there was no evidence for the idea that underconfidence in one’s performance could be the default mode in females, at least not when it comes to answering knowledge questions.
5.3. Discussion

Under conditions of flat rate payment, that is when neither performance nor confidence judgments were incentivized, I replicated the findings of the pretest on the difficulty levels of the knowledge questions. Participants answered easy questions more often correctly than questions of medium difficulty, while questions categorized as difficult were least often answered correctly. This categorization of knowledge questions into three levels of difficulty was additionally confirmed by a main effect of item difficulty on the response times on the answers of the knowledge test. Participants answered easy questions quickest, followed by questions of medium difficulty and they took longest for answering difficult questions. The latter were more quickly answered as medium and difficult questions (Hypothesis 1b confirmed). There was no gender effect on performance, females and males performed equally well on the knowledge questions and were equally fast in answering the questions.

This pattern of item difficulty was also confirmed by the confidence judgments. Participants judged their confidence on easy questions higher than on questions of medium difficulty and lowest on difficult questions. Hence, Hypothesis 1a was fully confirmed. The response times on the confidence judgments also reflected this pattern. The response
times on the confidence judgment were quickest on easy questions, slower on questions of medium difficulty and longest on difficult questions. This somewhat contradicts Koriat’s (2012a) assumption that confidence judgments are based on mnemonic cues (in this case the response time on the answer). If this was true the response times on the answers should vary depending on item difficulty but the response times on the confidence judgments should be constant since the information from the mnemonic cue (the response time on the answers) is always the same and should not result in additional deliberation on the confidence judgment. The correlation between the two response times was highly significant and positive. This indicates that the two processes (answering a question and judging the confidence in that answer) are strongly linked and that the confidence judgment is not exclusively based on the process information of the answer to knowledge question.

When putting performance in the knowledge test in relation to the confidence judgments by computing the confidence bias, I successfully replicated the hard-easy effect (Lichtenstein et al., 1982; see also Merkle, 2009). All participants were underconfident on easy questions and overconfident on difficult questions, while they were rather accurate in their confidence judgment on their performance on questions of medium difficulty. This result confirmed Hypothesis 3. Contrary to the expectations, gender did not impact the confidence bias which means that Hypothesis 5 was not confirmed. There was also no interaction between gender and item difficulty on this bias.

Following the argument that people have a good monitoring resolution (discrimination between correct and incorrect answers), participants judged their confidence in correct answers significantly higher than on incorrect answers. This was also supported by the pattern of response times on the answers meaning that correct answers were given significantly quicker than incorrect ones. Thus, Hypothesis 4a and b were confirmed. This demonstrates that people have a good monitoring resolution without being incentivized for performance or an accurate confidence judgment. Further, all participants were quicker in judging their confidence when giving correct answers compared to incorrect answers. Again, this supports the assumption that the confidence judgments are not exclusively based on mnemonic cues (response time on the answers) but are also informed by additional information. The fact that the response times on the confidence judgments are significantly longer on incorrect answers provides evidence for a post-decisional stage of evidence accumulation (see Moran et al., 2015).

Furthermore, I examined Koriat’s (2012a) argument that confidence judgments are based on mnemonic cues by calculating the correlation between the response times on
the answers and the confidence judgments. The correlation was negative which points to an experience-based judgment, meaning that participants used the information of mnemonic cues (response time on the answer) as indicator of their confidence (Hypothesis 2). The longer it took a person to answer a question, the less confident she was in the correctness of this answer. Note, that I only found a significant negative correlation between response times on the answers and the confidence judgments for females (not males). Females seem to rely more on their experience with the task for judging their confidence than males. All in all, the results indicate that the confidence judgments are indeed informed by mnemonic cues as for example the response times on the answers but not exclusively. The variation of the response times on the confidence judgment depending on item difficulty and the correctness of the answer supports the notion of a dual stage model of confidence judgments with a post-decisional stage of evidence accumulation (see Moran et al., 2015).

The analysis of response times on the confidence judgments revealed that, as expected, the confidence bias was negatively predicted by response times on the confidence judgments. This result suggests that the more time people invested in their confidence judgments on the correctness of their answers, the less overconfident they were, meaning they judged their performance more accurately. This finding is an indicator that rather controlled (slow) processes of confidence judgments support more accurate confidence judgments than automatic (quick) judgments.

In contrast to previous research on processes of confidence judgments, I analyzed the data separately for females and males. They revealed that males drove the link between response times on confidence judgments and the confidence bias. The less time males invested in their confidence judgments, the more they were overconfident in their performance. In case of slower confidence judgments, males were more accurate on judging their performance. This finding sheds more light on the link between response times and metacognitive accuracy. Note, that this stands in contrast to findings by Siedlecka et al. (2019) who found that longer response times on the confidence judgments were related to lower metacognitive accuracy (more bias). Yet, this finding by Siedlecka et al. (2019) was mainly based on female participants (Study 1: 71% female participants; Study 2: 85% female participants). This is a first indicator that the processes underlying confidence judgments differ between females and males. This link between response times on confidence judgments and the confidence bias was not observed for females. Furthermore, females did not show any signs of underconfidence in case of short response times on their confidence judgments. From a dual process perspective, it could be argued
that overconfidence might be the default mode in males as it relied on rather quick (automatic) processes, but for females underconfidence was not the default mode. Hence Hypothesis 6 was only confirmed for male participants.
6. Study 2: The Effects of Incentives for Performance on Confidence Judgments

This experiment was conducted at another university (Universität of Cologne) than Study 1. It investigated the effects of incentives on performance in the general knowledge test and their impact on non-incentivized confidence judgments and the confidence bias. Importantly, I scrutinized if the impact of the incentive on performance differed between genders. I also explored the effects of the incentive for performance on processes of answering the knowledge questions and processes of the confidence judgments and if it was distinct for females and males.

The same paradigm and knowledge questions were used as in Study 1. I compared the effect of incentives on performance in the knowledge test to those in a flat rate payment condition. In the former, participants received an incentive for each correct answer. In the latter, participants were paid for participating in the experiment, independent of their performance or accuracy of confidence judgments. The flat rate payment condition was supposed to replicate the findings of Study 1. Due to this design, it was possible to compare the effects of these two different payments on performance and its processes, and on confidence judgments, and their processes. An equal number of females and males was recruited as participants for the study to explore potential gender effects.

I expected to replicate the results of the pretest and Study 1 on the effect of item difficulty on the performance and the confidence judgments. Participants should answer more easy questions correctly than questions of medium difficulty followed by difficult questions. The confidence judgments should reflect the same pattern. Response times on the answers and on the confidence judgments should support the categorization of the questions. They should show the following pattern of findings: easy < medium difficult < difficult questions. Furthermore, I predicted to find a positive correlation between the two response times as in Study 1. I postulated a replication of the hard-easy effect.
observed in Study 1. I expected underconfidence on easy questions, a rather accurate judgment on questions of medium difficulty and overconfidence on difficult questions.

As in Study 1, I expected participants to have a good monitoring resolution (see Koriat, 2012b, 2012a) which is reflected by a link between the correctness of the answers and the confidence level. Participants should judge their confidence higher (and quicker) on correct answers than on incorrect answers in which the confidence judgments should be low (and slow). In addition, I predicted participants to give correct answers quicker than incorrect answers. I did not expect the incentive on performance to affect the monitoring resolution since the incentive depended only on performance and not on the confidence judgments (see Lebreton et al., 2018).

Participants in the incentive treatment were offered an incentive for each correctly answered question. They were expected to be more motivated to perform well in the knowledge test than participants in the flat rate payment condition (see Gneezy & Rustichini, 2000). Consequently, I claimed a higher performance in the incentivized condition than in the flat rate payment condition. I further predicted the response times on the answers to be longer in the incentive condition compared to the flat rate payment condition (see Achtziger & Alós-Ferrer, 2014; Koriat et al., 2006). Controlled processes should prevail under incentivization of performance because a higher motivation to perform well should result in higher time investments, for instance due to more effort and attention deployed to the knowledge test in order to improve one’s performance in this task. Since the incentive for performance was expected to prolong the response times on the answers, I did not necessarily expect to find the negative correlation between the response time on the answers and the confidence judgments of Study 1. The incentive for performance might be used as additional information for the confidence judgment so that it is not based on the mere time it took to give an answer. The question of whether the incentive on performance influences the response time on the confidence judgments remains open.

Based on the findings of Study 1, I expected that the less time participants invested in the confidence judgments, the higher their confidence bias (especially males were expected to show this behavior and reveal automatic overconfidence). Despite the lack of a gender effect on the confidence bias in Study 1, I expected to find a gender effect (i.e., overconfidence in males) in both conditions because Study 1 still provided some evidence that overconfidence could be the default mode in males. The incentive on performance is not expected to reduce the confidence bias because the incentive is paid completely independent of the accuracy of the confidence judgment (see Lebreton et al.,
To conclude, the hypotheses of Study 2 were the following:

**Hypothesis 1a**: The replication of the effect of item difficulty on the performance and the confidence judgments of Study 1 was expected (easy > medium > difficult).

**Hypothesis 1b**: The replication of the effect of item difficulty on the response times on the answers and the confidence judgments was expected (easy < medium < difficult).

**Hypothesis 1c**: A positive correlation between the response times on the answers and the response times on the confidence judgments as in Study 1 was predicted.

**Hypothesis 2**: The replication of the hard-easy effect as in Study 1 was expected.

**Hypothesis 3a**: Participants are expected to have a good monitoring resolution. Confidence judgments on correct answers should be higher than on incorrect answers.

**Hypothesis 3b**: The response time on the answers and on the confidence judgments should be quicker on correct answers than on incorrect answers.

**Hypothesis 4**: Males are expected to display overconfidence whereas females are expected to be underconfident.

**Hypothesis 5**: The response time on the confidence judgment is negatively related to the confidence bias (especially for males).

**Hypothesis 6a**: The incentive for performance should increase performance on the knowledge questions.

**Hypothesis 6b**: The response time on the answers should be longer in the incentive condition compared to the flat payment condition.
6.1. Methods

6.1.1. Participants
255 participants (127 females, 128 males) were recruited among the student population of a university using ORSEE (Greiner, 2015), a standard online recruitment system that allows for random recruitment from a predefined subject pool. They were between 18 and 58 years old, $M = 22.96$ ($SD = 3.64$). Students majoring in psychology were not allowed to participate. In exchange for participation, they received either a payment of 14 €, or a payment based on their performance on the knowledge questions, plus a show-up fee of 4 €. Participants in the treatments with incentives earned on average 13.45 € (range: 10 - 17 €).

6.1.2. Design
The study followed a $2 \times 2 \times 3$ (incentive, gender, difficulty) mixed model design. Dependent variables were percentage of correct answers (i.e., performance), confidence judgments, the confidence bias (see above), response times on the confidence judgments, and response times on the knowledge questions.

6.1.3. Materials
The knowledge questions and the confidence judgments scale were identical to those of Study 1.

6.1.4. Procedure
The study was conducted using SoSci Survey (Leiner, 2019). Students took part in group sessions of 30-32 individuals. A session lasted about 45 minutes. Participants in the incentive condition were instructed that they would receive 50 Cent for each correctly answered question in addition to a show-up fee of 4 Euros. Participants in the flat rate payment treatment were told, that they would receive 14 Euros for their participation. The rest of the procedure was identical to Study 1.

---

3I reduced the post-experiment questionnaires to the State Self-Control Scale (Ciarocco et al., 2004), the Strenuous Mental Activity Scale (Job et al., 2010), the Sex-Role Inventory (Berger & Krahé, 2013). These scales are not further mentioned due to a null effect on all dependent variables.
6.2. Results

6.2.1. Number of Correctly Answered Knowledge Questions

A mixed model ANOVA with the between-subjects factors gender (male vs. female), incentive (incentive vs. flat rate payment) and the within-subjects factor item difficulty showed the expected highly significant effect of item difficulty, $F(2, 502) = 965.06, p < .001, \eta^2_p = .79$. On average, participants answered $M = 86.55\% (SD = 12.13\%)$ of the easy questions, $M = 55.49\% (SD = 17.94\%)$ of the medium questions and $M = 34.04\% (SD = 17.70\%)$ of the difficult questions correctly. Subsequent $t$-tests showed that participants answered significantly more easy than medium questions correctly, $t(254) = 26.14, p < .001$, and significantly more medium than difficult questions correctly, $t(254) = 17.10, p < .001$. This validated the experimental material of the pretest and of Study 1.

A highly significant gender effect occurred, $F(1, 251) = 17.42, p < .001, \eta^2_p = .07$, with males answering significantly more questions correctly than females ($M_m = 61.64\%; SD_m = 10.62\%; M_f = 55.72\%; SD_f = 12.11\%)$. This finding indicated that in contrast to Study 1 (and to the gender-neutral selection of the questions based on the pretest), in Study 2, a gender bias was revealed. Yet, because I was mainly interested in confidence judgments and their processes, this was not so problematic because the confidence bias (see below) puts performance and confidence judgment in relation to each other. There was no effect of incentive on performance, $F(1, 251) = .75, p = .388$. Hence, the incentive did not increase performance in the knowledge test. None of the interactions of this ANOVA reached significance, $Fs \leq 1.51, ps \geq .220$.

6.2.2. Response Times on the Answers to the Knowledge Questions

I logarithmized the response times as they were skewed. First, I compared the response times on the answers across gender, incentive, and item difficulty. In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factors gender and incentive and the within-subjects factor item difficulty yielded the predicted effect of item difficulty, $F(1.85, 464.09) = 362.97, p < .001, \eta^2_p = .59$. Participants answered easy questions in $M = 7576.12 \text{ ms} (SD = 5220.75 \text{ ms})$, medium questions in $M = 10229.75 \text{ ms} (SD = 3301.15 \text{ ms})$, and difficult questions in $M = 11884.14 \text{ ms} (SD = 5205.17 \text{ ms})$. Subsequent $t$-tests showed that easy questions were answered significantly quicker than medium questions, $t(254) = 19.54, p < .001$, and
medium questions significantly quicker than difficult questions, $t(254) = 8.32, p < .001$. The analysis showed no gender effect, $F(1, 251) = .78, p = .379$.

As hypothesized, a highly significant incentive effect emerged, $F(1, 251) = 16.14, p < .001, \eta^2_p = .06$ (see Figure 6.1). Participants in the incentive condition took longer to answer the knowledge questions ($M = 10728.54$ ms; $SD = 4189.66$ ms) than participants in the flat rate payment condition ($M = 9071.29$ ms; $SD = 2446.67$ ms). Incentivized participants also showed a much higher variance in their response times on answers than participants in the flat rate payment condition.

Further, the analysis showed a significant interaction between item difficulty and gender, $F(1.85, 464.09) = 6.50, p = .002, \eta^2_p = .03$. T-tests revealed that females and males differed in their response time on easy questions, $t(253) = 2.30, p = .022$. Males took $M = 7438.61$ ms ($SD = 6948.24$ ms) to answer these questions, while females took $M = 7714.72$ ms ($SD = 2499.08$ ms). There was no gender difference on response times on answers on medium, $t(253) = .50, p = .619$, or on difficult questions, $t(253) = -.78, p = .437$. No other interaction reached significance, $Fs \leq .35, ps \geq .555$.

Next, I compared response times on correct answers with those on incorrect answers. A mixed model ANOVA with the between-subjects factor gender and incentive and the within-subjects factor correctness (correct vs. incorrect answers) yielded the postulated effect of correctness, $F(1, 251) = 382.12, p < .001, \eta^2_p = .60$ (see Figure 6.2). Participants gave correct answers quicker, $M = 8923.41$ ms ($SD = 4430.71$ ms), than incorrect answers, $M = 12577.37$ ms ($SD = 4142.51$ ms). No gender effect emerged, $F(1, 251) = .09, p = .768$, but a significant interaction between correctness and gender, $F(1, 251) = 5.35, p = .022, \eta^2_p = .02$. T-tests showed that the difference in response times between correct and incorrect answers was significant in females, $t(126) = -17.81, p < .001$. Women took $M = 8820.55$ ms ($SD = 2362.22$ ms) for correct answers and $M = 12140.55$ ms ($SD = 3616.77$ ms) for incorrect answers. The difference in response times was also significant for males, $t(127) = -12.58, p < .001$. Males took $M = 9025.48$ ms ($SD = 5805.62$ ms) for correct answers and $M = 13010.78$ ms ($SD = 4578.50$ ms) for incorrect answers. There was no significant difference between females and males in the response time on correct, $t(253) = .76, p = .447$, or incorrect answers, $t(253) = -1.43, p = .153$. Again, there was a significant effect of the incentive, $F(1, 251) = 15.89, p < .001, \eta^2_p = .06$. Participants were slower in answering the questions in the incentive condition compared to the flat rate payment condition (see above). This effect did neither depend on the correctness of the answer nor on gender. No other interaction reached significance, $Fs \leq .27, ps \geq .604$. 

45
Figure 6.1.: Study 2: Response time on answers by incentive condition

Figure 6.2.: Study 2: Response time on answers by correctness
6.2.3. Confidence Judgments

The mean confidence judgments were compared across gender, incentive, and item difficulty. In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factors gender and incentive and the within-subjects factor item difficulty discovered the expected significant effect of item difficulty, $F(1.92, 481.85) = 1044.50, p < .001, \eta^2_p = .81$. On average, participants judged their confidence on easy questions with $M = 80.16$ ($SD = 13.22$), on questions with medium difficulty with $M = 51.57$ ($SD = 18.92$) and on difficult questions with $M = 43.46$ ($SD = 17.61$). Subsequent $t$-tests showed that the confidence judgment on easy questions was significantly higher than on medium questions, $t(254) = 30.74, p < .001$, and the confidence judgments on medium questions were significantly higher than that on difficult questions, $t(254) = 10.51, p < .001$.

A significant gender effect occurred, $F(1, 251) = 38.26, p < .001, \eta^2_p = .13$, and a significant interaction between item difficulty and gender emerged, $F(1.92, 481.85) = 6.27, p = .002, \eta^2_p = .02$. $T$-tests showed that males and females differed significantly in their confidence judgments for easy questions, $t(229.67) = 4.64, p < .001$, medium questions, $t(253) = 5.05, p < .001$, and difficult questions, $t(253) = 6.50, p < .001$. Males assessed their confidence in the correctness of their answers always higher than females. This gender difference even increased with higher levels of difficulty ($\Delta M_{\text{easy}} = −7.39$, $\Delta M_{\text{medium}} = −11.42$, $\Delta M_{\text{difficult}} = −13.29$).

There was no significant effect of the incentive, $F(1, 251) = .90, p = .343$, but a weak interaction between incentive and difficulty, $F(1.92, 481.85) = 2.46, p = .087, \eta^2_p = .01$. $T$-tests showed that the incentive condition and the flat rate payment condition differed significantly in their confidence judgment for easy questions, $t(253) = 2.08, p = .039$. Participants in the incentive condition judged their confidence on easy questions on average lower ($M = 78.45; SD = 13.91$) as participants in the flat payment condition ($M = 81.86; SD = 12.91$). The $t$-tests showed no difference between the incentive condition and the flat rate payment condition for medium questions, $t(253), p = .867$, and difficult questions, $t(253) = .59, p = .559$. None of the other interactions reached significance, $Fs \leq .15, ps \geq .859$.

Following Koriat (2012a) I calculated the correlation between the confidence judgments and the response time on the answer to the knowledge questions. The idea was that if this correlation was negative, the confidence judgments were experience-based and relied on mnemonic cues (response time on the answers). It turned out that this correlation was significant and negative as in Study 1, $r = −.27, p < .001$. I then calculated this
correlation separately for females and males to scrutinize potential gender effects. The correlation was negative and significant both for females, \( r = -0.26, p = .003 \), and males, \( r = -0.31, p < .001 \).

For a more detailed picture I calculated the correlations for females and males for the two payment conditions (flat rate payment vs. incentive for performance) separately. The correlation between the confidence judgments and the response time on the answers in the flat rate payment condition was neither significant for females, \( r = -0.13, p = .310 \), nor for males, \( r = -0.02, p = .840 \). But in the incentive condition, the correlation between the confidence judgments and the response times on the answers reached significance for females, \( r = -0.37, p = .003 \), and males, \( r = -0.47, p < .001 \). A Z-test showed no significant difference between the correlation coefficients, \( Z = 0.67, p = .252 \). Hence, the incentive for performance prompted all participants to rely their confidence judgments more on experience (response time on the answers) (see Koriat, 2012b). Yet, in the flat rate payment condition, response times on answers were not systematically linked to the confidence judgments in females as in Study 1.

Following Lebreton et al. (2018), I explored the confidence judgments for correctly and incorrectly answered questions separately. A mixed model ANOVA with the between-subjects factors gender and incentive, and the within-subjects factor correctness (correct vs. incorrect) revealed a significant main effect of the correctness of the answers on the confidence judgments, \( F(1,251) = 1554.88, p < .001, \eta^2_p = .86 \) (see Figure 6.3). Not surprisingly, the confidence judgments for correctly answered questions \( M = 72.35 \) (\( SD = 13.62 \)) were higher than for incorrect answers \( M = 38.61 \) (\( SD = 17.19 \)). Again, this result validated the experimental material and replicated the findings of Study 1. It also supported the idea by Lebreton et al. (2018) that confidence judgments are based on noisy evidence and Koriat’s (2012a) idea that people have overall a good monitoring resolution. This finding showed that people have quite a good idea of being right or wrong in answering knowledge questions.

There was a significant gender effect, \( F(1,251) = 27.34, p < .001, \eta^2_p = .10 \) (see Figure 6.3). Males were more confident in the correctness of their answers than females (\( M_{female} = 53.02, SD_{female} = 14.88, M_{male} = 63.73, SD_{male} = 12.69 \)). In addition, there was a significant interaction between the correctness of an answer (correct vs. incorrect) and gender, \( F(1,251) = 657.55, p = .008, \eta^2_p = .03 \). On correct answers, males judged their confidence higher (\( M = 75.53; SD = 11.20 \)) than females (\( M = 69.14; SD = 15.05 \)), \( t(232.72) = 3.85, p < .001 \). Interestingly, on incorrect answers, males also judged their confidence higher (\( M = 44.05; SD = 16.03 \)) than females (\( M = 33.13; SD = 16.36 \)),
Note that this gender difference was bigger for incorrect than for correct answers ($\Delta M_{\text{correct}} = -6.34$, $\Delta M_{\text{incorrect}} = -10.92$), leading to the interaction between gender and correctness of answers. The difference in confidence judgments of correct and incorrect answers was significant for females, $t(126) = 30.41$, $p < .001$, and males, $t(127) = 25.64$, $p < .001$. These results suggested that ironically males were especially more confident in the correctness of their answers in case of incorrectly answered questions than females. The incentive on performance did not affect confidence judgments on correctly and incorrectly answered questions differently than the flat rate, $F(1, 251) = .66$, $p = .418$. This means that the incentive on performance did not improve the monitoring resolution compared to the flat rate payment condition. No other interaction reached significance $Fs \leq .27$, $ps \geq .602$.

Figure 6.3.: Study 2: Confidence judgments by correctness and gender

6.2.4. Response Times on the Confidence Judgments

I also compared the response time on the confidence judgments (also logarithmized) across gender, incentive, and item difficulty. In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factors gender and incentive, and the within-subjects factor item difficulty yielded a significant
effect of item difficulty, $F(1.85, 463.89) = 318.50, p < .001, \eta^2_p = .56$. Participants judged their confidence on easy questions quicker $M = 9555.26$ ms ($SD = 5472.11$ ms) than on medium questions $M = 12204.55$ ms ($SD = 3582.45$ ms) and were slowest with their confidence judgments on difficult questions $M = 13922.49$ ms ($SD = 5600.14$ ms). T-tests confirmed that the response time for easy questions was significantly quicker than for medium questions, $t(254) = 18.13, p < .001$, and for medium questions significantly quicker than for difficult questions, $t(254) = 6.21, p < .001$. This pattern of response times on confidence judgments replicated the findings of Study 1.

The ANOVA did not reveal a gender effect, $F(1, 251) = .02, p = .877$, but a significant main effect of incentive, $F(1, 251) = 10.23, p = .002, \eta^2_p = .04$. Participants in the incentive treatment $M = 12654.86$ ms ($SD = 4484.79$ ms) took longer to judge their confidence on the correctness of their answers than participants in the flat rate payment treatment $M = 11139.28$ ms ($SD = 2922.73$ ms). Thus, controlled judgment processes prevailed more in the incentive treatment even if the incentive was provided for performance in the knowledge test and not for being accurate in the confidence judgment. This is somewhat surprising for the reason that Koriat (2012a) argues that confidence judgments rely on mnemonic cues (time needed to retrieve an answer). Hence the incentive for performance should not have an influence on the response time for the confidence judgment. In this ANOVA a significant interaction between item difficulty and gender emerged, $F(1.85, 463.89) = 11.70, p < .001, \eta^2_p = .05$. T-tests showed that males and females differed in their response time on confidence judgments on easy questions, $t(253) = 2.25, p = .025$. Males took less time $M = 9381.48$ ms ($SD = 7203.12$ ms) for their confidence judgment on easy questions than females $M = 9730.40$ ms ($SD = 2829.33$ ms). There was no gender difference for medium, $t(253) = -.49, p = .626$, or difficult questions, $t(253) = -1.55, p = .122$. No other interaction reached significance, $Fs \leq .55, ps \geq .578$.

Next, I compared response times for confidence judgments on correct answers with those on incorrect answers. A mixed model ANOVA with the between-subjects factor gender and incentive and the within-subjects factor correctness (correct vs. incorrect answers) yielded the expected effect of correctness, $F(1, 251) = 164.97, p < .001, \eta^2_p = .40$ (see Figure 6.3). Participants gave their confidence judgments quicker on correct answers $M = 10970.32$ ms ($SD = 4605.98$ ms) than on incorrect answers $M = 13319.48$ ms ($SD = 4227.03$ ms). No gender effect emerged, $F(1, 251) = .48, p = .489, \eta^2_p < .01$, but a marginally significant interaction between correctness and gender, $F(1, 251) = 3.59, p = .059, \eta^2_p = .01$. Subsequent t-tests showed that for females the difference in response
times for confidence judgments between correct and incorrect answers was significant, \(t(126) = -10.06, p < .001\). Females took \(M = 10786.44\) ms (SD = 2769.98 ms) for correct answers and \(M = 12864.07\) ms (SD = 3647.82 ms) for incorrect answers. The difference in response times was also significant for males, \(t(127) = -8.79, p < .001\). Males took \(M = 11152.77\) ms (SD = 5894.95 ms) for correct answers and \(M = 13771.32\) ms (SD = 4703.25 ms) for incorrect answers. T-tests confirmed that there was no difference between females and males in the response time on confidence judgments for correct answers, \(t(253) = .14, p = .890\), or incorrect answers, \(t(253) = -1.48, p = .139\).

As in the previous analysis, there was a significant effect of the incentive, \(F(1,251) = 10.80, p = .001, \eta^2_p = .04\). Participants made quicker confidence judgments in the flat rate payment condition \(M = 11139.28\) ms (SD = 2922.73 ms) compared to the incentive condition \(M = 12654.86\) ms (SD = 4484.79 ms). This effect did not depend on the correctness of an answer or on gender (all interactions: \(F's \leq 1.57, ps \geq .211\)).

As in Study 1, I calculated the correlation between the response time on the answers to the knowledge questions and the response time on the confidence judgments. The correlation was highly significant and positive, \(r = .96, p < .001\). Again, this highlights the strong connection between these two processes.
6.2.5. Confidence Bias

This bias was calculated as in Study 1. A mixed model ANOVA with the between-subjects factors gender, and incentive, and the within-subjects factor item difficulty resulted in a highly significant main effect of item difficulty, $F(2, 502) = 99.86, p < .001, \eta^2_p = .29$. Like in Study 1, all participants were underconfident on easy questions, $M = -6.39$ ($SD = 13.06$), rather accurate in their confidence on the correctness of answers on the questions of medium difficulty, $M = -3.92$ ($SD = 19.55$), and were overconfident on difficult questions, $M = 9.42$ ($SD = 20.71$). T-tests suggested that the confidence bias differed significantly from zero for easy questions, $t(254) = -7.81, p < .001$, for medium questions, $t(254) = -3.21, p = .002$, and for difficult questions, $t(254) = 7.27, p < .001$. T-tests confirmed that the confidence bias on easy questions differed significantly from the confidence bias on medium questions, $t(254) = 2.13, p = .034$, and the confidence bias on medium questions differed significantly from the bias on difficult questions, $t(254) = 10.89, p < .001$. Overall, the hard-easy effect (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982) observed in Study 1 was successfully replicated, except that participants in Study 2 were also underconfident on their performance on questions of
medium difficulty (see Figure 6.6).

There was a significant gender effect of, $F(1, 251) = 7.27, p = .008, \eta^2_p = .03$ (see Figure 6.7). As predicted, males’ confidence bias was clearly higher than that of females ($M_m = 2.09; SD_m = 12.83, M_f = -2.70; SD_f = 15.34$). The means of the confidence bias suggested overconfidence in males (as it was positive) and underconfidence in females (as it was negative). Note that this gender difference was not due the better performance of males compared to females as the computation of the confidence bias accounted for the performance in the knowledge test. A deeper analysis revealed that males’ and females’ confidence biases differed from zero, but in opposite directions. The confidence bias deviated marginally from zero for males, $t(127) = 1.84, p = .068$, and significantly from zero for females, $t(126) = 1.98, p = .050$. This showed that males tended to be overconfident while females were definitely underconfident in their own performance. There was no effect of the incentive on the confidence bias, $F(1, 251) = .06, p = .815$. None of the interactions reached significance, $Fs \leq .90, ps \geq .406$.

Figure 6.6.: Study 2: The hard-easy effect
To investigate the impact of gender, incentive, and response times on the confidence bias, I ran linear regressions. All variables were z-standardized, except for the response times on confidence judgments which were logarithmized (due to their skewness). Gender (male = 1, female = 0) and incentive (incentive for performance = 1, flat rate payment = 0) were dummy-coded. Regressions on the confidence bias are presented in Table 6.1. Model 1 accounted for the main variables, gender, incentive, and the interaction between both. I observed the predicted gender effect with males displaying significantly higher scores on the confidence bias than females. Neither the incentive on performance or the interaction between incentive and gender were significant predictors for the confidence bias in Model 1. This is in line with the findings by Lebreton et al. (2018) that an incentive on performance did not significantly affect the confidence bias (in their case there was only a trend for an increased bias when performance was incentivized). Koriat et al. (2006) also found that their participants were overall overconfident but did not display a difference in bias between the incentive conditions (Experiment 7). Model 2 added response times on confidence judgments as a predictor of the confidence bias. The effect of gender remained significant and turned out to be quite robust. Congruent to Model 1, the incentive on performance and the interaction between incentive and gender did not predict the confidence bias in Model 2. Response times on confidence
judgments negatively predicted the confidence bias. Participants, who invested more time to judge their confidence on the correctness of their answers in the knowledge test were less overconfident in their confidence judgment (i.e., were more realistic in judging their performance) than participants who quickly judged their confidence. Thus, I replicated the findings on response times on confidence judgments as predictors of the confidence bias of Study 1 and expanded them to females. This is an indicator that confidence judgments are more realistic when more time is invested meaning that maybe not only mnemonic cues from the experience with the task are taken into account when making the confidence judgment.

Model 3 added the interaction between response times on the confidence judgments and the incentive on performance as a predictor. The effect of gender remained significant and was robust. The incentive on performance positively predicted the confidence bias, meaning that in the incentive condition, participants were more biased than in the flat rate payment condition. Interestingly, the interaction between the incentive and the response times on confidence judgments negatively predicted the confidence bias. A simple slope analysis showed that the response times on confidence judgments did not predict the confidence bias under flat rate payment (simple effects; \( b = .01, SE = .39, p = .978 \)). But in the incentive condition the response times on confidence judgments negatively predicted the confidence bias (simple effects; \( b = -1.21, SE = .30, p < .001 \)). This finding implied that the slower participants in the incentive condition judged the confidence in their performance, the lower was their confidence bias (i.e., the more accurate they were in their confidence judgment).

In order to illustrate that the negative prediction of the confidence bias by response times on confidence judgments in the incentive condition was distinct for both gender, I run a median split of response times on confidence judgments separately for males and females only for the incentive condition (see Figure 6.8; see also Study 1). It revealed that incentivized males who took longer to judge their confidence were less overconfident than males who quickly gave their confidence judgments. Thus, only for participants incentivized for performance in the knowledge test I replicated the finding of Study 1 for males, but not for flat rate paid males. The figure also indicated that incentivized females who slowly rated their confidence in their performance were more underconfident than females who quickly made their confidence judgments. Put differently, females in the incentive condition who quickly rated their confidence were fairly accurate with their confidence judgment, but if they thought too long about their performance, they showed an underconfidence bias. Hence the moderating effect of incentives on performance on
the relationship between response time on the confidence judgments and the accuracy of confidence judgments is mostly improving males’ judgments and not females'.

Table 6.1: Study 2: Linear regressions on confidence bias

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male = 1)</td>
<td>.44**</td>
<td>.45**</td>
<td>.44**</td>
</tr>
<tr>
<td>Incentive for performance (yes = 1)</td>
<td>.07</td>
<td>.15</td>
<td>8.68**</td>
</tr>
<tr>
<td>Gender × Incentive for performance</td>
<td>-.21</td>
<td>-.21</td>
<td>-.20</td>
</tr>
<tr>
<td>Response time on confidence judgments</td>
<td>-.67***</td>
<td>-.11</td>
<td></td>
</tr>
<tr>
<td>Incentive × Response time on confidence judgments</td>
<td></td>
<td></td>
<td>-.92*</td>
</tr>
<tr>
<td>Constant</td>
<td>-.20*</td>
<td>5.97***</td>
<td>.82</td>
</tr>
</tbody>
</table>

| N                                             | 255       | 255       | 255       |
| R²adj.                                        | .02       | .05       | .06       |
| F                                             | 2.67**    | 4.16***   | 4.13***   |

Notes. Standard errors in brackets. *p < .10, **p < .05, ***p < .01.
6.3. Discussion

I replicated the findings of the pretest and Study 1 on the difficulty levels of the knowledge questions. Even when half of participants were paid for performing well in the knowledge test the categorization of the questions into three levels of difficulty was confirmed. All participants answered questions that I categorized as easy questions based on the pretest and on Study 1 more often correctly than questions formerly categorized as medium difficult, while questions previously categorized as difficult were least often answered correctly. The incentive did not impact these observations. In contrast to Study 1, males answered more questions correctly than females. This was a main effect and did not interact with item difficulty. Note that this was only a difference of 6% (males: 62% correct answers over all questions, females: 56%). As Study 2 was conducted at another university, this gender bias might be caused due to specificities of the population of this university. It was not problematic for the testing of the main research questions, because the computation of the confidence bias accounted for actual performance in the knowledge task.

As in Study 1, this categorization of knowledge questions in three difficulty levels
was additionally supported by a main effect of item difficulty on the response time on the answers of the knowledge test. All participants took longer for answering difficult questions as for answering questions of medium difficulty and easy questions. The latter were more quickly answered as medium and difficult questions. There was a significant interaction between gender and difficulty level on response times on answers. It turned out that the interaction was exclusively based on the finding that females took more time (i.e., were more cautious) for answering easy questions than males. There were no gender differences on response times on answering medium difficult and difficult questions, though. Note that females did not better than males on easy knowledge questions. Thus, their higher time investment did not result in better performance but might simply reflect more caution in females even on easy questions. This pattern of results did not interact with the incentive, but the incentive increased response times on answers on all questions (i.e., independent of their difficulty) compared to the flat rate payment condition which confirmed Hypothesis 6b. In addition, this supports the finding by Koriat et al. (Koriat et al., 2006) that an incentive for performance increases the response time on the answers. Thus, participants put more effort in answering the knowledge questions correctly in the incentive condition than participants who were paid flat rate. Further, much higher standard deviations on response times on the answers on the knowledge questions emerged in the incentive condition compared to the flat rate condition. That is, under incentivization of performance, people varied strongly in the time they invested in answering a question. This might reflect that some participants were strongly motivated to perform well due to the incentive and invested more time, while others were not very motivated by the incentive and did not invest more time, which finally ended up in high standard deviations in the incentive condition.

Like in Study 1, confidence judgments indicated that all participants were more confident in the correctness of their answers on easy questions than on medium difficult questions, and finally they were less confident in the correctness of their answers on difficult questions than on easy and on medium difficult questions. Hence, Hypothesis 1a was confirmed. I also observed a gender difference on confidence judgment. Males were more confident in the correctness of their answers than females. This effect increased in strength from easy to medium to difficult questions and gave rise to a gender x item difficulty interaction. Thus, males were more confident on their performance as females in general, but specifically so on their performance on difficult questions. In this sense, I discovered a gender bias on confidence judgments, but this also depended partly on the difference in performance (see above). The incentive for performance did not influence
Also in replication of Study 1, all participants judged the confidence on the correctness of their answer quicker on easy questions, followed by response times on confidence judgments on questions of medium difficulty, and they were slowest when judging their confidence on the correctness of their answer on difficult questions. Thus, Hypothesis 1b was confirmed. This also spoke for the idea that people have a good feeling of whether they might be correct or incorrect, because they accounted for difficulty levels of the questions as indicated by response times on confidence judgments. Again, as in Study 1, this conflicts with Koriat’s (2012a) argument that confidence judgments are based on mnemonic cues. The variation of the response time speaks for a post-decisional evaluation of confidence. What I also discovered was that males gave their confidence judgments on their performance on easy questions quicker than females. This gender difference did not appear for medium difficult questions and for difficult questions. Surprisingly, the incentive for performance increased the response time on the confidence judgments compared to the flat rate payment. This is interesting, since the incentive only depended on the performance on the knowledge questions and not at all on the accuracy of the confidence judgments. As in Study 1 and in support of Hypothesis 1c, I found a highly significant and positive correlation between the response time on the answers and the response times on the confidence judgments. This confirmed the assumption that the two processes (answering the question and judging the confidence in that answer) are strongly connected.

On the confidence bias, I replicated the hard-easy effect (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982) that already occurred in Study 1. All participants were underconfident on easy questions and overconfident on difficult questions. Note that in Study 2, participants were also slightly underconfident in their performance in the questions of medium difficulty, which was not the case in Study 1. The hard-easy effect (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982) was not affected by the incentive on performance and it was alike for females and males. Hence, Hypothesis 2 was confirmed except for the slight underconfident on questions of medium difficulty.

Like in Study 1, participants showed a good monitoring resolution. The confidence judgments on correctly answered questions were significantly higher than judgments on incorrectly answered questions in both genders. This finding confirmed Hypothesis 3a. Not surprisingly however, males’ confidence judgments were significantly higher than females’ confidence judgments. This was especially true for incorrectly answered questions, an observation that implied males’ overconfidence and suggested less valid
feelings regarding the correctness of their answers compared to females. Thus, males were either less motivated or able to discriminate between correct and incorrect answers than females. Whether this observation was due to a lack of motivation to accurately assess their confidence in males or whether it had other reasons like being less able to rely confidence judgments on evidence in males than in females could not be answered by the present data. As expected, the incentive on performance did not affect confidence judgments on correct and incorrect answers differently. That means that the incentive did not increase the sensitivity to the evidence of being correct or incorrect compared to the flat rate payment condition.

The pattern of response times confirmed the finding on the monitoring resolution. Participants gave correct answers quicker than incorrect answers and they judged their confidence on correct answers quicker than on incorrect answers. Again, this replicates the findings of Study 1 and confirmed Hypothesis 3b. There was an interaction between gender and correctness of the answers for both response times (on the answers and the confidence judgments). Yet, this can be neglected, because subsequent t-test confirmed that there was no significant difference in the response times (for the answer and the confidence judgment) on correct answers incorrect answers between females and males. The same was true for the response times on incorrect answers.

As in Study 1, the findings supported Koriat’s (2012a) argument that confidence judgments are (at least partly) based on mnemonic cues (response time on the answer). I found a negative correlation between the response times on the answers and the confidence judgments but surprisingly only for incentivized participants. This indicates that the response times on the answers (fluency of retrieving the answer to the question) are used as a mnemonic cue for the confidence judgments (see Koriat, 2012a) but only if the performance is incentivized. In other words, the longer it takes a person to retrieve the answer to a question, the lower the confidence that this answer is correct. When performance was incentivized, participants relied more on the response time on their answers in order to judge their confidence in the correctness of these answers. One could argue, that since participants put more effort in answering the questions when they were incentivized (indicated by longer response times in the incentive condition compared to the flat rate payment condition), they also relied more on the information their receive from their response time on the answers for their confidence judgments. This seems not to be true for participants who received a flat rate payment which is somewhat surprising since the flat rate payment condition was a replication of Study 1. Since Study 2 was conducted at another university than Study 1, this difference between the findings in
both studies might be due to the sample.

The research aim of Study 2 was to investigate the influence of monetary incentive for performance on confidence judgments and the underlying processes. Contrary to the expectations (Hypothesis 6a), the incentive on performance did not increase the number of correctly answered questions in the knowledge test. A possible explanation for the missing increase of performance in the incentive condition could be that the incentive was not salient enough (see Cerasoli, Nicklin, & Nassrelgrgawi, 2016). That would mean that the incentive might have been so small that participants ignored that their earnings in the experiment would depend on their performance. But the fact that participants in the incentive condition invested more time for their answers, ruled out this explanation. The following explanation of the missing incentive effect on performance was more plausible: Even if participants invested more time in answering the knowledge questions in the incentive condition as in the flat rate payment condition, they could not do better in this task as they either knew the correct answer or not. So, their performance could not be improved by putting more effort into the task. This finding conflicts with literature claiming that offering incentives for good performance in tests (even IQ tests) increases performance because of higher motivation to do well in this task (e.g., see Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). The present findings indicate that the dominance of controlled processes, presumably due to being highly motivated to retrieve the correct answer from long-term memory, did not result in a better performance in the incentive condition compared to the flat rate payment condition.

In Study 2, I observed a gender effect on the confidence bias over all levels of item difficulty. Males tended to be overconfident in their performance (when testing their confidence bias against zero), while females definitively were underconfident (i.e., their confidence bias significantly differed from zero). This result confirmed Hypothesis 4. Note that this gender effect did not interact with the payment, thus is was not influenced by the incentive. Hence, the incentive for performance did not reduce the confidence bias in both genders (overconfidence in males and underconfidence in females).

As in Study 1, I investigated the processes underlying confidence judgments by analyzing the relationship between the response time on the confidence judgments and the confidence bias. In line with the results of Study 1, I obtained some evidence that males’ often reported overconfidence in own performance might rely on rather automatic processes and hence be the default mode for this gender. Yet, different to Study 1 in which all participants were paid flat rate, I observed evidence for this hypothesis only in males who were incentivized for performance, but not in males who were paid flat rate.
Put differently, only in the incentive condition, males who invested more time in their confidence judgments were more accurate (i.e., less overconfident) in their confidence judgments than males who quickly judged their confidence. This confirmed Hypothesis 5 only partly. Why this effect did not occur in the flat rate payment condition was unclear because this condition was the same as Study 1 where I did not introduce any incentives but paid all participants flat rate only. Since Study 2 was conducted at another university than Study 1, this difference between the findings in both studies might be due to the sample.

Females tended to show a contrary pattern in case of being incentivized for performance. Only in the incentive condition, females who invested more time (and effort) in their confidence judgment tended to be underconfident, while females who judged their confidence quickly were accurate in their confidence judgments. Hence, in contrast to the assumption that females’ miscalibration that resulted in underconfidence might be grounded in automatic processes and hence qualified as a default response, things were even worse. When being incentivized for performance, females who deliberated carefully on the correctness of their answer (indicated by prolonged response times on confidence judgments) were less confident than females who judged their confidence quickly. One could argue that paying females for their performance makes them more critical and insecure regarding their performance after they thought carefully how well (or poor) they did in a knowledge test. Therefore, females miscalibrated their confidence judgments by (inaccurately) reducing them. This finding is in line with the results by Siedlecka et al. (2019). They found a negative relation between response times on the confidence judgments and meta-cognitive accuracy (longer response times, lower accuracy) but in a mainly female sample.

These gender differences on confidence judgments in the incentive condition supported the argument that when investigating the sources of confidence judgements, one should distinguish between a pre-decision phase and a post-decision/pre-confidence judgment phase (see Lebreton et al., 2018). For instance, models of post-decisional evidence accumulation (Navajas, Bahrami, & Latham, 2016) claim that if post-decisional information processes (i.e., processes after an answer was chosen in the test) are not biased, confidence judgments will be rather accurate. Yet, if these processes are biased, confidence judgments will degrade over time. In case of incentivized males, post-decisional information processes seemed to be not biased, because they became better the more time, they invested in the incentive condition (for flat rate payment, this increase in accuracy was not observed for longer confidence judgement times). In females, post-decision information processes
were extremely biased, because their judgment was miscalibrated over time and ended up in underconfidence (if performance was incentivized).
7. Study 3: The Effects of Incentives for Accuracy on Confidence Judgments

Since in Study 2 only performance in the knowledge test was incentivized, but not confidence judgments’ accuracy, the effects of incentives on processes of confidence judgments were only indirect. In Study 3, I therefore tested whether incentivizing the accuracy of the confidence judgment directly would result in more accurate confidence judgments and fewer gender differences in the confidence bias than observed in Study 2. The same knowledge questions as in Studies 1 and 2 were presented to measure task performance and confidence judgments on the correctness of the answers.

I replaced the scale for confidence judgments used in Studies 1 and 2 with the random binary choice elicitation mechanism (RBCEM) suggested by Healy (2016). This method had the advantage that it was incentive compatible. This means that the incentive was consistent with following the instructions (i.e., being accurate in one’s confidence judgments). By using this method, there was no motivation for participants to lie (give false confidence judgments) or to deliberately perform badly in order to increase the accuracy of the confidence judgment. The method used in this study is similar to the incentivizing methods used in several other studies (see Lebreton et al., 2018; Sanchez & Dunning, 2018). Another advantage of this method was that it was more implicit than the confidence rating scales used in Study 1 and 2. With the RBCEM (Healy, 2016) people are instructed to choose between the option to bet on the correctness of their answer on a question (e.g., winning 10€ if the answer to the question is correct) or the option of winning an incentive in a lottery (e.g., winning 10€ with the probability X; see below for detailed description). Through this mechanism I incentivized the accuracy of the confidence judgments instead of the performance in the knowledge test.

In Study 3, I also manipulated the level of the incentive itself. Participants could win either a low or a high monetary incentive. The probability of winning this incentive
depended on the option participants chose. Participants could increase the probability by judging their performance accurately. By offering a high incentive I intended to increase the incentive’s effect on the confidence bias (see Gneezy & Rustichini, 2000). The dependent variables were the same as in Study 1 and 2 (response times on the answers, performance in the knowledge test, response times on the confidence judgment, confidence judgments, and the confidence bias).

First, I intended to replicate the effect of item difficulty on the performance and the confidence judgments of Studies 1 and 2. The performance and the confidence judgments should display the following pattern: easy > medium > difficult. Further, I expected the same pattern of response times on the answers and the confidence judgments as in Studies 1 and 2 (i.e., easy < medium < difficult questions). In addition, the hard-easy effect observed in Studies 1 and 2 should replicate. The argument was that since I suggested the RBCEM (Healy, 2016) to validly measure confidence in performance, this method would provide the same covariation of confidence judgments with item difficulty observed in Study 1 and 2. The hard-easy effect and the effect of item difficulty on the performance, the confidence judgment and the response confirms the validity of the RBCEM (Healy, 2016) for measuring confidence.

In line with Studies 1 and 2, I predicted that participants show a good monitoring resolution meaning that the confidence should be higher for correctly answered questions and lower for incorrectly answered questions (see Koriat, 2012a; Lebreton et al., 2018). In addition, I expected incentive effects on the monitoring resolution. The high incentive should promote a higher metacognitive sensitivity compared to the low incentive (see Lebreton et al., 2018). This means that the high incentive should improve the ability to discriminate between correct and incorrect answers. Consequently, I expected higher confidence in correct answers in the high incentive condition compared to the low incentive. In conclusion, the confidence in incorrect answers should be lower in the high incentive condition compared to the low incentive condition.

The response times were also expected to be influence by the correctness of an answer. The response times on correct answers should be quicker than on incorrect answers. This should be independent of the incentive condition. The response times on the confidence judgments should also be quicker on correct answers than on incorrect answers. In addition, I expected an effect of the incentive condition. In line with the prediction on the monitoring resolution, I expected the response times on the confidence judgments for correct answers to be quicker in the high incentive condition compared to the low incentive. The reversed pattern was expected for the response times on confidence
judgments for incorrect answers (high incentive > low incentive). In line with the results of Study 1 and 2 and Koriat’s (2012a) assumptions, I predicted a negative relationship between the response time on the answers and the confidence judgment. Longer response times on the answers should be related to lower confidence judgments. This was assumed to be independent of the incentive for an accurate confidence judgment.

Regarding the confidence bias, I predicted that males would be overconfident while females would reveal underconfidence in the low incentive condition. I expected that the high incentive would support accurate confidence judgments more than the low incentive (see Blavatskyy, 2009; Hoelzl & Rustichini, 2005; Clark & Friesen, 2008). Camerer and Hogarth (1999) reported that high monetary incentives improved performance in judgment tasks that included judgments in own skills. Moreover, I found that longer response times on the confidence judgments were predictive for more accurate confidence judgments (in incentivized males) in Study 2. Hence, I expected the high incentive to result in longer response times on the confidence judgments which in turn has a positive effect on the accuracy of the confidence judgment (mainly in males). This hypothesis is supported by the assumptions of dual stage models of confidence judgments (see Moran et al., 2015). Given that post-decisional information processes (i.e., processes after an answer was chosen in the test) are not biased, then confidence judgments should be rather accurate (Navajas et al., 2016). Charles and Yeung (2019) as well as Siedlecka et al. (2019) found that metacognitive accuracy is improved if confidence judgments are informed before and after the initial decision. The high incentive for accuracy increases the likelihood of post-decisional information accumulation and hence a more accurate confidence judgment. Note that there are also finding that support the hypothesis that incentives for accuracy would increase the confidence bias (overconfidence) (see Lebreton et al., 2018; Sanchez & Dunning, 2018).

In Study 2, in the incentive for performance condition, I observed that prolonged response times on confidence judgments affected the confidence bias differently in females and males. The longer response time on confidence judgments improved males’ judgments but females who took longer were more biased than females who judged their confidence more quickly. Hence, I expected the effect of the high incentive for an accurate judgment to be more beneficial for males than females. Since this time the incentive depended on the accuracy of the confidence judgments, I did not expect the response times on the answers in the knowledge test to be longer in the high incentive condition compared to the low incentive condition (see also Lebreton et al., 2018). Hence, I expected to find a weaker positive correlation between the response times on the answers and the response
times on the confidence judgments compared to Study 1 and 2.

To conclude, the hypotheses of Study 3 were the following:

**Hypothesis 1a:** The replication of the effect of item difficulty on the performance and the confidence judgments of Studies 1 and 2 was expected (easy > medium > difficult).

**Hypothesis 1b:** The replication of the effect of item difficulty on the response times on the answers and the confidence judgments of Studies 1 and 2 was expected (easy < medium < difficult).

**Hypothesis 1c:** A (weak) positive correlation between the response time on the answers and the response time on the confidence judgments as in Studies 1 and 2 was predicted.

**Hypothesis 2:** The replication of the hard-easy effect as in Studies 1 and 2 was expected.

**Hypothesis 3a:** Participants are expected to have a good monitoring resolution (as in Studies 1 and 2). Confidence judgments on correct answers should be higher than on incorrect answers.

**Hypothesis 3b:** The monitoring resolution should be higher in the high incentive condition compared to the low incentive condition.

**Hypothesis 3c:** The response time on the answers and on the confidence judgments should be quicker on correct answers than on incorrect answers.

**Hypothesis 3d:** The response time on the confidence judgments should be quicker on correct answers and slower on incorrect answers in the high incentive condition compared to the low incentive condition.

**Hypothesis 4:** The response time on the answers serves as mnemonic cue for the confidence judgment which is shown by a negative correlation between the two variables (as in Studies 1 and 2).
Hypothesis 5: Males are expected to display overconfidence whereas females are expected to be underconfident in the low incentive condition.

Hypothesis 6a: The high incentive should decrease the confidence bias in both genders but more so for males.

Hypothesis 6b: The response time on the confidence judgments should be longer in the high incentive condition compared to the low incentive condition.

Hypothesis 7: The response time on the confidence judgment is negatively related to the confidence bias.

7.1. Methods

7.1.1. Participants

249 participants (128 females, 119 males, 2 without specification) were recruited among the student population of a university (University of Cologne) using ORSEE (Greiner, 2015). I excluded participants who majored in psychology. They were between 17 and 41 years old \( M = 22.70 \) \( (SD = 3.26) \). In exchange for participation, they received a payment of 7 or 10.50 € (depending on the condition). In addition, participants could win a bonus of 5 or 10 € based on how accurate they judged their performance (indicated by the RBCEM choices). Participants in the low incentive condition earned on average 14.32 € and participants in the high incentive condition 12.71 €.

7.1.2. Design

The study followed a 2 (incentive, between: low incentive vs. high incentive) × 2 (gender, between: female vs. male) × 3 (difficulty, within: easy vs. medium vs. difficult) mixed model design. Dependent variables were percentage of correct answers, switch points in the RCBEM (Healy, 2016) (see below for description), response times on answering the knowledge questions (just as in Study 1 and 2), and response times on the switch points (see below).

The switch point, an implicit measure of confidence, indicated the point at which participants switched from betting on the correctness of their answer to the lottery in
order to win the incentive (see detailed description below) in the RBCEM (Healy, 2016).
The response time on the switch points was measured from the moment at which the
answer to the knowledge question was given (by mouse click on one answer option) until
a switch point was indicated by mouse click on a continuous scale from 0 to 100. That
procedure was identical to the measurement of response time on confidence judgments in
Study 1 and 2.

7.1.3. Materials

Knowledge questions

The knowledge questions were identical to those of Study 1 and 2.

Confidence Judgments: Random Binary Choice Elicitation Mechanism (Healy, 2016)

I implemented the RBCEM (Healy, 2016) to measure participants’ beliefs that their
answers were correct. For each knowledge question, participants were presented with the
following table (example for the high incentive condition):

<table>
<thead>
<tr>
<th>Question</th>
<th>Option A</th>
<th>Option B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Would you rather have: 10€ if your answer to the question is correct</td>
<td>or 6% chance of 10€</td>
</tr>
<tr>
<td>1</td>
<td>Would you rather have: 10€ if your answer to the question is correct</td>
<td>or 1% chance of 10€</td>
</tr>
<tr>
<td>2</td>
<td>Would you rather have: 10€ if your answer to the question is correct</td>
<td>or 2% chance of 10€</td>
</tr>
<tr>
<td>3</td>
<td>Would you rather have: 10€ if your answer to the question is correct</td>
<td>or 3% chance of 10€</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>99</td>
<td>Would you rather have: 10€ if your answer to the question is correct</td>
<td>or 99% chance of 10€</td>
</tr>
<tr>
<td>100</td>
<td>Would you rather have: 10€ if your answer to the question is correct</td>
<td>or 100% chance of 10€</td>
</tr>
</tbody>
</table>

For each knowledge question, participants were instructed to choose for every row
of the table between Option A and Option B in order to measure their confidence in
the correctness of their answers. Following Healy (2016), I claimed that participants
would choose Option A in the first few rows of the table but would eventually switch
to choosing Option B in the remaining rows. For the first few choices, it would not be
reasonable to switch from betting on the correctness of the answer to a very low chance
of winning the incentive (for example 0, 1 or 2%). At one point, the participant would
trade Option A (winning if the answer was correct) to Option B (for example at a 90%
chance of winning the incentive).

The very first row, in which a participant switched from choosing Option A to choosing
Option B provided the switch point. The switch point meant that participants were
willing to trade the option of receiving the incentive if their answer was correct (Option A) for a probability indicated in Option B. Hence, the switch point indicated the participants confidence in the correctness of their answer. To facilitate working on the RBCEM task, the participants were only requested to indicate their switch point for each knowledge question (ranging from 0 to 100).

The participants were informed that at the end of the knowledge test, one of the knowledge questions would be randomly chosen by a computer for the payout. Then, one row of the corresponding table would be randomly chosen, and the choice indicated there (either Option A or Option B) would determine the payout. If a row below the switch point was randomly chosen, Option A would determine the payout. That means the participant would receive the incentive if the answer to the knowledge question was correct. If a row above the switch point was randomly chosen, Option B would determine the payout. That means that the participant would receive the incentive with the probability indicated in the randomly chosen row. The instructions highlighted that there was no incentive to lie on the RBCEM task. The RBCEM (Healy, 2016) was introduced as a decision task. Participants were also told that the best chance to win the bonus was an honest decision on choosing the switch point.

The following example illustrates the method. If a knowledge question was easy and the participant was quite sure that her/his answer was correct (e.g., approximately 80%), then the participant had to decide what the probability of winning the bonus by chance had to be to switch from Option A (winning the incentive if the answer was correct) to Option B (winning the incentive with a probability of x or higher). Assuming the participant would choose the switch point 80, a random selection of rows 0 to 79 would result in the incentive (low or high depending on the condition) if the answer was correct. In case of an incorrect answer, the participants would not get the incentive, but would only earn the show-up fee. If a row between 80 and 100 would be selected, then the participant would win the incentive with the probability indicated by Option B in that specific row. That means that if row 90 was randomly chosen, the participants would receive the incentive with a probability of 90%.

7.1.4. Procedure

The study was conducted in the same laboratory as Study 2 using SoSci Survey (Leiner, 2019). Students participated in group sessions of 30-32 individuals and a session lasted about 45 minutes. They were randomly assigned to one of the two incentive conditions (low incentive vs. high incentive). Participants in the high incentive condition were
instructed that they would receive 7 Euros for their participation and a chance of winning a bonus of 10 Euros depending on their choices in the RBCEM (Healy, 2016). Participants in the low incentive condition were told that they would receive 10.50 € for participation and a chance of winning a bonus of 5 € depending on their choices in the RBCEM (Healy, 2016). I varied the show-up fee between the low and high incentive condition so that overall the participants received equal payment. Again, the participants were instructed that they could maximize their chance of winning the incentive by choosing the switch points truthfully. The show-up fee differed between the two conditions, so that the participants in the two groups received on average the same amount for their participation (following the university’s behavioral economics’ lab rules). Following this instruction for the RBCEM (Healy, 2016), participants answered control questions concerning their understanding of the RBCEM (Healy, 2016). Then, they answered the same knowledge questions used in Study 1 and 2 and chose a switch point on the RBCEM (Healy, 2016) for each answer. The switch point operationalized the confidence in the correctness of an answer measure (see above). After answering all knowledge questions, participants filled out some questionnaires. Then, the computer program chose one of the knowledge questions randomly for the payout.

7.2. Results

7.2.1. Number of Correctly Answered Knowledge Questions

The percentage of correctly answered questions was compared across gender, incentive, and item difficulty. A mixed model ANOVA with the between-subjects factors incentive and gender, and the within-subjects factor item difficulty showed a significant main effect of item difficulty, $F(2, 486) = 1026.77$, $p < .001$, $\eta^2 = .81$. On average, participants answered $M = 86.14\%$ ($SD = 12.20\%$) of the easy questions, $M = 56.83\%$ ($SD = 16.75\%$) of the medium questions and $M = 32.49\%$ ($SD = 16.74\%$) of the difficult questions correctly. T-tests showed that participants answered significantly more easy than medium questions correctly, $t(248) = 25.08$, $p < .001$, and more medium than difficult questions, $t(248) = 20.42$, $p < .001$. This was the successful replication of the earlier result on the performance on the knowledge questions of Study 1 and 2.

There also was a significant main effect of gender, $F(1, 243) = 6.39$, $p = .012$, $\eta^2 = .03$.

\footnote{Faith in Intuition questionnaire (Keller et al., 2000) and questions regarding their statistical knowledge. Since participants self-ratings on the Sex-Role Inventory (Berger & Krahé, 2013) did not have any influence on performance and confidence judgments in Study 1 and 2, I did not include it in Study 3.}
with males answering on average more questions correctly than females ($M_f = 57\%, M_m = 60\%$). The difference between the low and high incentive condition was marginally significant, $F(1,243) = 2.99$, $p = .085$, $\eta^2 = .01$, while the interaction between item difficulty and incentive was significant, $F(2,486) = 3.48$, $p = .032$, $\eta^2 = .01$. When I analyzed this interaction, it turned out that participants in the low incentive condition answered more difficult questions correctly ($M = 35.04\%$ ($SD = 16.66\%$)) than participants in the high incentive condition ($M = 30.00\%$ ($SD = 16.49\%$), $t(247) = 2.40$, $p = .017$. The two incentive conditions did not differ in the percentage of correct answers for medium, $t(247) = -.38$, $p = .707$, and easy questions, $t(243,76) = 1.09$, $p = .279$.

### 7.2.2. Response Times on the Answers to the Knowledge Questions

I logarithmized the response times on the answers since they were skewed. I compared the response times on the answers across gender, incentive, and item difficulty. In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factors gender, incentive, and the within-subjects factor item difficulty yielded a significant main effect of item difficulty, $F(1.88,455.83) = 381.14$, $p < .001$, $\eta^2 = .61$. Participants answered easy questions in $M = 8162.11$ ms ($SD = 2438.12$ ms), medium questions in $M = 11674.39$ ms ($SD = 3410.39$ ms), and difficult questions in $M = 13186.08$ ms ($SD = 4454.50$ ms). T-tests confirmed that participants were significantly quicker on easy than on medium questions, $t(248) = 20.52$, $p < .001$, and quicker on medium than on difficult questions, $t(248) = 7.09$, $p < .001$. These response times on the answers again validated the categorization of the knowledge questions into three difficulty levels. Further, I replicated the findings of Study 1 and 2 on response times on the answers in the knowledge test. This analysis showed no gender effect, $F(1,243) = .05$, $p = .832$, and no effect of incentive, $F(1,243) = 1.03$, $p = .311$. None of the interactions reached significance, $Fs \leq 2.03$, $ps \geq .133$. The analysis implied that the RBCEM (Healy, 2016) did not affect responses on the knowledge questions or processes of answering the questions.

Next, I compared response times on correct answers with those on incorrect answers. A mixed model ANOVA with the between-subjects factor gender and incentive and the within-subjects factor correctness (correct vs. incorrect answers) yielded the postulated effect of correctness, $F(1,243) = 644.84$, $p < .001$, $\eta^2 = .73$ (see Figure 7.1). Participants gave correct answers quicker $M = 9677.62$ ms ($SD = 2642.84$ ms) than incorrect answers $M = 15498.92$ ms ($SD = 5374.05$ ms). There was no effect of gender, $F(1,243) < .01$, $p = .712$ or incentive, $F(1,243) = .64$, $p = .425$. There was a marginally significant three-
way interaction between correctness, gender and incentive, \( F(1, 243) = 2.88, p = .091 \). No other interaction reached significance, \( Fs \leq 1.68, ps \geq .196 \).

To analyze this interaction, I looked at females and males separately. For females, an ANOVA with the between-subjects factor incentive and the within-subjects factor correctness showed a significant effect of correctness, \( F(1, 126) = 398.93, p < .001 \), \( \eta^2 = .76 \). There was no effect of incentive, \( F(1, 126) = .35, p = .554 \), but a significant interaction between correctness and incentive, \( F(1, 126) = 5.37, p = .022 \), \( \eta^2 = .04 \). Subsequent \( t \)-tests showed that the response time for correct and incorrect answers differed significantly for the low incentive condition, \( t(66) = -13.61, p < .001 \), as well as for the high incentive condition, \( t(60) = -14.50, p < .001 \). In the low incentive condition, females took \( M = 10026.76 \text{ ms} \) (\( SD = 2932.80 \text{ ms} \)) for correct answers and \( M = 15361.30 \text{ ms} \) (\( SD = 6118.61 \text{ ms} \)) for incorrect answers. In the high incentive condition, females took \( M = 9211.93 \text{ ms} \) (\( SD = 2566.60 \text{ ms} \)) for correct answers and \( M = 15599.10 \text{ ms} \) (\( SD = 4880.98 \text{ ms} \)) for incorrect answers. This shows that for females the difference in response times between correct and incorrect answers is bigger in the high incentive condition than in the low incentive condition.

For males, an ANOVA with the between-subjects factor incentive and the within-subjects factor correctness revealed an effect of correctness, \( F(1, 117) = 263.12, p < .001 \), \( \eta^2 = .69 \). There was no effect of incentive, \( F(1, 117) = .29, p = .591 \) and no interactions between correctness and incentive, \( F(1, 117) = .07, p = .794 \). Then, I ran an ANOVA on response times on correct answers with the between-subjects factors gender and incentive. There was no effect of gender, \( F(1, 243) = .33, p = .568 \), incentive, \( F(1, 243) = 2.12, p = .146 \), or an interaction between the two, \( F(1, 243) = .90, p = .343 \). I ran the same ANOVA on response times on incorrect answers and found no significant effects (gender, \( F(1, 243) = .01, p = .908 \), incentive, \( F(1, 243) < .01, p = .949 \), interaction between the two, \( F(1, 243) = .52, p = .470 \)).

I further compared the incentive conditions separately. An ANOVA for the low incentive condition with the between-subjects factor gender and the within-subjects factor correctness showed an effect of correctness, \( F(1, 120) = 301.28, p < .001 \), \( \eta^2 = .72 \). There was no gender effect, \( F(1, 120) = .05, p = .824 \), or an interaction between gender and correctness, \( F(1, 120) = .86, p = .355 \). The same ANOVA for the high incentive condition revealed also an effect of correctness, \( F(1, 123) = 344.70, p < .001 \), \( \eta^2 = .74 \), but no gender effect, \( F(1, 123) = .09, p = .764 \), or an interaction between gender and correctness, \( F(1, 123) = 2.15, p = .146 \).
7.2.3. Confidence Judgments

The switch points in the RBCEM (Healy, 2016) were used as implicit measures for confidence in the correctness of an answer. A higher switch point indicated more confidence in the correctness of an answer since participants were not willing to switch from Option A (i.e., winning the incentive if the answer was correct) to Option B (winning the incentive with a certain probability). This meant that they were not willing to trade the probability of having a correct answer (Option A) for a lottery (Option B) that did not depend on the correctness of their answer but on chance.

The mean switch points were compared across gender, incentive, and item difficulty. In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factors gender and incentive, and the within-subjects factor item difficulty showed a highly significant main effect of item difficulty, $F(1.52, 369.08) = 216.84, p < .001, \eta^2 = .47$. On average, participants chose a rather high switch point on easy questions $M = 75.51$ ($SD = 18.68$), a lower one on questions with medium difficulty $M = 58.90$ ($SD = 14.26$), and finally, the lowest switch point on difficult questions $M = 52.82$ ($SD = 16.20$). $T$-tests confirmed that participants chose a significantly higher switch point on easy than on medium questions, $t(248) = 14.33,$

---

Figure 7.1.: Study 3: Response time on answers by correctness
p < .001, and a higher switch point on medium than on difficult questions, $t(248) = 7.78$, $p < .001$. The analysis revealed a significant gender effect, $F(1, 243) = 5.40$, $p = .021$, $\eta^2 = .02$ (see Figure 7.2). Males chose on average higher switch points $M = 64.28$ ($SD = 13.52$) than females $M = 60.77$ ($SD = 12.40$), that is they were more confident in the correctness of their answers as females. A significant difference between the low and high incentive condition occurred, $F(1, 243) = 6.66$, $p = .010$, $\eta^2 = .03$ (see Figure 7.2). Participants in the low incentive condition chose higher switch points in the RBCEM (Healy, 2016) than participants in the high incentive condition ($M_l = 64.34$; $SD_l = 13.02$; $M_h = 60.53$; $SD_h = 12.78$). This finding validated the RBCEM (Healy, 2016) and underlined its usefulness for measuring confidence in performance. It also illustrated that the manipulation of different incentive levels worked quite well and that the RBCEM (Healy, 2016) is quite sensitive towards differing incentive levels. None of the interactions reached significance, $Fs \leq 1.34$, $ps \geq .248$.

Following Koriat (Koriat, 2012a) I calculated the correlation between the confidence judgments and the response time on the answer to the knowledge questions. Contrary to the findings of Study 1 and 2, the correlation was positive and significant, $r = .14$, $p = .024$. I then calculated the correlation separately for females and males. The correlation was not significant for females, $r = .03$, $p = .702$, but highly significant for
males, $r = .23$, $p = .005$. As in Study 2, I followed up with the correlations for females and males in the high and low incentive condition. In the low incentive condition, the correlation for females was not significant, $r = -.003$, $p = .980$, but significant and positive for males, $r = .35$, $p = .009$. In the high incentive condition, the correlation for females was also not significant, $r = .06$, $p = .633$, but for males the correlation did also not reach significance, $r = .18$, $p = .154$. Hence, the positive relationship between the response time on the answers and the confidence judgments was driven by males in the low incentive condition. The positive correlation (in contrast to the negative correlation in Study 1 and 2) reflects that males in the low incentive condition did not rely on the mnemonic cue (the response time on the answers) to make their confidence judgments.

As in Study 1 and 2, I examined confidence in the correctness of an answer (the switch points in the RBCEM; Healy, 2016) for correctly and incorrectly answered questions separately. A mixed model ANOVA with the between-subjects factors gender (female vs. male) and incentive (low vs. high) and the within-subjects factor answer (correct vs. incorrect) showed a significant effect of the correctness of the answers on the switch points, $F(1, 247) = 265.54$, $p < .001$, $\eta^2 = .52$ (see Figure 7.3). Identical to the results of Study 1 and 2, the confidence in the correctness of an answer was higher for correctly answered questions $M = 70.78$ ($SD = 16.50$) than for incorrect answers $M = 49.56$ ($SD = 15.53$). This validated the paradigm and underlined that confidence judgments are evidence based and that on a meta-cognitive level, participants have a valid feeling of being correct.

There was also a significant gender effect, $F(1, 247) = 4.01$, $p = .046$, $\eta^2 = .02$, and a significant effect of the level of incentives, $F(1, 247) = 6.54$, $p = .011$, $\eta^2 = .03$ (see Figure 7.3). The main effect of gender implicated that both on correct and incorrect answers males ($M_{correct} = 72.54; SD_{correct} = 17.86; M_{incorrect} = 50.74; SD_{incorrect} = 15.48$) chose on average a higher switch point than females ($M_{correct} = 69.15; SD_{correct} = 15.00; M_{incorrect} = 48.47; SD_{incorrect} = 15.57$). This replicated in part the result of Study 2 but without an interaction between correctness and gender. The main effect of incentives revealed that participants in the low incentive condition were more confident on both correct ($M_{correct} = 71.69; SD_{correct} = 16.32$) and incorrect answers ($M_{incorrect} = 52.43; SD_{incorrect} = 14.54$) than participants in the high incentive condition ($M_{correct} = 69.90; SD_{correct} = 16.68; M_{incorrect} = 46.76; SD_{incorrect} = 16.01$). This ANOVA also yielded a weak three-way interaction between the correctness of an answer, gender, and incentive, $F(1, 247) = 3.03$, $p = .083$, $\eta^2 = .01$. None of the other interactions reached significance, $Fs \leq 1.95$, $ps \geq .164$.
To analyze the three-way interaction further, I first looked at females and males separately. For females, an ANOVA with the between-subjects factor level of incentive and the within-subjects factor correctness revealed an effect of correctness, $F(1, 126) = 165.43$, $p < .001$, $\eta^2 = .57$. There was no effect of level of incentive, $F(1, 126) = 2.30$, $p = .132$, but a significant interaction between correctness and level of incentive, $F(1, 126) = 6.41$, $p = .013$, $\eta^2 = .05$. Subsequent $t$-tests showed that for the low incentive condition the confidence judgments for correct and incorrect answers differed significantly, $t(66) = 8.71$, $p < .001$. The average confidence for correct answers was $M = 68.72$ ($SD = 14.50$) and for incorrect answers $M = 51.96$ ($SD = 13.21$). For the high incentive condition the confidence judgments for correct and incorrect answers also differed significantly, $t(60) = 9.37$, $p < .001$. The average confidence judgment for correct answers was $M = 69.61$ ($SD = 15.64$) and for incorrect answers $M = 44.63$ ($SD = 17.10$). When comparing the confidence judgments for correct answers for the low and high incentive condition, a $t$-test showed no significant difference, $t(126) = -.33$, $p = .740$. However, there was a significant difference in the confidence judgments for incorrect answers between the low and the high incentive condition, $t(126) = 2.73$, $p = .007$. Participants in the low incentive condition judged their confidence on incorrect answers on average higher, $M = 51.96$ ($SD = 13.21$) than participants in the high incentive condition, $M = 44.63$. 

77
(SD = 17.10). This result illustrated that females in the low incentive condition did care less about being accurate in their confidence judgments in particular on incorrect answers as they had less to lose than participants in the high incentive condition in case they were not accurate. Put differently, the low incentive made females more risk seeking in case of incorrect answers than participants in the high incentive condition.

For males, an ANOVA with the between-subjects factor level of incentive and the within-subjects factor correctness showed an effect of correctness, $F(1, 117) = 109.76$, $p < .001$, $\eta^2 = .48$, and an effect of level of incentive, $F(1, 117) = 4.36$, $p = .039$, $\eta^2 = .04$, but no significant interaction between correctness and level of incentive, $F(1, 117) = .05$, $p = .830$. The average confidence judgment on correct answers was $M = 72.54$ (SD = 17.86) and on incorrect answers $M = 50.74$ (SD = 15.48). The average confidence judgment in the low incentive condition was $M = 67.56$ (SD = 13.43) and in the high incentive condition $M = 61.46$ (SD = 13.05).

To analyze the three-way interaction further I ran an ANOVA on confidence judgments on correct answers with the between-subjects factors gender and level of incentive. There was a weak effect of gender, $F(1, 243) = 2.91$, $p = .089$, $\eta^2 = .01$. Females judge their confidence on correct answers on average $M = 69.15$ (SD = 15.00) and males $M = 72.54$ (SD = 17.86). There was no effect of level of incentive, $F(1, 243) = 1.02$, $p = .314$, or an interaction between the two, $F(1, 243) = 2.05$, $p = .153$. I ran the same ANOVA on confidence judgments on incorrect answers and found no significant gender effect, $F(1, 243) = 1.78$, $p = .183$, but a significant effect of level of incentive, $F(1, 243) = 8.76$, $p = .003$, $\eta^2 = .04$. In the low incentive condition, the average confidence judgment on incorrect answers was $M = 52.35$ (SD = 14.51) and in the high incentive condition $M = 46.64$ (SD = 16.00). As mentioned in the previous analysis (separate for females and males) this effect is driven by female participants. The interaction between gender and level of incentive did not reach significance, $F(1, 243) = .64$, $p = .424$.

I further compared the incentive conditions separately. An ANOVA for the low incentive condition with the between-subjects factor gender and the within-subjects factor correctness showed an effect of correctness, $F(1, 120) = 119.32$, $p < .001$, $\eta^2 = .50$, a weak gender effect, $F(1, 120) = 3.15$, $p = .078$, $\eta^2 = .03$, but no interaction between the two, $F(1, 120) = 2.39$, $p = .125$. The mean confidence judgment on correct answers was $M = 71.57$ (SD = 16.30) and on incorrect answers $M = 52.35$ (SD = 14.51). Females judged their confidence on average $M = 61.88$ (SD = 12.19) and males $M = 67.56$ (SD = 13.43). An ANOVA for the high incentive condition with the between-subjects factor gender and the within-subjects factor correctness showed an effect of correctness,
$F(1,123) = 146.96, \ p < .001, \ \eta^2 = .54$, no gender effect, $F(1,123) = 1.15, \ p = .287$, and no interaction between the two, $F(1,123) = .88, \ p = .349$. The average confidence judgment on correct answers was $M = 69.92$ ($SD = 16.62$) and on incorrect answer $M = 46.64$ ($SD = 16.00$).

7.2.4. Response Times on the Confidence Judgments

I logarithmized the response times on the confidence judgments indicated by the switch points in the RBCEM (Healy, 2016) since they were skewed. I compared the response times on the choices in the RBCEM across gender, incentive, and item difficulty. In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factors gender, incentive, and the within-subjects factor item difficulty yielded a significant main effect of item difficulty, $F(1.83, 445.61) = 264.28, \ p < .001, \ \eta^2 = .52$. Participants judged their confidence on easy questions in $M = 11887.82$ ms ($SD = 3703.36$ ms), on questions with medium difficulty in $M = 15955.98$ ms ($SD = 4789.13$ ms), and on difficult questions in $M = 17489.63$ ms ($SD = 5868.97$ ms). T-tests confirmed that participants judged their confidence quicker on easy than on medium questions, $t(248) = 16.54, \ p < .001$, and quicker on medium than on difficult questions, $t(248) = 5.74, \ p < .001$. These findings approved that the RBCEM (Healy, 2016) was a valid measure of confidence. There was no gender effect, $F(1,243) = .33, \ p = .568, \ \eta^2 < .01$, and no incentive effect, $F(1,243) = 1.33, \ p = .249, \ \eta^2 < .01$. None of the interactions reached significance, $Fs \leq 1.07, \ ps \geq .343$.

Then I analyzed the response time on the confidence judgments for correct and incorrect answers. I ran an ANOVA with the between-subjects factors gender and level of incentive and the within-subjects factor correctness. The analysis revealed an effect of correctness, $F(1,243) = 191.06, \ p < .001, \ \eta^2 = .44$. The average response time on the confidence judgments for correct answers was $M = 13668.82$ ms ($SD = 3717.04$ ms) and on incorrect answers $M = 17367.14$ ms ($SD = 5959.89$ ms). This result supports the assumption that participants have overall a good monitoring resolution and hence are able to discriminate between correct and incorrect answers. Interestingly, the standard deviation of response times on confidence judgments was also much higher for incorrect than for correct answers. One could conclude that in case of incorrect answers, response times are much noisier in terms of interindividual variances as in case of correct answers. There was no gender effect, $F(1,243) = .04, \ p = .842$, no effect of level of incentive, $F(1,243) = 1.43, \ p = .232$. None of the interactions reached significance, $Fs \leq 1.72, \ ps \geq .191$. Surprisingly, the level of incentive did not impact this pattern of response times on correct vs. incorrect answer.
The effects are illustrated in Figure 7.4. I then calculated the correlation between the response time on the answers and the response time on the confidence judgments. As in Study 1 and 2, the correlation was positive and highly significant, \(r = .87, p < .001\). This illustrates that there is a strong link between the response times for the answers to the knowledge questions and the corresponding response time for the confidence judgments.

![Figure 7.4: Study 3: Response time on confidence judgments by correctness and incentive level](image)

### 7.2.5. Confidence Bias

The switch points ranged from 0 to 100 like the confidence rating scales in Study 1 and 2. A low switch point indicated low confidence whereas a high switch point indicated high confidence. I therefore calculated the confidence bias using the switch points as confidence measure in the same manner as in Study 1 and 2.

The confidence bias was compared between gender, incentive, and item difficulty. In case of a violation of sphericity, the Greenhouse-Geisser correction was used. A mixed model ANOVA with the between-subjects factors gender and incentive (low vs. high incentive), and the within-subjects factor item difficulty revealed the predicted effect of item difficulty, \(F(1.78, 431.33) = 189.70, p < .001, \eta^2 = .44\). Like in Study 1 and 2, participants were underconfident on easy questions, \(M = -10.54 \ (SD = 19.62)\),
rather accurate in their confidence on the correctness of answers on the questions of medium difficulty, $M = 2.18$ ($SD = 18.20$), and were overconfident on difficult questions, $M = 20.48$ ($SD = 20.20$). $t$-tests suggested that the confidence bias significantly differed from zero for easy questions, $t(248) = -8.57$, $p < .001$, and for difficult questions, $t(248) = 15.87$, $p < .001$, but in opposite directions. The confidence bias for medium questions did not differ from zero, $t(248) = 1.79$, $p = .075$, and hence indicated accurate confidence judgments. Overall, the hard-easy effect (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982) was again replicated (see Figure 7.5). These findings asserted that the RBCEM (Healy, 2016) was a valid measure of confidence as it provided similar results as the confidence rating scales of Study 1 and 2. There was no gender effect, $F(1, 243) = .03$, $p = .856$, and no incentive effect, $F(1, 243) = 1.23$, $p = .268$. None of the interactions reached significance, $Fs \leq 1.97$, $ps \geq .146$. To determine the direction of the confidence bias, $t$-tests of males’ and females’ confidence biases were computed. They revealed overconfidence in both genders, the confidence bias differed significantly from zero for males, $t(118) = 3.94$, $p < .001$, and for females, $t(127) = 3.18$, $p = .002$ (see Figure 7.6).

Figure 7.5.: Study 3: The hard-easy effect

<table>
<thead>
<tr>
<th></th>
<th>Confidence Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td>female</td>
</tr>
<tr>
<td>medium</td>
<td>female</td>
</tr>
<tr>
<td>difficult</td>
<td>female</td>
</tr>
</tbody>
</table>
To investigate the effects of gender, incentive and response times on the confidence bias, I ran hierarchical regressions. All variables were z-standardized, except for the response times on confidence judgments which were logarithmized (due to their skewness). Gender (male = 1, female = 0) and incentive (high incentive = 1, low incentive = 0) were dummy-coded. Hierarchical regressions on the confidence bias are presented in Table 7.1. Model 1 accounted for the main variables, i.e. gender and incentive. Incentive and gender were not significant predictors of the confidence bias in Model 1. Model 2 added response times on confidence judgments as a predictor. The response time on confidence judgments did not predict the confidence bias. Model 3 added the interaction between incentive and response time on confidence judgments as predictor which was also not significant. I calculated the correlation between response times on the confidence judgments and the confidence bias separately for males and females as in Study 1 and 2. The correlation for males was positive but not significant, $r = .15$, $p = .101$ (females: $r = -.06$, $p = .486$).
Table 7.1.: Study 3: Linear regressions on confidence bias

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male = 1)</td>
<td>.02</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>(.13)</td>
<td>(.13)</td>
<td>(.13)</td>
</tr>
<tr>
<td>Incentive for accuracy (high = 1)</td>
<td>-.14</td>
<td>-.14</td>
<td>-.61</td>
</tr>
<tr>
<td></td>
<td>(.13)</td>
<td>(.13)</td>
<td>(4.87)</td>
</tr>
<tr>
<td>Response time on confidence judgments</td>
<td>.08</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.25)</td>
<td>(.35)</td>
<td></td>
</tr>
<tr>
<td>Incentive × Response time on confidence judgments</td>
<td>.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.51)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>.07</td>
<td>-.69</td>
<td>-.47</td>
</tr>
<tr>
<td></td>
<td>(.11)</td>
<td>(2.44)</td>
<td>(3.39)</td>
</tr>
<tr>
<td>N</td>
<td>249</td>
<td>249</td>
<td>249</td>
</tr>
<tr>
<td>R²adj.</td>
<td>-.003</td>
<td>-.01</td>
<td>-.01</td>
</tr>
<tr>
<td>F</td>
<td>.64</td>
<td>.46</td>
<td>.34</td>
</tr>
</tbody>
</table>

Notes. Standard errors in brackets. *p < .10, **p < .05, ***p < .01.

7.2.6. Exploratory Analysis of the Confidence Bias on Questions of Medium Difficulty

The analysis of the hard-easy effect revealed no incentive effect or interaction of incentive and item difficulty. But the descriptive statistics indicated that participants in the high and low incentive condition differed in their confidence bias on questions of medium difficulty. Participants in the high incentive condition showed a confidence bias of $M = -0.54$ ($SD = 18.17$) compared to $M = 4.72$ ($SD = 17.87$) in the low incentive condition. Therefore, I ran an exploratory analysis on the confidence bias of questions of medium difficulty. It delivered a significant effect of the incentive, $F(1, 243) = 5.59$, $p = .019$, $\eta^2 = .02$. The participants in the high incentive treatment were well calibrated and judged their performance quite accurately. Yet, participants in the low incentive treatment displayed overconfidence on questions of medium difficulty (see Figure 7.7).

$T$-tests confirmed that the confidence bias in the high incentive condition did not differ from zero, $t(125) = .33$, $p = .741$, whereas the confidence bias in the low incentive
condition deviated significantly from zero, \( t(122) = 2.93, p = .004 \). The analysis revealed no gender effect, \( F(1, 243) = .67, p = .414 \), and no interaction between gender and incentive, \( F(1, 243) = .40, p = .527 \). Thus, the high incentive supported primarily a good calibration of the confidence judgments on questions of medium difficulty.

Figure 7.7.: Study 3: Confidence bias on questions of medium difficulty by incentive level

7.3. Discussion

I replicated the findings of the pretest and of Study 1 and 2 on the difficulty levels of the knowledge questions. Even when offering incentives for a good calibration of the confidence judgments through the RBCEM (Healy, 2016), the categorization of the questions into three levels of difficulty remained valid. All participants answered questions previously categorized as easy questions more often correctly than questions categorized as questions of medium difficulty, while questions categorized as difficult in the preceding studies were least often answered correctly. Like in Study 2, males performed better than females in the knowledge test. Further, I found an interaction between the incentive for accuracy of the confidence judgment and the item difficulty on performance. Participants in the low incentive condition performed better on difficult questions compared to the
high incentive condition. There was no difference in performance for easy questions and questions of medium difficulty. Why varying incentives on the accuracy of confidence judgments impacted performance in the knowledge test at all remained unclear.

One should keep in mind that the participants were not aware whether a question was easy, medium or difficult, because they were not told that there were three difficulty levels. Further, the knowledge questions were presented randomly, that is people did not learn somehow that the difficult level might increase or decrease from question to question if they might have an implicit hypothesis about difficulty levels. Considering ideas about meta-cognitive processes of confidence judgments, one could argue that participants in the low incentive condition took it easier to be confronted with a difficult question, because they did not feel as much under pressure as participants in the high incentive condition to be accurate on the confidence judgments. For this reason, they might have had more cognitive resources available to retrieve the correct answer from long-term memory than participants in the high incentive condition. This explanation refers to research on choking under pressure (Baumeister & Showers, 1986; Mobbs et al., 2009). It describes that in case of a strong motivation to perform well, people might underperform. But the incentive was supposed to improve the accuracy of confidence judgments and not to increase performance in the knowledge test, thus this explanation remains speculative.

As in Study 1 and 2, the categorization of knowledge questions into three levels of difficulty was additionally supported by a main effect of item difficulty on the response time on the answers in the knowledge test. Response times were longer for answering difficult questions than for answering questions of medium difficulty and easy questions. The latter were more quickly answered as medium and difficult questions. These results were observed for both gender and both incentive conditions. Hence, measuring confidence through the RBCEM (Healy, 2016) and introducing different incentives did not influence response times patterns on the knowledge questions or processes of answering them.

Confidence was measured through the switch points in the RBCEM (Healy, 2016). The higher the switch point, the higher people’s confidence that their answer on a knowledge question was correct. On this variable, I first replicated Study 1 and 2 in that the difficulty level of a knowledge question determined confidence judgments. Switch points were higher for easy questions than for questions of medium difficulty and for difficult questions. They were also higher for medium questions compared to difficult questions. The response time on the confidence judgment supported this pattern. Participants judged their confidence quickest on easy questions followed by medium questions and
they took longest for difficult questions. Thus, Hypotheses 1a and b were confirmed by
the results. In addition, these findings proved that the RBCEM (Healy, 2016) was a
valid measure of confidence in own performance.

It also turned out that males displayed overall higher confidence than females. Further,
the confidence judgments were lower in the high incentive condition compared to the low
incentive condition. This finding implied that the RBCEM (Healy, 2016) was sensitive
towards differing levels of incentives. It also suggested that the low incentive resulted
in higher risk taking compared to the high incentive. People were more cautious with
choosing their switch points in the high incentive condition as in the low incentive
condition. Note that this effect was independent of item difficulty and hence can not
only be explained by the performance difference on difficult questions between the high
and low incentive condition.

When analyzing the confidence bias, I replicated the hard-easy effect observed in Study
1 and 2 which confirmed Hypothesis 2. Participants displayed underconfidence on easy
questions, overconfidence on difficult questions and were accurate on questions of medium
difficulty. Again, this validates the use of the RBCEM (Healy, 2016) as implicit measure
of confidence.

In line with the results of Study 1 and 2, I found higher confidence judgments (switch
points) on correct answer compared to incorrect answer. This result further validated
the RBCEM (Healy, 2016) as a measurement of confidence as it replicated the results of
Study 1 and 2 on confidence judgments on correct vs. incorrect answers. The finding
confirmed Hypothesis 3a that people have quite a good monitoring resolution (ability to
discriminate between correct and incorrect answers).

Note that the higher confidence judgments in the low incentive condition compared
to the high incentive condition were observed for correct and incorrect answers. One
could argue that this effect was due to the performance difference on difficult items of the
two incentive conditions (participants in the low incentive condition gave more correct
answers on difficult items than participants in the high incentive condition; see above),
but participants in the low incentive condition also judged their confidence for incorrect
answers higher than those in the high incentive condition over all three difficulty levels.
This finding partly supported the Hypothesis 3b that only the high, but not the low
incentive for accurate confidence judgments increased the sensitivity to evidence for being
correct or incorrect.

Interestingly, males’ higher confidence judgments compared to females’ was evident
for both correct and incorrect answers. In this sense males were overall more confident
than females which is not only explained by a better performance since this was also true for incorrect answers. A weak three-way interaction between gender, correctness and incentive showed that for females the confidence in incorrect answers was higher in the low incentive condition than the high incentive condition. I found no difference between incentive conditions for males.

The results on the response times supported the notion of a good monitoring resolution. The response time on correct answers was significantly quicker than on incorrect answers. In addition, the results showed that the difference between the response times on correct and incorrect answers was bigger in the high incentive condition compared to the low incentive condition for females. This indicates that females were more cautious in answering the knowledge questions in the high incentive condition, because the stakes were higher. The response times on the confidence judgments reflected the correctness of the answer as in Study 1 and 2. The response times on the confidence judgments on correct answers were quicker than on incorrect answers. Hence Hypothesis 3c was confirmed. Contrary to the prediction of Hypothesis 3d, this was independent of the incentive level. The high incentive did not increase the time on incorrect answers and decrease the time on correct answers compared to the low incentive condition. One explanation could be that compared to Study 2, in this study the level of incentive was varied. Hence, there was no condition in which the accuracy of the confidence judgment was not incentivized at all. The low incentive could already have increased the response time on the confidence judgments and the higher incentive in the high incentive condition did not prolong the response times significantly compared to the low incentive. The response time on the confidence judgments was also independent of gender.

As in Study 1 and 2, I tested Koriat’s (2012a) assumption that confidence judgments are based on mnemonic cues (response time on the answers). Surprisingly, the correlation between the confidence judgments and the response time on the answers was positive (compared to the negative correlation in Study 1 and 2) which contradicted Hypothesis 4. Further analysis showed that this correlation was only significant for males in the low incentive condition. The question why this correlation was only found for this subgroup remains unclear, but it still indicates that participants in that subgroup did not base their confidence judgments on mnemonic cues (response times on the answers). Overall, it can be said that there was no negative relationship between the response time on the answers and the confidence judgments, which indicates that the confidence judgments were based on different information than in Study 1 and 2. As in Study 1 and 2, I found a highly significant, positive correlation between the response times on the answers and
the response times on the confidence judgments which confirmed Hypothesis 1c. This finding seems to be quite robust and points to a strong link between these two processes.

In contrast to Study 2, there was no gender difference in the confidence bias. Males and females both displayed overconfidence over all difficulty levels and independent of the incentive level which contradicted Hypothesis 5. The incentivization of the accuracy of the confidence judgments prevented females from being underconfident and even turned them overconfident. On the other hand, this result suggested that even the high incentive on accurate confidence judgements did not eliminate males’ overconfidence in their performance, observed in Study 2 when performance was incentivized, but this confidence bias of males remained in Study 3.

There was no effect of the incentive levels on the confidence bias which disproved Hypothesis 6a. In both the low and the high incentive condition participants were overconfident and hence missed the aim of judging their performance accurately. One possible explanation could be that the incentive (high and low) increased the post-decisional information accumulation but this information was not unbiased. Since the monetary incentive (in both conditions) depended on the choice participants made between a lottery and the correctness of their answer, participants were rather inclined to bet on themselves than leave it to chance. This result is additionally supported by the response time data of both incentive conditions. Participants took overall approximately three seconds longer to judge their confidence (choose a switch point: $M = 15290$ ms, $SD = 4422$ ms) than participants in Study 2 ($M = 11894$ ms, $SD = 3850$ ms) and Study 1 ($M = 11713$ ms, $SD = 3004$ ms). Further, the standard deviations of the response times on the confidence judgments were quite high in the present study. This hindered statistically significant differences between the incentive conditions on response times which disproved Hypothesis 6b. Related to this observation, I did not replicate the finding of Study 1 and 2 that male participants who took longer for their confidence judgment were less overconfident than males who quickly judged their confidence. Thus, Hypothesis 7 was not confirmed. Again, this could be as all participants in Study 3 were rather slow in their confidence judgments and the entire sample had quite big standard deviations of their response times on the switch points.

Since earlier research on the impact of incentives on confidence judgments did not vary the item difficulty (see Koriat et al., 2006; Lebreton et al., 2018; Sanchez & Dunning, 2018), I took a closer look at the separate difficulty levels. Based on the descriptive statistics, I scrutinized the questions of moderate difficulty separately. I found that the high incentive reduced overconfidence compared to the low incentive on these questions.
so strongly, that participants in the high incentive condition were perfectly accurate on questions of moderate difficulty. In contrast, participants in the low incentive condition were overconfident on these questions. The low incentive therefore was not motivating enough to overcome overconfidence. I concluded that offering a high incentive for accurate confidence judgments led to an increase in the accuracy of this judgment primarily if the knowledge question was neither easy, nor difficult, but when the chance of correctly answering the question was between 50 and 75% (i.e., medium difficult questions).

These findings deviated from results of Lebreton et al. (2018). They did not manipulate varying levels of item difficulty but presented items (visual contrasts) with a constant level of difficulty. Lebreton et al. (2018) varied the frame of the incentive for accurate confidence judgments (gain vs. loss). Participants were more overconfident if they could win an incentive for being accurate as if they lost money when not being accurate. For the gain frame, they found that the higher the magnitude of the incentive (10 Cents vs. 1€ vs. 2€) the more overconfident the participants. A possible explanation for these contrary findings could be that their observations are mainly valid for visual tasks with two alternative choices, but not for general knowledge questions with five answer options.
8. General Discussion

The aim of the present research was to investigate the processes of confidence judgments on performance and the respective confidence bias. More specifically, I focused on gender differences in confidence judgments and the underlying processes as well as the potentially beneficial effect of monetary incentives on the confidence bias. In order to fully understand the effect of monetary incentives on confidence judgments, I scrutinized monetary incentives that were based on the performance in the task and monetary incentives that were based on the accuracy of the confidence judgments. Previous research on confidence judgments and the confidence bias has demonstrated that gender (Larrick et al., 2007; Soll & Klayman, 2004) and item difficulty (Brenner, 2003; Grieco & Hogarth, 2009; Griffin & Tversky, 1992; Larrick et al., 2007; Fischhoff et al., 1977; Merkle, 2009; Moore & Healy, 2008) may distort results on confidence judgments. Consequently, I pretested my research materials and selected gender-neutral general knowledge questions with varying item difficulty (easy, medium, difficult).

Over three studies, I observed the hard-easy effect (Brenner, 2003; Grieco & Hogarth, 2009; Griffin & Tversky, 1992; Larrick et al., 2007; Fischhoff et al., 1977; Merkle, 2009; Moore & Healy, 2008). Participants were underconfident on easy questions, were quite accurate in their judgment of their performance on questions of medium difficulty and were overconfident regarding their performance on difficult questions. This effect was consistent despite some variations of the design in each study. For instance, in Study 1, no incentive was introduced, and all participants were paid flat rate. This finding suggested that the emergence of the hard-easy effect did not depend on any kind of incentivization. In Study 2, the incentive was provided for performing well in the knowledge test and in Study 3 the incentives were offered for accurate confidence judgments. In both studies, the hard-easy effect occurred as in Study 1. These findings and the fact that the data were collected at two different universities and with rather big sample sizes (Study 1: N = 123; Study 2: N = 255; Study 3: N = 249) suggest that the hard-easy effect (Brenner, 2003; Grieco & Hogarth, 2009; Griffin & Tversky, 1992; Larrick et al., 2007; Fischhoff et al., 1977; Merkle, 2009; Moore & Healy, 2008) is a quite robust phenomenon. I claim that
the occurrence of the hard-easy effect could not be attributed to a biased selection of test questions as argued by Gigerenzer, Hoffrage, and Kleinbölting (1991). The selection of the general knowledge questions in our research was based on a pretest \(N = 71\) in which performance on 84 questions was examined and tested whether a gender bias occurred for performance on each of the items. The questions used in all three studies hence were not biased in the sense of Gigerenzer, Hoffrage, and Kleinbölting (1991) that researchers are prone to choose harder-than-normal questions for experiments on overconfidence. Juslin, Winmann, and Olson (2000) argued that the existence of the hard-easy effect cannot be proven by two-alternative general knowledge questions. For this reason, I used multiple choice questions with five options for the present research. Hence, the hard-easy effect observed in the present studies cannot be attributed to a faulty item selection.

The categorization of questions in difficulty levels was not only supported by the occurrence of the hard-easy effect, but also by the process data (response times). Over all three studies, the response times on the answers as well as the response times on the confidence judgments reflected the item difficulty (see also Wixted & Mickes, 2010). Participants answered easy questions faster than questions of medium difficulty and the latter faster than difficult questions. More interestingly, the same pattern was found for the response times on the confidence judgments. Participants judged their confidence on easy questions quickest, followed by questions of medium difficulty and they were slowest on difficult questions. The variation of response time on the confidence judgments depending on item difficulty speaks for a dual stage account of confidence (see Moran et al., 2015). Dual stage models assume that confidence is not only affected by the initial choice but also by novel information collected after the initial choice (in this case choosing an answer to the knowledge question). The effect of item difficulty shows that the evidence accumulation for the confidence judgment is ongoing after the initial choice (the answer to the knowledge question) is made. Hence, this contradicts single stage models of confidence (see for example Koriat, 2012a) which assume that confidence is based on the same information that underlies the initial decision. Furthermore, the effect of item difficulty on the response time on the confidence judgment supports the notion that people have overall a good monitoring resolution meaning they can discriminate between correct and incorrect answers. When the probability of being incorrect was higher (on difficult questions), participants took longer to select an answer and to judge their confidence than when they were very likely correct (on easy questions). In all three studies, I found a highly significant positive correlation between the response time on the answers and the response time on the confidence judgments (see also Moran et al.,
2015 for similar results). This finding supports the idea that these two processes are strongly linked and that the process of confidence formation does not solely depend on the process of the initial decision.

In order to investigate the processes underlying confidence judgments, I examined participants’ monitoring resolution meaning the discrimination between correct and incorrect answers. I investigated monitoring resolution in all three studies and examined the influence of incentives (Study 2 and 3) on the monitoring resolution. The assumption that people have overall a good monitoring resolution was supported by the analysis of confidence judgments on correct and incorrect answers (Koriat, 2012a, 2018; Lebreton et al., 2018). In all three studies, participants judged their confidence on correct answers significantly higher than on incorrect answers. Hence, they had overall a good feeling whether their answer was correct or incorrect. The analysis of the process data confirmed the findings on the monitoring resolution. In all three studies, participants gave correct answers significantly faster than incorrect answers. More interestingly, the response times on the confidence judgments showed the same pattern. The confidence on correct answers was judged significantly faster than on incorrect answers. This finding is supported by the collapsing confidence boundary model (CCB) proposed by Moran et al. (2015). The authors argue that the confidence judgment process describes a tradeoff between the desire of high confidence and the time that needs to be invested for evidence accumulation to support such a high confidence level. In case of being correct, participants need less time to accumulate evidence for the correctness of their answer than in case of being incorrect.

In addition to the general ability to discriminate correct and incorrect answers, I also investigated the effect of incentives (on performance in Study 2 and on accuracy of the confidence judgment in Study 3) on the monitoring resolution. Based on the findings of Lebreton et al. (2018), I expected the incentive for accurate confidence judgments (Study 3) to improve the monitoring resolution. In Study 2, I found no effect of incentives for performance on the monitoring resolution. In line with results by Lebreton et al. (2018), the incentive for performance did not increase the discrimination between correct and incorrect answers by increasing (or decreasing) the confidence in the answers. Nonetheless, the incentive for performance increased the response time on the confidence judgments. Following the CCB model (Moran et al., 2015) the longer response time in the incentive condition should have resulted in more evidence accumulation which in turn should have increased the monitoring resolution. Moran et al. (2015) demonstrated that evidence accumulation after the initial choice improved the monitoring
resolution. One key difference between the present investigation and the experiments conducted by Moran et al. (2015) and also Lebreton et al. (2018) is the type of task used to investigate confidence judgments. Moran et al. (2015) (as well as Lebreton et al., 2018) used two alternative forced-choice perceptual tasks whereas the present work investigated confidence judgments in multiple-choice general knowledge questions. The information that can be accumulated after the initial choice in a perceptual task might contain more helpful information than in general knowledge questions. It is possible that general knowledge questions pose more of a ceiling-effect (see Campbell, Crumbaugh, Knouse, & Snodgrass, 1970) for evidence accumulation after the initial choice than perceptual tasks. In Study 3 in which accurate confidence judgments were incentivized, I also did not find an effect of incentive (high vs. low) on the monitoring resolution which stands in contrast to the results of Lebreton et al. (2018). Participants in the high incentive condition did not display a better monitoring resolution than participants in the low incentive condition. There are two possible reasons for this. First, as discussed before, the response time on the confidence judgments did not differ between the high and low incentive condition. Hence, participants spent the same amount of time to judge their confidence independent of the incentive level. It is possible that the low incentive already encouraged participants to invest more time in the evidence accumulation so that there was no difference found between the two incentive conditions. Secondly, the type of task used in the present work (general knowledge questions) might not be as sensitive to prolonged response times on the confidence judgments. More specifically, there is less evidence to accumulate after the initial choice compared to a perceptual task. I also investigated gender differences in monitoring resolution since there is already some evidence that confidence judgments and confidence formation might differ between females and males (see Barber & Odean, 2001; Hügelschäfer & Achtziger, 2014; Pulford & Colman, 1997). In Study 1, there was no difference between females and males in their monitoring resolution. In Study 2, I found that males judged their confidence higher than females especially on incorrect answers. It remains an open question whether males were less motivated to collect more evidence for their confidence judgment or were less able. In Study 3, males judged their confidence overall (on correct and incorrect answer) higher than females which could also be explained by males' higher performance on the knowledge questions. The results of the present work on potential gender differences in the monitoring resolution are not robust enough to result in a clear conclusion. Nevertheless, they indicate a potential gender difference which ought to be investigated in future research.

Based on the assumptions of the self-consistency model of subjective confidence (Koriat,
2012a), I investigated the relationship between the response time on the answers and the confidence judgments. Following Koriat’s (2012a) assumption, confidence judgments are based on mnemonic cues which is information that is derived from the process of the task (in this case the mnemonic cue is the response latency). If confidence judgments are in fact experience-based (see Koriat, 2012a) then confidence should be higher when the answer to the question is given quickly and low when it took a long time to retrieve the answer from memory. In Study 1, I found a negative correlation between the response time on the answers and the level of the confidence judgments for male participants (but not for females). In Study 2, the correlation was significant for females and males but only in the incentive condition. The same relationship was not found in participants of the flat rate payment condition. In Study 3, the correlation was surprisingly positive but only significant for males in the low incentive condition. The results of Study 1 and 2 indicate that the process information from the task (meaning the first decision) informs the confidence judgment. In Study 3, this relationship was not found but rather invers (for males in the low incentive condition). The longer male participants in the low incentive condition took to choose an answer to the knowledge question the more confident they were in its correctness. The results on the relationship between the response time on the answers and the level of confidence judgments indicate that the process information of the task informs the confidence judgment in some way, but the result is not as robust as to confirm the assumption that confidence judgments are exclusively based on mnemonic cues.

Next to the monitoring resolution, I was interested in the calibration meaning the absolute discrepancy between confidence and performance. I investigated whether biased judgments (overconfidence or underconfidence) were based on automatic processes meaning they were the default response. Following from this hypothesis, the automatic process would have to be overwritten by controlled processes to achieve accurate confidence judgments. In addition, I scrutinized whether the processes differed for males and females and whether incentives influenced these processes. First, I examined the relationship between the response time on the confidence judgments and the confidence bias. In Study 1, I found that the response times on the confidence judgments negatively predicted the confidence bias. When calculating the correlation between the response time on the confidence judgments and the confidence bias, I found it only significant for males. A median split showed that males who judged their confidence quickly were rather overconfidence whereas males who took more time were accurate in their judgment. In Study 2, the result was replicated but only for incentivized males. In addition, the
correlation was significant for females. However, the median split for females revealed that females who took longer to judge their confidence were rather underconfident. In contrast, females who judged their confidence quickly were quite accurate. In Study 3, I did not find a relationship between the response time on the confidence judgments and the confidence bias which could be explained by the overall prolonged response times on the confidence judgments due to the implemented paradigm (RBCEM, Healy, 2016). This finding highlights the importance to discriminate between a pre- and post-decisional phase of confidence judgments (see Moran et al., 2015) and especially separately for females and males. The information accumulated after the initial decision clearly differs between females and males. Apparently, males accumulate more evidence in favor of a realistic confidence judgment whereas females tend to accumulate biasing information which harms their accuracy. Navajas, Bahrami and Latham (2016) claim that the accuracy of confidence judgments depends on the accuracy of the accumulated evidence in the post-decisional phase. The difference between genders has important implications for potential interventions to increase accuracy in confidence judgments. Additional time for the confidence judgment seems to be much more beneficial for males whereas it is important to avoid females’ tendency to accumulate biasing information. Especially since females tend to collect information that leads them to doubt themselves.

Since a biased confidence judgment results in unfavorable effects (see Barber & Odean, 2001; Goel & Thakor, 2008; Malmendier & Tate, 2005, 2008; Zant & Moore, 2013; Camerer & Lovallo, 1999), I was interested in the potentially beneficial effect of incentives on the confidence bias. I claimed that monetary incentives should increase the probability of controlled processes (see Achtziger & Alós-Ferrer, 2014). More specifically, providing an incentive should result in the investment of more time and effort, and the deployment of attention on the incentivized task. Especially in case of incentivizing accurate confidence judgments, the longer time invested in post-decisional evidence accumulation should result in more accurate confidence judgments (see Moran et al., 2015).

As expected, the incentive for performance in Study 2 did not decrease the confidence bias. Males displayed overconfidence whereas females displayed underconfidence. This was not surprising since the incentive in Study 2 depended only on the performance on the task and not the accuracy of the confidence judgment. However, the analysis of the process data showed that the incentive for performance increased the response time on the confidence judgments. The prolonged response time on the confidence judgment should have resulted in more post-decisional evidence accumulation which in turn would have improved the confidence judgment. It is possible, that the incentive for performance
increased participants’ diligence overall (which is also seen in the increased response time on the answers in the incentive condition), but the focus was not on the accuracy of the confidence judgment. In Study 3, there was no difference in the confidence bias between the low and the high incentive condition. Contrary to my expectations, incentivizing the accuracy of the confidence judgement did not increase accuracy. Participants in both conditions displayed overconfidence. This finding is in line with the results of Sanchez and Dunning (2018) as well as Lebreton et al. (2018). Surprisingly, males as well as females displayed overconfidence. Incentivizing the confidence judgment encouraged female participants to be more confident and lead them to rely more (or even too much) on the correctness of their answer. This is an interesting finding, because it indicates that females are able to correct (or even overcorrect) their tendency to underestimate their abilities when money is at stake. In an exploratory analysis, I demonstrated that the high incentive increased accuracy on questions of medium difficulty whereas the low incentive did not have such a beneficial effect. This has important implications on the implementation of incentives to increase accuracy in confidence judgments. Their effectiveness seems to depend on the difficulty of the task and cannot be generalized to all levels of item difficulty.

To conclude, the present research demonstrated that there is strong evidence for dual stage accounts of confidence (see Moran et al., 2015). Confidence judgments are informed by additional information after the initial choice is made. However, the judgments seem not to be independent of the information provided by the experience of the task which is demonstrated by the link between the response times (on answer and confidence) and the relationship between the response time on the answers and the confidence level. The present work was able to show that the necessity to investigate gender in the context of confidence judgments and confidence biases. The processes of confidence judgments clearly differ between males and females which has widespread consequences for possible interventions to improve confidence judgments. Furthermore, the present research demonstrated that incentives in general are not necessarily an effective intervention to improve the accuracy of confidence judgments. They evidently increase the time and effort invested in the incentivized task, but incentives are only beneficial for the accuracy of confidence judgments when they explicitly depend on the accuracy and when the task is of medium difficulty.
8.1. Limits of the Present Research

One aim of this line of research was to investigate gender differences in confidence judgments in own skills and the confidence bias. I pretested the materials to test confidence judgments with gender-neutral knowledge questions. Nonetheless, I found gender differences in performance in two out of three studies. The finding of confidence biases in Study 2 (overconfidence in males and underconfidence in females) and 3 (overconfidence in males and females) can be interpreted independent of the difference in performance since the confidence bias reflects confidence relative to performance. Future research should focus even more on implementing gender-neutral materials. A clear conclusion on gender differences in confidence judgments and especially the underlying processes can only be drawn if the materials are not biased a priori.

Besides the focus on gender, I was interested in differentiating the effects of monetary incentives for performance and for an accurate judgment. It is noteworthy that the method of incentivizing in Study 2 and 3 are not completely comparable. In Study 2, I incentivized performance through paying small incentives for each correctly answered question whereas in Study 3 participants had the opportunity to win one big incentive. In addition, in Study 3, I compared incentive levels (low and high incentives) in contrast to a control condition with flat rate payment. Due to the implementation of the RBCEM (Healy, 2016) as measurement for confidence in Study 3, it was not possible to implement a comparable condition without incentive. Furthermore, in Study 2 we used an explicit method of incentivizing and in Study 3 an implicit one. Nonetheless, incentive compatibility in Study 3 was in my opinion of more importance than comparability to Study 2 which is why I chose the RBCEM (Healy, 2016) in Study 3. In contrast to Lebreton et al. (2018), I did only test the effects of incentives in a win frame in Study 2 and 3. The present studies should be repeated with a loss frame of the incentives.
9. The Effect of Intentions on Rational Decision Making: Evidence from Neural Correlates of Reinforcement Learning

Every day, we face situations in which we need to make a decision. This ranges from small, trivial decisions (for example what to wear) to very important decisions with long-term consequences (for example where to invest money). The traditional view in economics is that humans are rational decision makers (G. T. Huang, 2005). The perfect rational Bayesian optimizes the expected payoff of a decision by integrating new information with previous beliefs (Bayes & Price, 1763). There is however a great deal of empirical evidence that humans deviate from this normative model (e.g. see Charness et al., 2010; Erev et al., 2008; Fiedler et al., 2000; Tschan et al., 2009).

One particular decision strategy that conflicts with Bayesian Updating is the reinforcement heuristic (Charness & Levin, 2005). This heuristic is based on simple reinforcement learning (see Thorndike, 1911) which basically means that people tend to repeat successful behavior and avoid unsuccessful behavior. This conflict between the rational decision strategy and the simple reinforcement heuristic can be explained by dual process accounts (e.g. Alós-Ferrer & Strack, 2014; Evans, 2008; Kahneman, 2003; Strack & Deutsch, 2004). The reinforcement heuristic is an automatic process which is fast, unconscious and effortless whereas the rational Bayesian updating reflects a controlled process which is slower, effortful and demands consciousness.

Several studies have investigated the role of reinforcement in economic decision making (e.g., Achtziger & Alós-Ferrer, 2014; Achtziger, Alós-Ferrer, Hügelschäfer, & Steinhauser, 2015; Alós-Ferrer et al., 2016, 2017; Alós-Ferrer & Ritschel, 2018). Hügelschäfer and Achtziger (2017) for example investigated the effectiveness of goal intentions and implementation intentions in controlling the automaticity of the reinforcement heuristic.
They demonstrated that a specific goal intention and implementation intention that focused on the analysis of the decision feedback reduced reinforcement behavior and hence improved economic decision making in this paradigm. Reinforcement learning is not only observable in behavioral data but can also be detected in neural correlates. The feedback-related negativity (FRN), an event-related potential that reflects a reward prediction error (Holroyd & Coles, 2002), indicates reinforcement learning. A larger FRN is generated when the outcome of a decision deviates from expectations (especially after negative feedback). The aim of the present study was to replicate the findings of Hügelschäfer and Achtziger (2017) on the effectiveness of goal and implementation intentions in controlling the reinforcement heuristic. Additionally, and more importantly, the present work investigated the effect of goal and implementation intentions on the amplitude of the feedback-related negativity.

9.1. Rationality and the Reinforcement Heuristic

According to traditional economic theories, humans behave rationally which means they try to maximize their subjective expected utility (see G. T. Huang, 2005; Kahneman, 2003; Sanfey, 2007). A rational human being does so by integrating new information with previous beliefs by means of Bayes’ rule (Bayes & Price, 1763). People have prior beliefs regarding the likelihood of certain events. These beliefs are based on previous experiences and knowledge. Following Bayes’ rule, newly acquired information should then be taken into account, in order to update the probability of an event. The posterior probability of an event can then be mathematically derived from the prior probabilities and the new information. One critical assumption behind this model of a rational decision maker is that the decision maker has unlimited knowledge, time and cognitive resources.

However, there is a fundamental amount of empirical evidence that demonstrates that people are not rational Bayesian decision makers (for example base-rate neglect; Fiedler et al., 2000; Erev et al., 2008; conjunction fallacy; Charness et al., 2010; confirmation bias; Tschan et al., 2009). In these cases, the rational process of Bayesian updating is substituted by simpler decision rules, more specifically heuristics (see Tversky & Kahneman, 1974). Heuristics are simple, experience-based decision rules that work as cognitive shortcuts. In some cases, heuristics lead to good decisions, for example in complex decision situations or when only incomplete information is available (Gigerenzer, Todd, & the ABC Research Group, 1999; Hertwig, Hoffrage, & the ABC Research Group, 2013; Todd, Gigerenzer, & the ABC Research Group, 2012). In other cases, heuristics lead
to suboptimal decisions because they ignore the objective relevance of information due to their simplicity which contradicts normative prescriptions (Achtziger & Alós-Ferrer, 2014; Erev et al., 2008; Evans, 2006; Glöckner & Betsch, 2008). One typical heuristic that perturbs rational decision making is the reinforcement heuristic (Charness & Levin, 2005). As the name indicates, this heuristic is based on simple reinforcement learning (see Thorndike, 1911), which is the tendency to repeat successful behavior and avoid unsuccessful behavior. Based on this principle, decisions should improve over time by updating reward expectations according to the feedback of the outcome (Barto & Sutton, 1997). Yet, this restricted reliance on decision outcomes may lead to decision errors especially if the probability of an event should be taking into consideration (see Achtziger & Alós-Ferrer, 2014; Hügelschäfer & Achtziger, 2017).

The conflict between these two competing decision strategies (Bayesian Updating vs. Reinforcement Heuristic) is perfectly reflected by dual process theories (see, e.g. Alós-Ferrer & Strack, 2014; Evans, 2008; Kahneman, 2003; Strack & Deutsch, 2004). These theories postulate two types of processes. On the one hand, automatic processes which are immediate, fast, effortless and unconscious. On the other hand, controlled processes which are slow and demand cognitive resources and conscious intent. Achtziger and Alós-Ferrer (2014) demonstrated by measuring decision times in a probability updating task that Bayesian updating (compared to the reinforcement heuristic) is a rather controlled process. In contrast, the reinforcement heuristic represents a more automatic processes reflected by faster decisions. Evidence from neuroscience substantiates this argument by showing that neural correlates of reinforcement (FRN see Section 9.3) occurs extremely early in the decision process (Holroyd & Coles, 2002).

The relevance of reinforcement for economic decision making was investigated in several studies. Achtziger, Aós-Ferrer, Hügelschäfer, and Steinhauser (2015) investigated the neural correlates of reinforcement (FRN, see Section 9.3) in economic decisions and the influence of monetary incentives. Contrary to standard economic expectations, the results showed that high monetary incentives (compared to low incentives) reduced rational decision making and increased reinforcement-based decisions. The high incentive made the feedback valence (win or lose) of a first decision even more prominent because it was more relevant to the final pay-off. This led participants to rely even more on the reinforcement heuristic for a second decision. In the same paradigm, Alós-Ferrer, Hügelschäfer, and Li (2017) showed that framing influenced the reliance on the reinforcement heuristic. In a

\[\text{In the present study, there was no analysis of response times due to the fixed interval between the presentation of feedback and the second draw which was a necessity for EEG measurement (see Horstmann, Hausmann, & Ryf, 2010)}.\]
loss frame (instruction to avoid decisions that result in losing money), participants were more prone to shift away from an unsuccessful option than in a gain frame (instruction to make decisions that increase the amount of money). In a recent study, Alós-Ferrer and Ritschel (2018) moved beyond binary-choice tasks and investigated the reinforcement heuristic in a strategic economic setting by means of normal form games. The authors demonstrated that the reinforcement heuristic does not only apply to binary decisions, but also explains decisions in more complex situations (e.g. choices between three options). Hügelschäfer and Achtziger (2017) investigated the effectiveness of motivational interventions in fostering rational decision making (see the next section for more details). The results demonstrated that the automaticity of reinforcement can be controlled by setting specific goal intention and implementation intentions. Participants who were instructed to analyze the decision feedback carefully by means of goal intentions and implementation intentions displayed less reinforcement-based decisions than a control group. Based on this finding, the present study aimed to investigate the effectivity of motivational interventions on decision making more thoroughly by examining the neural correlates of reinforcement (FRN, see Section 9.3). The next section gives an overview over the evidence on the effectiveness of goal and implementation intentions.

9.2. Goal Intentions and Implementation Intentions: An Intervention to Control Automatic Processes

Setting goal intentions or forming implementation intentions (P. M. Gollwitzer, 1999) is a simple, quick and low-cost intervention to improve rational decision making. By setting goal intentions, people form a mental representation of a desired end state they want to achieve. For example, if a person wants to lose weight, a corresponding goal intention could be: ‘I want to lose two pounds!’ Although, there are some findings that strong goal intentions support goal striving (Ajzen, 1991; Armitage & Conner, 2001), research has shown that there is a significant gap between goal intentions and goal striving (see P. M. Gollwitzer & Sheeran, 2006; von Suchodoletz & Achtziger, 2011). A meta-analysis by Webb and Sheeran (2006) demonstrated that a medium-to-large (J. Cohen, 1992) change in commitment to the goal affected the corresponding behavior only in a small way ($d = .33$). This suggests that setting a goal and committing to it does not suffice to attain this goal. Another meta-analysis demonstrated that this failure in goal attainment is not due to lack of commitment but to the failure to act on the intentions (Sheeran, 2002).
A more promising approach of goal achievement is the formation of implementation intentions in addition to goal intentions (P. M. Gollwitzer, 1999, 2014). Implementation intentions specify when, where, and how an action is to be performed in order to attain a goal. This is done in the form of an if-then plan. For example, an if-then plan to lose weight could be formulated as: 'If I see unhealthy food, then I will not eat it!'. In a meta-analysis, Gollwitzer and Sheeran (2006) demonstrated that the effect of implementation intentions on goal attainment is much larger \((d = .65)\) than the effect of simple goal intentions. Implementation intentions foster a strong associative link between a specific situation and the goal-directed behavior which is specified in the then-part of the implementation intention (Achtziger, Gollwitzer, & Sheeran, 2008; P. M. Gollwitzer, 1999). This results in a heightened perceptual readiness to initiate the goal-directed behavior in the specified situation. More specifically, by forming an implementation intention the information processing shifts from top-down to bottom-up which automates goal striving. This allows for the specified behavior to be initiated immediately, quickly, effortlessly and without much conscious intent (Bayer, Achtziger, Gollwitzer, & Moskowitz, 2009; Hügelschäfer, Jaudas, & Achtziger, 2016).

Implementation intentions have been shown to be especially helpful in goal achievement when the goal is difficult to attain, or the goal-directed behavior is unpleasant (P. M. Gollwitzer & Sheeran, 2006). A study by Bayer and Gollwitzer (2007) demonstrated that forming a self-efficacy strengthening implementation intention in addition to a goal intention improved female students’ performance on a math test (Study 1) and male student’s performance on an analytical reasoning test. A recent field study by Achtziger, Glas, Kenning, and Rudolph (2019) showed that an implementation intention intervention reduced unhealthy snack consumption in employees of a retailer. More interestingly, participants who formed an implementation intention performed as well as participants who received a financial incentive in reducing unhealthy snack consumption and the effect of the implementation intention lasted longer. Research on the control of negative inner states also demonstrated a positive effect of forming implementation intentions. Tennis players reported increased concentration after forming implementation intentions to control their nervousness (Achtziger et al., 2008). Forming an implementation intention to inhibit temptation also helped students to reduce the influence of test anxiety on their performance in a math exam (Parks-Stamm, Gollwitzer, & Oettingen, 2010).

It is important to note that implementation intentions do not exclusively reduce existing automatic processes but can also intensify them. A study by Gollwitzer, Schwörer, Stern, Gollwitzer and Bargha (2017) demonstrated that depending on the wording of the if-then
plan the implementation intention either reduced social projection ('If I’m asked to estimate what percent of other people agree with me, then I will remember that other people are different!') or increased it ('If I’m asked to estimate what percent of other people agree with me, then I will remember that other people are similar!').

Another example of the complex effect of goal and implementation intentions on decision processes is work by Hügelschäfer and Achtziger (2017). In three studies, the authors investigated whether goal intentions and implementation intentions improve economic decision making by reducing less optimal automatic responses. Participants worked on a probability-updating task in which the optimal decision rule (Bayes’ rule) conflicted with a simple decision rule (reinforcement heuristic). The participants’ goal was to draw as many balls of a certain color (blue or green) from two urns containing six balls each. The color of the balls was concealed to the participants, but they received information on two possible distributions of the colored balls. One trial consisted of two draws which allowed the participants to evaluate the information of the first draw in order to make the choice for the second draw. The results of Studies 1 and 2 showed that forming a goal intention or an implementation intention to analyze the decision feedback improved decision making (reduced decision errors caused by relying on the reinforcement heuristic). Contrary, a goal intention to suppress disappointment over negative feedback had no effect on reinforcement behavior and a corresponding implementation intention even increased reinforcement errors. In Study 3, participants set a goal intention to keep cool in response to negative feedback or formed the corresponding implementation intention. The goal intention as well as the implementation intention to keep cool reduced unwanted reinforcement behavior. Again, this work underlines that goal and implementation intentions can reduce or increase automatic responses depending on the specified goal and if-then plan.

The effect of implementation intentions can not only be demonstrated in behavioral data but can also be detected in event-related-potentials of electrocortical activity (see the review by Wieber, Thürmer, & Gollwitzer, 2015). A study by Paul et al. (2007) investigated the effect of implementation intentions on the performance in a Go/no-go task in children with ADHD. The children in the implementation intention condition displayed better response inhibition and an increased P300 (which is associated with conflict monitoring) compared to the children in the control condition. Schweiger Gallo (2009) showed that participants who formed an implementation intention to ignore fear-inducing stimuli reported less negative affect in response to pictures of spiders compared to a control condition. Moreover, they demonstrated a lower positivity in
the P1 in response to pictures of spiders. Negative stimuli typically elicit larger P1 amplitudes. A study by Hügelschäfer, Jaudas, and Achtziger (2016) demonstrated that forming an implementation intention inhibited automatic gender categorization. Participants were instructed to categorize pictures according to gender. The N170 was compared across three conditions (control vs. goal intention vs. implementation intention). This early event-related potential is associated with face processing. In the implementation intention condition (compared to the control condition and the goal intention condition), participants displayed an equal N170 amplitude for gender-congruent and gender-incongruent pictures which reflects less automated processing of social information.

In the present study, the aim was to investigate the effect of goal intentions and implementation intentions (see Hügelschäfer & Achtziger, 2017) on the feedback-related negativity (FRN) which is the neural correlates of reinforcement. The next section gives an overview of this event-related potential.

9.3. The Feedback-Related Negativity: An Indicator of Reinforcement Learning

In the field of neuroeconomics, neuroscientific methods are used to test economic theories of human behavior (see Camerer, 2007; Camerer, Loewenstein, & Prelec, 2005; Loewenstein, Rick, & Cohen, 2008). The predominant brain imaging technique in the field of neuroeconomics is the electroencephalography (EEG) which is a direct measure of neural activity with very high temporal resolution. This high resolution allows for investigations of decision-making processes in the human brain via event-related potentials (ERPs) which are time-locked to certain events (or stimuli) and occur in preparation of a response (Fabiani, Gratton, & Coles, 2000; Luck, 2014).

A highly investigated ERP is the feedback-related negativity (FRN) which is a component that represents a reward prediction error (Holroyd & Coles, 2002). The FRN occurs 200-350 ms after feedback and is strongest at the frontocentral midline. Several studies have identified the anterior cingulate cortex (ACC) as the source of the FRN (e.g. Bellebaum & Daum, 2008; Gehring & Willoughby, 2002; Hauser et al., 2014) which is consistent with research findings confirming the role of the medial frontal cortex in performance monitoring and regulation of cognitive control (see Gläscher et al., 2012; Kerns et al., 2004). The amplitude of the FRN is typically larger following negative (unfavorable) feedback (e.g., errors or loss of money) than positive (favorable) feedback.
Many researches have argued that the feedback-related negativity constitutes a neural indicator of reinforcement learning (Chase et al., 2011; Holroyd & Coles, 2002; Holroyd et al., 2002, 2005). The FRN is linked to the basal ganglia and the dopamine system which are involved in reinforcement learning (Schultz, 2010; Schultz, Dayan, & Montague, 1997). The outcome of a decision is evaluated against previously held expectations by the basal ganglia. If the outcome deviates from the expectations (better or worse), then this information is transmitted to the anterior cingulate cortex via an increase or decrease in dopamine in the midbrain (see Daw & Doya, 2006; Montague, Hyman, & Cohen, 2004; Schultz, 2002, 2004). According to the Reinforcement Learning Theory (Holroyd & Coles, 2002) the strength of this dopamine signal determines the amplitude of the FRN (see also a recent study by Hauser et al. (2014) on the influence of surprise on the FRN modulation). Hence, the FRN reflects the prediction error of reward.

There are several studies supporting the claim that the FRN indicates reinforcement learning. Work by Cohen and Ranganath (2007) showed that larger FRN amplitudes after losing to a computer opponent in a strategic economic game preceded behavioral switches meaning switching to the other of two stimuli after losing. This supports the assumption that reinforcement learning is used to evaluate decision options. Furthermore, it demonstrates that the FRN reflects automatic reinforcement learning because it occurs before a conscious decision is made. Yasuda, Sato, Miyawaki, Kumano, and Kubok (2004) found that the FRN amplitude was highest when participants’ expectations were violated (expecting a positive outcome and receiving a negative one). Moreover, they found a positive correlation between the FRN amplitude and the rate of response switching which indicates reinforcement learning.

There are different factors modulating the FRN amplitude. Several studies have demonstrated that the FRN amplitude decreases with learning (Holroyd & Coles, 2002; Nieuwenhuis et al., 2002; Pietschmann, Simon, Endrass, & Kathmann, 2008). This is in line with the assumption that the FRN reflects reinforcement learning. The longer people learn, the more evidence they accumulate for their decision strategy. In turn, the decision outcomes are less unexpected which results in a decrease of the FRN amplitude. Work by Bellebaum, Kobza, Thiele, and Daum (2010) showed that the processing of feedback information depends crucially on the relevance of one’s own action. Part of the participants actively played a learning task where monetary incentives depended on the performance on the task. The other half of the participants observed. Both groups
demonstrated a similar learning rate but the FRN amplitude after negative feedback was significantly smaller in observers. Another factor modulating the amplitude of the FRN are the magnitude of monetary incentives. Achtziger, Alós-Ferrer, Hügelschäfer, and Steinhauser (2015) showed that the FRN amplitude after negative feedback was significantly higher in case a high monetary incentive was at stake (compared to a low incentive) for participants who committed strongly to the reinforcement heuristic (high reinforcement error rate). The high incentive lead participants to rely even more on reinforcement learning which was then reflected in a higher FRN amplitude. Work by Benett, Sasmita, Maloney, Murawski, and Bodel (2019) supports this finding. They investigated belief updating and FRN amplitudes in a simple reward learning task. The results showed larger belief updating and larger FRN amplitudes in the monetary feedback condition compared to an instructive feedback condition. This supports the notion that monetary incentives influence inference strategies in reward learning which is also reflected in the electrophysical correlates of belief updating. The present work aims at investigating whether setting goal intentions or forming implementation intentions also modulate (specifically decrease) the FRN amplitude by reducing the reliance on reinforcement learning.

9.4. Research Questions and Hypotheses

The aim of the present study was to investigate the influence of goal intentions and implementation intentions on the feedback-related negativity (FRN), a neural correlate of reinforcement (Holroyd & Coles, 2002). Hügelschäfer and Achtziger (2017) demonstrated that specific goal intentions and implementation intentions can control the automaticity of the reinforcement heuristic in economic decision making. The authors showed that the goal intention and implementation intention that focused on analyzing the decision feedback carefully reduced the reliance on the reinforcement heuristic and hence improved economic decision making. Based on these findings, the present study investigated the effect of the same goal intention and implementation intention on the feedback-related negativity in the same paradigm as Hügelschäfer and Achtziger (2017).

One of the research aims was to replicate the findings on the behavioral data of Hügelschäfer and Achtziger (2017). In case of alignment of the reinforcement heuristic and Bayes’ rule, participants should commit less decision errors than in case of conflict (Hypothesis 1). Further, I expected to replicate the effect of the intentions manipulation on decision making. Participants in the goal intention and implementation intention
conditions were expected to commit less reinforcement errors than participants in the control condition (Hypothesis 2). In line with the findings of Hügelschäfer and Achtziger (2017), no difference between the two intention conditions in the reinforcement error rate was expected. As mentioned earlier, the amplitude of the FRN reflects reinforcement learning (Holroyd & Coles, 2002). Hence, I expected the FRN amplitude to match the pattern of the behavioral data. More specifically, since reinforcement is expected to be controlled by the goal intention and the implementation intention, the FRN amplitude of the two intention conditions should be smaller than the FRN amplitude of the control condition (Hypothesis 3). No difference in the FRN amplitude was expected between the goal intention and implementation intention condition.

9.5. Methods

9.5.1. Participants

Eighty-six right-handed participants (39 females, 47 males) were recruited at the University of Cologne and the Zeppelin University Friedrichshafen. Students majoring in economics were excluded from participation. The participants received a show-up fee of 12 € plus a monetary bonus that depended on the outcome of the decision task (total amount: $M = 23.75\, €, \, SD = .81\, €$). Eleven participants were excluded from the EEG data analysis due to an excessive number of artifacts which resulted in too few valid trials for averaging the ERP waveform.

9.5.2. Design

The study followed a 3 between (intentions: control condition vs. goal intention condition vs. implementation intention condition) × 2 between (counterbalance: blue as rewarded color vs. green as rewarded color) × 3 within (first draw: forced left vs. forced right vs. free) mixed-model design. The dependent variables were error rates in the decision task (see below for more detailed description) and the feedback-related negativity measured during the decision task.

9.5.3. Materials

The experiment was run on a personal computer using Presentation® Software 12.2 (Neurobehavioral Systems, Albany, CA). The stimuli for the decision task were images of
two urns containing six balls each (image size 154 x 93 pixels) as well as images of blue and green balls (image size 40 x 40 pixels). The colors of the balls were selected based on the equality of their brightness in order to control the effect of color brightness on the EEG data. The distance between the two urns were kept close together to reduce eye movements during the EEG recording. The participants were seated at a distance of approximately 50 cm to the monitor.

9.5.4. Decision Task

The decision task was identical to the task developed in Achtziger, Alós-Ferrer, Hügelschäfer and Steinhauser (2015) and was based on the two-draw paradigm introduced by Charness and Levin (2005). Participants were presented with two urns containing six balls each. The color of the balls (either blue or green) was concealed to the participants. Depending on counterbalancing, the participant received 18 Cents for each blue (or green) ball. Participants chose an urn by pressing one of two keys of the keyboard (‘F’ for the left urn and ‘J’ for the right urn) and then the computer program drew a ball randomly from the selected urn. Each trial consisted of two draws with replacement. After the first draw the color of the ball is revealed and then the participant chose the urn for the second draw. There were 60 trials in total. For the first 30 trials the first draw was forced by the computer program either left or right and the second draw was free. This was visualized for the participant by greying out the urn that could not be chosen. For the last 30 trials both draws could be chosen freely by the participants.

The possible distributions of the balls in the urns was known to the participants but it varied depending on the unknown state of the world (up or down). That is, participants were informed of the distribution of the balls in both states of the world and they knew that the prior probability of both states was 50%. They were not informed about the state of the world during the actual trial. Further, they received the information that the state of the world was selected at random at the beginning of each new trial and was held constant during the trial (across both draws). In case the participant earned a monetary incentive for drawing blue balls, the up state of the world consisted of six blue balls in the right urn and two green and four blue balls in the left urn. The down state of the world consisted of six green balls in the right urn and two blue and four green balls in the left urn (see Figure 9.1). In case participants were rewarded for green balls, the distribution of the balls was vice versa.
9.5.5. Decision Rules and Decision Errors

The design of the decision task makes it possible to distinguish two types of decision rules: Rational decisions and decisions based on the reinforcement heuristic. A rational decision maker (in Bayesian terms) would use the information from the first draw (color of the ball) to update her prior probability of the state of the world. The second draw would then be based on the posterior probability in order to maximize the expected payoff.

A decision maker who relies fully on the reinforcement heuristic would follow the simple decision rules of ‘win-stay/lose-shift’. If the first draw would result in a rewarded ball, the decision maker would draw again from the same urn (‘win-stay’). If the first draw would reveal a losing/non-rewarded ball, the decision maker would switch to the other urn for the second draw (‘lose-shift’). Hence the decision maker who follows the reinforcement heuristic bases the decision of the second draw solely on the feedback of the first draw and ignores the prior probability.

After a first draw from the right urn, the decision rules of Bayes’ rule and the reinforcement heuristic are aligned. The feedback of the color reveals the state of the world (since the right urn only contains one color). If a rewarding ball is drawn from the right urn, the posterior probability of the up state of the world is 1 and the same urn should be selected for the second draw. The reinforcement heuristic prescribes the same behavior. If the first draw from the right urn results in a positive feedback (rewarded color), the second ball should be drawn from the same urn. The opposite is true if the first draw from the right reveals a non-rewarded ball. Then the posterior probability
of the up state of the world is 0, and the payoff is maximized by switching to the left urn for the second draw. According to the reinforcement heuristic, the decision maker should also switch to the left urn for the second draw after the negative feedback from the first draw from the right urn. A (rare) mistake after a first draw from the right urn was categorized as an understanding error since it signals that the participant did not fully understand the task or did not pay attention during the first draw. Since a first draw from the right urn reveals the state of the world and hence improves the chances of a rewarded ball for the second draw, a first draw from left is considered a first-draw-error. Note that this only applies to the trials with a free first draw (second half).

After a first draw from the left urn, the decision rules of Bayes’ rule and the reinforcement heuristic are in conflict. In contrast to the first draw from the right urn, a first draw from the left urn does not reveal the state of the world. Nonetheless, the information of the first draw (color of the ball) can be used to update the prior probability of the state of the world. If a rewarded ball is drawn from the left urn, the posterior probability of the up state of the world is \( \frac{2}{3} \). Hence, Bayes’ rule would prescribe to switch to the right urn (which only contains rewarded balls in the up state of the world) for the second draw to maximize the expected payoff. If a non-rewarded ball is drawn from the left urn, the posterior probability of the up state of the world is \( \frac{1}{3} \). Following Bayes’ rule, the decision maker should stay with the left urn to maximize the probability of drawing a rewarded ball. To summarize, Bayes’ rule would prescribe a win-shift/lose-stay strategy after a first draw from the left.

As mentioned above, this conflicts with the reinforcement heuristic. Since the reinforcement heuristic is solely based on the feedback of the first draw, the decision rule after the first draw from the left is the same as after the first draw from the right. If a rewarded ball is drawn from the left urn, the same urn is selected for the second draw (’win-stay’). If a non-rewarded ball is drawn from the left urn, the right urn is selected for the second draw (’lose-shift’). Since following the reinforcement heuristic does not maximize the expected payoff in this case, mistakes after the first draw from the left are categorized as reinforcement errors.

9.5.6. Experimental Manipulations

In the control condition, participants received no further instruction and started with the decision task immediately after the practice trials. The instructions for the participants in the goal intention and the implementations intentions conditions were printed on paper and handed to them after the practice trials. Participants in the goal intention were
instructed to set the following goal for the decision task: 'I want to analyze the drawn balls carefully!'. They were told this goal would be helpful in the upcoming task. The participants were instructed to repeat this goal intention silently and were asked to write it down three times. In the implementation intentions condition, participants received the same instructions. In addition, they were told that in order to achieve this goal they should form the following implementation intention: 'If a drawn ball is presented, then I will analyze it carefully!'. Again, they were instructed to repeat this implementation intention silently and to write it down three times.

9.5.7. Procedure

The experiment took place in individual sessions and were all conducted by a female experimenter and with the help of a student assistant. The individual sessions took approximately 90 minutes (10-15 minutes for the decision task). Upon arrival, participants were greeted and asked to sit down in front of the PC. First, they gave written informed consent for the experiment and the EEG measurement. The experimenter tested the handedness with a short questionnaire to ensure that the participant was right-handed.

After the electrodes for the EEG acquisition were applied, participants read the instructions of the decision task. The instructions explained the decision task in detail and were illustrated with screenshots to foster the participants’ understanding. In addition, it was explained that it was important for the EEG acquisition that participants move as little as possible, keep their fingers on the keys of the keyboard and fixate the fixation cross. After participants finished reading the instructions, they were first asked if they had any questions concerning the task. Then they were asked to explain the decision task and the rules in their own words. The experimenter assessed the participants’ understanding by means of an experimental protocol (see Appendix A.3). The experimenter answered questions concerning the understanding of the decision task and the rules of the task but did not answer questions concerning possible decision strategies.

Then, participants completed three practice trials under the supervision of the experimenter. These trials gave an example for the first 30 trials (forced first draw) and the second 30 trials (free draws) in order to familiarize participants with the task. The experimenter made sure that there were no more open questions and the participant understood the task. Then, participants in the goal intention and implementation intention condition received written instructions to form either goal or implementation intentions (see experimental manipulations). The participants in the control condition started right
away after the practice trials.

Then the EEG recording and the experiment was started. Before the decision task, a baseline EEG was measured for one minute. Participants were instructed to look straight at a fixation point on the screen during this time span. They were told to relax and to think about nothing in particular. Again, they were informed that it was important that they sat still for the EEG acquisition. After the baseline measurement, the decision task began.

The experiment consisted of 60 trials. At the beginning of each trial, an overview of the two state of the world (possible distributions of balls) was presented until the participant pressed the space bar (Figure 9.2). Then a fixation cross was presented for an inter-stimulus-interval of 500 ms. Following, the two urns were presented until the participant indicated her choice by pressing a key on the keyboard (either 'F' = left or 'J' = right). After an inter-stimulus-interval of 1000 ms, the feedback of the first draw was presented to the participant for another 1000 ms. Then the participant was instructed to press the space bar to proceed to the second draw. Again, the participants indicated her second choice by pressing either 'F' or 'J' on the keyboard while the feedback of the first draw (colored ball) remained visible. After an inter-stimulus-interval of 1000 ms, the feedback of the second draw was presented for 1000 ms. Then, an instruction to proceed to the next trial by pressing the space bar was presented at the bottom of the screen. Figure x gives an overview of the sequence of one complete trial.

As mentioned above, for the first 30 trials the first draw was forced by the computer program. For trials with an odd number participant were forced to draw first from the left urn and for trials with an even number they were forced to draw first from the right urn. This constraint was a necessity to ensure sufficient observations of a first draw from the left urn (see Achtziger & Alós-Ferrer, 2014; Achtziger, Alós-Ferrer, Hügelschäfer, & Steinhauser, 2015). These are the situations in which the decision rules (Bayes’ rule and the reinforcement heuristic) are in conflict.

After the 60 trials, the amount of money the participants earned (in addition to the show-up fee) was displayed on the screen. Then, the cap and the external electrodes were removed. Next, the participants filled out a questionnaire (paper-pencil; see Appendix A.3) on their understanding of the task, the difficulty of the task, the effort invested in the task, the importance of their performance on the task and questions concerning the intention manipulation (importance, commitment, helpfulness). In addition, the questionnaire measured the participants’ subjective value of money (see Brandstätter & Brandstätter, 1996), the self-reported skills in statistics and probabilities, the high school
grades, and demographic information (gender and age). All self-reported measures were rated on a 10 cm continuous analog scale. Lastly, participants indicated their valence and arousal of the two balls on a discrete scale from 1 to 5 (Self-Assessment Manikin [SAM]; Hodes, Cook III, & Lang, 1985; Bradley & Lang, 1994). Following the questionnaire, participants had the opportunity to wash their hair. At the end they were thanked and paid.

![Figure 9.2.: Overview of the sequence of one trial](image)

### 9.5.8. EEG Acquisition and Analysis

The EEG data was recorded using BioSemi Active II system (BioSemi, Amsterdam, The Netherlands, www.biosemi.com) and the analysis was performed using Brain Electrical Source Analysis (BESA) software (BESA GmbH, Gräfelfing, Germany, www.besa.de).

For the EEG acquisition 64 Ag-AgCl pin-type active electrodes were used which were fixed on an elastic cap (ECI). The electrodes were arranged according to the 10-20 system which is a relative positioning system to ensure comparable results despite variations in skull size. Additionally, two electrodes were placed at the right and left mastoid. Further, two electrodes were placed at the left eye to monitor eye movements and blinks by electro-oculogram (EOG) signals. One electrode was placed approximately 1 cm to the left side of the left eye to record vertical eye movements (blinks) and another electrode was placed approximately 1 cm below the left eye to record horizontal eye movements (from left to right). The Common Mode Sense active (CMS) electrode and the Driven Right Leg passive (DRL) electrode formed the ground electrode during acquisition. It was ensured that the electrode impedance was below 20kΩ. The EEG and the EOG were sampled at a 512 Hz rate. The data was re-referenced off-line to a mastoid reference. Ocular artifacts were corrected by an eye-movement correction algorithm implemented in BESA software.
The data was feedback-locked to the feedback of the first draw (positive or negative). The data was segmented into epochs from 500 ms before to 1000 ms after stimulus onset. The time span before the feedback (500 ms) was used for baseline correction. The epochs were averaged separately following positive and negative feedback resulting in two average ERP waveforms per participant. Epochs including EEG or EOG voltage exceeding +/-100 μV were omitted from averaging to avoid excessive electromyogram (EMG) or other noise transients. The resulting waveforms were filtered with a high-pass filter of 0.1 Hz and a low-pass filter of 30 Hz. As mentioned above, eleven participants were excluded from the ERP analysis due to too few valid epochs. On average 4.7% of the epochs were excluded from the average waveform from the non-excluded participant. The averaging of the waveforms across all participants resulted in the Grand Average (separately following positive and negative feedback).

In accordance with previous research (Holroyd, Krigolson, Baker, Lee, & Gibson, 2009) and in accordance with theoretical considerations (Luck, 2014), the feedback-related negativity was analysed using the difference wave. The ERPs following positive feedback were subtracted from the corresponding ERPs following negative feedback result in one difference wave for each participant. The FRN was quantified by the most negative peak occurring in the time window 200-350 ms after feedback onset (see Sambrook & Goslin, 2015). The FRN was evaluated at channel FCz were it is normally maximal (Holroyd et al., 2009; Holroyd & Krigolson, 2007).

9.6. Results

9.6.1. Equivalence of Groups

Before examining differences in decision errors, I tested for differences between the experimental conditions (control condition vs. goal intention condition vs. implementation intention condition) regarding several control variables. The scores of all self-reported variables ranged continuously between 0 and 10. ANOVAs with the between-subjects factor intentions (control vs. goal intention vs. implementation intention) revealed no difference between the experimental conditions in the self-reported understanding of the task, the difficulty of the task, the effort invested in the task, the importance to perform well, the subjective value of money, $F_{s} \leq 1.44$, $p_{s} \geq .242$. Overall, the reported understanding of the task was quite high $M = 8.64$ ($SD = 1.22$) and the perceived difficulty quite low $M = 1.94$ ($SD = 1.78$). Participants reported high effort $M = 8.79$
(SD = 1.37) and a reasonable high importance to perform well on the task \( M = 7.92 \) (SD = 1.73). The experimental groups did also not differ in their high school grades, the self-reported statistical skills and the skills in calculating probabilities, \( F_s \leq .56, ps \geq .574 \). On average, participants reported moderate statistical skills, \( M = 5.08 \) (SD = 2.25) and moderate skills in calculating probabilities, \( M = 5.50 \) (SD = 2.05).

I further compared the counterbalancing conditions. Approximately half of the participants (N = 44) were rewarded for drawing blue balls and the other half for drawing green balls (N = 42). As expected, the counterbalancing conditions differed in the valence of the balls. Participants who were paid for blue balls rated these more positive than participants who were paid for green balls and vice versa (blue balls: \( t(84) = 17.21, p < .001 \); green balls: \( t(83) = 18.01, p < .001 \)). The arousal for blue balls was significantly higher in participants who were paid for green balls than participants who were paid for blue balls, \( t(82) = 2.28, p = .025 \). There was no significant difference in the arousal for green balls, \( t(83) = 1.16, p = .248 \). For the descriptive statistics see Table 9.1.

The comparison of the counterbalancing conditions did not yield any differences regarding the understanding of the task, the effort invested in the task, the statistical skills or the skills in calculating probabilities, \( ts \leq 1.59, ps \geq .115 \). Participants who were rewarded for blue balls reported a higher subjective value of money \( M = 7.33 \) (SD = 1.42) than participants who were rewarded for green balls \( M = 6.54 \) (SD = 1.61), \( t(84) = 2.43, p = .017 \).

Next, I compared the experimental groups regarding their goal commitment, the importance of the goal and the helpfulness of the goal. The ANOVAs with the between-subjects factor intentions (control vs. goal intention vs. implementation intention) revealed no difference for goal commitment between the control condition, the goal intention condition and the implementation intention condition, \( F(2, 83) = 2.27, p = .109 \), and also no difference for the importance of the goal, \( F(2, 83) = 1.21, p = .303 \). The analysis showed a difference in the helpfulness of the goal, \( F(2, 83) = 3.43, p = .037 \).
\eta^2_p = .08. Surprisingly, the participants in the control condition rated the helpfulness of the goal on average \( M = 6.03 \) (SD = 2.56), in the goal intention condition \( M = 5.90 \) (SD = 2.65) and in the implementation intention condition \( M = 4.40 \) (SD = 2.65). Subsequent t-tests showed that the helpfulness of the goal differed between the control condition and the implementation condition, \( t(55) = 2.34, p = .022 \), and between the goal intention condition and the implementation intention condition, \( t(53) = 2.18, p = .034 \). There was no difference between the control condition and the goal intention condition, \( t(58) = .19, p = .847 \). It is somewhat curious that participants in the control condition rated the helpfulness of the goal intention as high as participants in the goal intention condition. A possible explanation could be that participants in the control condition misinterpreted the question and rated the potential helpfulness of such a goal intention (since they did not form such an intention in advance of the task).

Then, I compared the commitment, importance and helpfulness of the goal intention with those of the implementation intention within the implementation intention condition. A t-test showed a marginal significant difference in commitment between the goal and the implementation intention, \( t(25) = 1.75, p = .092 \), and a significant difference in the importance of the goal and the implementation intention, \( t(25) = 2.70, p = .012 \). Participants in the implementation intention condition committed less to the implementation intention, \( M = 5.97 \) (SD = 2.69) than to the goal intention \( M = 6.54 \) (SD = 2.41). They also rated the implementation intention as less important \( M = 6.03 \) (SD = 2.37) then the goal intention \( M = 7.07 \) (SD = 2.37). There was no difference in the helpfulness of the goal intention and the implementation intention, \( t(25) = .53, p = .602 \).

### 9.6.2. Error Rates

As mentioned above, we distinguished two types of errors: reinforcement errors and understanding errors. Reinforcement errors which occur after a first draw from the left, can be differentiated into 'win-stay' and 'lose-shift' errors. Understanding error (occurring after a first draw from the right) consist of 'win-shift' and 'lose-stay' errors. Table 9.2 gives an overview of the error rates of the different types of errors. Figure 9.3 displays the error frequency over time. As can be seen, the frequency of reinforcement errors (win-stay, lose-shift) and of understanding errors (win-shift, lose-stay) does not vary much over time.
Table 9.2.: Descriptive statistics of error rates by error type

<table>
<thead>
<tr>
<th>Intention condition</th>
<th>Reinforcement errors</th>
<th>Understanding errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Win-Stay Error</td>
<td>Lose-Shift Error</td>
</tr>
<tr>
<td>Control</td>
<td>45.5%</td>
<td>38.8%</td>
</tr>
<tr>
<td>Goal Intention</td>
<td>48.0%</td>
<td>47.3%</td>
</tr>
<tr>
<td>Implementation Intention</td>
<td>44.7%</td>
<td>56.7%</td>
</tr>
<tr>
<td>Total</td>
<td>46.1%</td>
<td>47.1%</td>
</tr>
</tbody>
</table>

Figure 9.3.: Error frequency over time

When participants drew first from the right urn, 250 of 3114 (8.1%) decision were incorrect resulting in an understanding error. When beginning with the left urn, 965 of 2047 (47.1%) decisions were incorrect which led to a reinforcement error. As expected, the participants committed much more reinforcement errors than understanding errors. Next, I tested whether the intentions manipulations affected the error rates in the predicted direction. First, the error rates of understanding errors were compared across the three intentions conditions (control vs. goal intention vs. implementation intention) by means of a Kruskal-Wallis test. The test revealed no difference between the intention conditions for the error rates of understanding errors, $H(2) = .77$, $p = .681$ (see Figure 9.4). Participants in the control condition committed on average $M = 5.9\%$ ($SD = 10.7\%$)
understanding errors. In the goal intention condition, the average understanding error rate was $M = 14.2\%$ ($SD = 23.1\%$) and in the implementation intention condition $M = 7.5\%$ ($SD = 11.6\%$). Separate tests showed also no significant difference for win-shift errors, $H(2) = 1.79$, $p = .408$, or lose-stay errors, $H(2) = .35$, $p = .838$. This did not come as a surprise since the error rates for understanding errors were already overall low and there were no expectations of the intentions manipulation to affect understanding errors.

Then, I tested the effect of the intentions manipulation on the error rates of reinforcement errors. A Kruskal-Wallis test revealed no significant difference between the intention conditions for reinforcement error, $H(2) = 1.15$, $p = .562$ (see figure 9.4). Participants in the control condition committed on average $M = 42.5\%$ ($SD = 30.7\%$) reinforcement errors. In the goal intention condition, the average reinforcement error rate was $M = 46.4\%$ ($SD = 27.9\%$) and in the implementation intention condition $M = 50.0\%$ ($SD = 28.8\%$). There was also no significant difference between conditions when the errors were differentiated in win-stay, $H(2) = .11$, $p = .948$, and lose-shift errors, $H(2) = 3.92$, $p = .141$.

Figure 9.4.: Understanding and reinforcement errors by condition

![Image of bar chart showing understanding and reinforcement errors by condition]
Each participant made 60 second-draw decisions which resulted in a strongly balanced panel. This panel data was analyzed by means of random-effects probit regressions on second-draw decisions (error vs. correct decision), assuming non-independence of observations within participants. The self-reported questionnaire data (goal commitment and statistical skills) were z-standardized. The variable time was calculated by dividing the trial number by 60. Table 9.3 show the results of the regression analysis. In line with the results of the Kruskal-Wallis test, Model 1 demonstrates that the goal and implementation intention condition did not significantly differ from the control condition in the amount of errors. As expected, more errors were committed in situations under conflict (first draw from the left). Not surprisingly, participants committed fewer errors with time. This effect is also due to the fact that participants could choose the first urn freely after the 31st trial. Gender and goal commitment (Model 2) did not predict errors and neither did the FRN amplitude or the interaction between conflict and the FRN amplitude (Model 3).
Table 9.3.: Random-effects panel probit regression on second-draw error (0 = correct choice, 1 = error)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal Intention Dummy</td>
<td>.19</td>
<td>.28</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>(.23)</td>
<td>(.24)</td>
<td>(.29)</td>
</tr>
<tr>
<td>Implementation Intention Dummy</td>
<td>.17</td>
<td>.24</td>
<td>.39</td>
</tr>
<tr>
<td></td>
<td>(.24)</td>
<td>(.24)</td>
<td>(.28)</td>
</tr>
<tr>
<td>Counterbalance (blue = 1)</td>
<td>-.23</td>
<td>-.24</td>
<td>-.14</td>
</tr>
<tr>
<td></td>
<td>(.19)</td>
<td>(.19)</td>
<td>(.21)</td>
</tr>
<tr>
<td>Conflict (1 = yes)</td>
<td>1.53***</td>
<td>1.53***</td>
<td>1.44***</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
</tr>
<tr>
<td>First Success (1 = yes)</td>
<td>.01</td>
<td>.01</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.05)</td>
</tr>
<tr>
<td>Time</td>
<td>-.57***</td>
<td>-.57***</td>
<td>-.52***</td>
</tr>
<tr>
<td></td>
<td>(.09)</td>
<td>(.09)</td>
<td>(.09)</td>
</tr>
<tr>
<td>Gender (male = 1)</td>
<td>-.25</td>
<td>-.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.20)</td>
<td>.(23)</td>
<td></td>
</tr>
<tr>
<td>Goal Commitment</td>
<td>-.10</td>
<td>-.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.10)</td>
<td>(.11)</td>
<td></td>
</tr>
<tr>
<td>FRN</td>
<td></td>
<td></td>
<td>.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.07)</td>
</tr>
<tr>
<td>FRN × Conflict</td>
<td></td>
<td></td>
<td>-.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.03)</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.49***</td>
<td>-1.41***</td>
<td>-1.48***</td>
</tr>
<tr>
<td></td>
<td>(.20)</td>
<td>(.22)</td>
<td>(.28)</td>
</tr>
<tr>
<td>log likelihood</td>
<td>-1874.05</td>
<td>-1872.93</td>
<td>-1564.51</td>
</tr>
<tr>
<td>Wald chi2</td>
<td>921.68***</td>
<td>922.96***</td>
<td>745.29***</td>
</tr>
<tr>
<td>Number of observations</td>
<td>5160</td>
<td>5160</td>
<td>4440</td>
</tr>
</tbody>
</table>

Notes. Standard errors in brackets. *p < .10, **p < .05, ***p < .01.
9.6.3. ERP Analysis – FRN

The grand-average difference wave from channel FCz over all participants is shown in Figure 9.5. It displays a substantial FRN. Figure 9.7 shows the corresponding topography which illustrates a negativity at the frontocentral midline. Figure 9.6 shows the grand-average difference wave separately for each condition. The FRN amplitude was compared across conditions by means of an analysis of variance (ANOVA) with the between-subjects factor intention (control vs. goal intention vs. implementation intention). The ANOVA revealed a significant main effect of intention, F(2, 71) = 4.12, p = .020, η² = .10. Participants in the control condition displayed an average FRN amplitude of M = −.98 (SD = 1.75). The average FRN in the goal intention condition was M = −1.94 (SD = 1.55) and in the implementation intention condition M = −2.29 (SD = 1.69). Subsequent t-tests showed a significant difference in FRN amplitudes between the control condition and the goal intention condition, t(47) = 2.02, p = .049, and the control and the implementation intention condition, t(48) = 2.69, p = .010. The FRN amplitude of the goal and implementation intention condition did not differ significantly, t(47) = .76, p = .451.

Next, I examined whether the FRN amplitude differed depending on the reinforcement error rate. By means of a median split (separately per condition) of the reinforcement error rates, participants were classified as having high and low reinforcement error rates. An ANOVA with the between-subjects factors intention (control vs. goal intention vs. implementation intention) and error rate (low vs. high) also revealed a significant main effect of intention, F(2, 68) = 4.26, p = .018, η² = .11, but no main effect of error rate, F(1, 68) = .25, p = .620. The interaction between intention and error rate did also not reach significance, F(2, 68) = .96, p = .389.

To confirm the effect of the intention manipulation on the FRN amplitude while controlling for individual differences, a regression analysis was conducted (see Table 9.4). The goal intention and the implementation intention were dummy coded with the control condition as reference group. Gender was also dummy coded (male = 1). The self-reported variables (subjective value of money, statistical skills, goal commitment) were z-standardized. To facilitate the interpretation of the results, the absolute value of the FRN amplitude was entered as depended variable. Linear regressions on the absolute value of the FRN amplitude confirmed the effect of the goal intention and the implementation intention. Model 1 showed that the goal intention as well as the implementation intention

Note that due to the averaging and differences in latency between participants, the peak amplitude in the figure differs from the reported FRN amplitude.
positively predicted the FRN amplitude meaning that the (absolute) FRN amplitude was higher in the goal intention and the implementation intention condition compared to the control condition. The counterbalancing did not affect the FRN amplitude. Model 2 added the error rate (high = 1) as predictor and confirmed that the reinforcement error rate does not predict the FRN amplitude. Model 3 added the personal characteristics gender, the subjective value of money, the statistical skills and the goal commitment as predictors. Gender (male = 1) negatively predicted the FRN amplitude, implying that the FRN amplitude was higher in females. Goal commitment was also a negative and marginally significant predictor of the FRN amplitude. This indicates that higher goal commitment is associated with a smaller FRN amplitude.

Table 9.4.: Linear regressions on FRN amplitude (absolute values)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal Intention Dummy</td>
<td>.94* (.48)</td>
<td>.93* (.48)</td>
<td>1.43*** (.52)</td>
</tr>
<tr>
<td>Implementation Intention Dummy</td>
<td>1.30*** (.48)</td>
<td>1.29*** (.48)</td>
<td>1.54*** (.48)</td>
</tr>
<tr>
<td>Counterbalance (blue = 1)</td>
<td>.14 (.39)</td>
<td>.12 (.40)</td>
<td>.11 (.41)</td>
</tr>
<tr>
<td>Error rate (1 = high)</td>
<td>.19 (.40)</td>
<td>0.06 (.40)</td>
<td></td>
</tr>
<tr>
<td>Gender (male = 1)</td>
<td>-.93** (.46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjective value of money</td>
<td>-.15 (.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical skills</td>
<td>-.01 (.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal Commitment</td>
<td>-.14* (.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>.78 (.20)</td>
<td>.73 (.22)</td>
<td>3.14** (.28)</td>
</tr>
</tbody>
</table>

N 74 74 74
R²adj. .07 .6 .10
F 2.75** 2.10* 1.98**

Notes. Standard errors in brackets. *p < .10, **p < .05, ***p < .01.
Figure 9.5.: Grand-average of difference wave (all participants)

Figure 9.6.: Grand-average of difference wave by condition
9.7. Discussion

The research aim of the present study was to investigate the modulating effect of goal intentions and implementation intentions on the FRN, an event-related potential that indicates reinforcement learning. I used the same paradigm as in Hügelschäfer and Achtziger (2017) while reducing the experimental manipulations to the goal intention and implementation intention that focused on the analyses of the decision feedback (in comparison to a control condition). The conceptual replication of Hügelschäfer and Achtziger (2017) ensured the comparability of the behavioral results while additionally exploring the effect of goal intentions and implementation intentions on the FRN.

In line with the results of Hügelschäfer and Achtziger (2017) (see also Achtziger, Alós-Ferrer, Hügelschäfer, & Steinhauser, 2015; Achtziger & Alós-Ferrer, 2014), participants committed fewer errors in case of alignment than in case of conflict (Hypothesis 1). As expected, the error rate in case of alignment was very low (8.1%), which confirms that participants had a basic understanding of the decision task. In contrast to Hypothesis 2 and the findings of Hügelschäfer and Achtziger (2017), the intention manipulations (goal intention and implementation intention) did not reduce the reinforcement error rate compared to the control condition (Hypothesis 2). Both, the goal intention as well as the implementation intention that focused on the analyses of the decision feedback failed at controlling the automaticity of reinforcement learning. The descriptive statistics even
indicated that participants in the goal intention (46%) and the implementation intention (50%) committed more errors than participants in the control condition (42%) although this difference did not reach significance.

One possible explanation for the lack of this effect is the relatively small sample size (N = 86) in this study due to the measurement of the EEG. In Study 1, Hügelschäfer and Achtziger (2017) demonstrated the effectiveness of the goal intention to reduce reinforcement errors, but the implementation intention was not successful in controlling reinforcement. The effectiveness of the implementation intention was proven in Study 2 with a bigger sample (N = 155). There are two arguments that contradict the explanation that the small sample in the present study is the reason for the lack of an effect. First, the present study does also not find the expected effect of the goal intention condition on the reinforcement error rate although the sample size of the present study is comparable to Study 1 of Hügelschäfer and Achtziger (2017) (N = 75). Secondly, when only considering the descriptive statistics, the goal intention and implementation intention rather increase reinforcement errors than decrease errors compared to the control condition (but again, this difference was not significant).

Another possible explanation could be the difference in the procedure. Due to the measurement of the EEG, there was a break between the experimental manipulation and the decision task. First, participants set the goal intention (or additionally fostered an implementation intention) and then a one-minute break followed during which the baseline EEG was measured. During the baseline EEG measurement, participants were specifically instructed to relax and not to think about anything in particular. The decision task started after the baseline measurement. It is possible that the instruction for the baseline EEG measurement counteracted the instruction of the goal intention and the implementation intention. Implementation intentions are supposed to facilitate goal achievement by making the mental representation of a specific situation highly activated and easily accessible (P. M. Gollwitzer, 1999; Achtziger & Gollwitzer, 2010). In addition, the heightened cognitive accessibility is supposed to facilitate the recall of this specific situation and subsequently the goal-directed behavior that is to be enacted (Achtziger & Gollwitzer, 2010). Maybe, the instruction for the baseline EEG measurement prevented the formation of the link between the specific situation (‘If a drawn ball is presented, …’) with the desired goal-directed behavior (‘… then I will analyze it carefully!’). The same explanation is applicable for the lack of an effect of the goal intention. By setting a goal intention, people form a mental representation of a desired end state (in this case: ‘I want to analyze the drawn ball carefully!’). The break between the goal intention
formation and the actual task with the specific instruction not to think about anything might have prevented the formation of the mental representation. Previous research has already demonstrated that mere goal intentions are often not sufficient to attain the specified goal (P. M. Gollwitzer & Sheeran, 2006; Sheeran, 2002).

When comparing the overall reinforcement error rate to the results of studies using the same paradigm (Achtziger, Alós-Ferrer, Hügelschäfer, & Steinhauser, 2015: 58.3%; Hügelschäfer & Achtziger, 2017: 57.2%), the reinforcement error rate of the present study is smaller (47.1%). This could be partly explained by the difference in the procedural setting. Compared to Hügelschäfer and Achtziger (2017), the data for the present study was collected in individual sessions. One could argue that participants are more inclined to ask questions about the task in individual settings compared to a group setting. However, in Achtziger et al. (2015), the data was also collected in individual settings. Another explanation could be that in the present study participants could ask questions after the practice trials which could have improved their understanding. It is important to note, that no questions concerning any decision strategies were answered.

In addition to the replication of the behavioral results of Hügelschäfer and Achtziger (2017), the second aim of the present study was to investigate the effect of goal intentions and implementation intentions on the FRN amplitude. Based on the findings of Hügelschäfer and Achtziger (2017) on a behavioral level, I expected the goal intention and the implementation intention to reduce the amplitude of the FRN compared to the control condition (no difference between the two intentions conditions was expected) (Hypothesis 3). Surprisingly, the results showed the opposite pattern. The FRN amplitude in the goal intention condition and the implementation intention condition was significantly higher than in the control condition. There was no difference in the amplitude between the two intention conditions. This result is surprising because the FRN is considered to be a neural indicator of automatic reinforcement learning. This means the goal intention and the implementation intention, contrary to the expectations, increased the reliance on the reinforcement heuristic at least on a neural level. In the present paradigm, the reinforcement heuristic prescribes a simple decision strategy in the form of 'win-stay' and 'lose-shift'. The decision for the second draw is exclusively based on the feedback of the first draw. The goal intention ('I want to analyze the drawn ball carefully!') and the implementation intention ('If a drawn ball is presented, then I will analyze it carefully!') might have been misinterpreted by the participants. It is possible that the phrasing of the goal intention as well as the implementation intention increased the participants' focus on the feedback of the first draw while neglecting the information about the distribution.
of the balls (state of the world). Contrary to the expectations, the goal intention and the implementation intention made the feedback valence (win/loss) of the first draw more prominent, which led participants to rely even more on the reinforcement heuristic. This effect is similar to the effect of high monetary incentives shown by Achtziger et al. (2015). The authors demonstrated that high monetary incentives (in the same paradigm) did not increase rational decision making, but instead the reliance on reinforcement as a decision strategy. The behavioral data of the present study does not fully support this notion (since there was no significant difference between the conditions in the reinforcement error rate), but it also does not contradict it.

Furthermore, the regression analysis on the FRN amplitude showed that gender is a significant predictor. Gender negatively predicted the (absolute) FRN amplitude meaning that males displayed a smaller FRN amplitude than females. This effect was independent of the intentions condition. This indicates that females rely more on the reinforcement heuristic than males. Nonetheless, this result should be interpreted with caution because it is not supported by the behavioral data. In the literature, there is no clear consensus on how individual characteristics influence the amplitude of the FRN. There are studies showing the opposite effect, meaning that males display a higher FRN amplitude than females (Larson, South, & Clayson, 2011; Yi et al., 2012). However, there are also research findings demonstrating that females have a higher FRN amplitude than males. Santesso, Dzyundzyak, and Segalowitz (2011) showed that females had a higher FRN amplitude than males in a gambling task in which participants could either win or lose money based on a binary choice. Yet, this difference was explained by the self-reported sensitivity to punishment. High punishment sensitivity was related to increased FRN amplitudes. Work by Fukushima and Hiraki (2006) showed that females’ FRN is more sensitive to socio-emotional information than males’ FRN. Participants played a competitive gambling task in which one’s monetary gain resulted in the opponent’s loss. Females perceived the opponent’s loss as negative indicated by the FRN amplitude, even if this meant that they gained money. These findings demonstrate that the picture on the gender differences in the FRN is rather complex. The design of the present study cannot provide sufficient information to explain the gender difference in the FRN amplitude.

The regression analysis showed a weak effect of goal commitment on the FRN amplitude. Higher commitment resulted in a smaller FRN amplitude which means less reliance on the reinforcement heuristic. Perhaps, participants who felt more committed to the goal (‘I want to analyze the drawn ball carefully!’), were more inclined to think beyond the mere feedback of the drawn ball, but analyze the additional information the color
of the first ball provided. Yet, this effect should be interpreted with caution since it is not supported by the behavioral data. Higher goal commitment did not lead to a decrease in the reinforcement error rate. Although the present study offers a first inside into the effect of goal intentions and implementation intentions on the neural correlate of reinforcement learning (FRN), it is not without limitations. For once, the sample size is quite limited due to the between-subjects design and the EEG measurement. As mentioned above this could be a possible explanation why the present study was not able to replicate the findings of Hügelschäfer and Achtziger (2017) on the behavioral data. It is arguable that the difference between the conditions in the reinforcement error rate could reach significance with a bigger sample size.

Furthermore, future research on this topic should investigate the effect of different phrasings of the goal intention and the implementation intention in the same paradigm. As discussed above, the specific phrasing might have encouraged participants to rely more on the reinforcement heuristic which was reflected in the difference in the FRN amplitude between the conditions. Hügelschäfer and Achtziger (2017) demonstrated that a goal intention and an implementation intention that focused on the suppression of disappointment over negative feedback either had no effect on the reinforcement error rate (goal intention) or even increased the reliance on the reinforcement heuristic (implementation intention). Work by Gollwitzer, Schwörer, Stern, Gollwitzer, and Bargha (2017) also demonstrated that depending on the phrasing, intentions have a decreasing or increasing effect on automaticity. In Study 2, the implementation intention either decreased or increased social projection depending on whether it focused on the differences or similarities of people.

Finally, future studies should account more extensively for individual characteristics that influence the modulation of the FRN (as for example gender). The limitations of the present study made it difficult to conclusively explain the gender difference in the FRN amplitude between males and females, especially since it was not reflected in the behavioral data. To conclude, results of the present research imply that the effect of goal intentions and implementation intentions on rational decision making goes beyond behavioral observations and can be detected on a neural level. Surprisingly, the intention manipulations increased the FRN indicating stronger reliance on reinforcement. Future research is needed to conclusively examine the effect of intention manipulations on rational decision making in order to be informed in which cases and under which conditions they are truly beneficial.
10. Conclusion

The aim of the present dissertation was twofold. First, I investigated confidence judgments in own skills and the confidence bias, the processes underlying these confidence judgments, and the influences of gender and monetary incentive on confidence judgments. The present research provided strong evidence for a dual stage account of confidence (as for instance the Collapsing Confidence Boundary Model by Moran et al., 2015). The confidence judgment process data (response times) indicated that there is a second stage of evidence accumulation that occurs after the initial decision is made. However, the confidence judgment is evidently not independent of the first stage (the initial choice) because the confidence level was influenced by the process information of the initial decision (see the Self-Consistency Model of Subjective Confidence by Koriat, 2012a).

The present dissertation provided novel evidence regarding the impact of monetary incentives on confidence judgments. The findings demonstrated that monetary incentives for performance increased the time invested in both tasks, meaning answering the knowledge question as well as the judging the confidence in the correctness of the answer. This suggests that monetary incentives for performance increased the effort invested in the task. Yet, the increased effort did not translate into better performance in terms of confidence judgment accuracy. Incentives for accurate confidence judgments also increased the time invested in the judgment task but incentivizing accuracy proved effective only in reducing the confidence bias in questions of medium difficulty and had no effect on the monitoring resolution (the ability to discriminate between correct and incorrect answer). This finding has two important implications for the use of monetary incentives to improve the accuracy of confidence judgments. First, offering incentives based on the performance in the task at hand has no effect on the accuracy of confidence judgments. Second, incentivizing the accuracy directly seems to be effective, but only for tasks of medium difficulty. Hence, the effectiveness of monetary incentives for improving the accuracy of confidence judgments depends strongly on the way the incentives are implemented as well as the difficulty of the task they depend on.

The second focus of this line of research was on gender differences in confidence
judgments. The present dissertation shed light on the differences between females and males in the process of judging the confidence in their own skills. The findings indicated that more controlled processes of confidence judgments (indicated by more time invested) were beneficial for the accuracy of males’ confidence judgments, decreasing overconfidence in males. The opposite was true for females. The higher contributions of automatic processes (indicated by faster responses) in females’ confidence judgments lead to rather accurate confidence judgments, whereas more controlled (slower) confidence judgments resulted in underconfidence. These findings contribute to the understanding of the processes of confidence judgments and especially the confidence bias, and they highlight the necessity to take gender into account for future research on this topic.

The second aim of the present dissertation was the investigation of the influence of goal and implementation intentions on rational decision making and how this influence is reflected in the neural correlate of reinforcement learning (FRN; see Holroyd & Coles, 2002). A recent study by Hügelschäfer and Achtziger (2017) demonstrated the effectiveness of goal and implementation intentions that focused on the decision feedback in reducing undesirable reinforcement-based decisions. The present dissertation investigated the neural basis of this effect by analyzing event-related potentials. The behavioral results of Hügelschäfer and Achtziger (2017) were not replicated. However, the analysis of the FRN gave new insights into the neural processes underlying the effect of goal and implementation intentions. Surprisingly, the manipulation of goal and implementation intentions, hypothesized to reduce reinforcement-based suboptimal decisions, increased the amplitude of the FRN. This indicated a stronger reliance on the reinforcement heuristic. One possible explanation for this unexpected finding is that the phrasing of the intentions in the instructions might have promoted the reliance on the reinforcement heuristic by pulling the participants focus more on the decision feedback which was used for reinforcement learning (for evidence on the intensifying effect of goal and implementation intentions on automatic processes see also A. Gollwitzer et al., 2017). This finding has strong implications for the use of implementation intentions as an intervention to foster rational decision making. Even though intentions can serve as a readily available, low-cost intervention, they should not be taken lightly. Diligent consideration of the intentions’ potential effects on choices and decision processes is required to warrant their successful implementation and to avoid undesired side-effects. The present findings demonstrated that goal and implementation intention effects on decision processes are rather complex, and that future research is needed to fully understand the processes behind these intentions.
In sum, the findings of the present dissertation contributed to the fields of psychology and (neuro-) economics by providing new insights into the processes underlying confidence judgments and rational decision making. The present research advanced the debate about possible interventions, to improve the accuracy of confidence judgments. Furthermore, novel evidence was presented that informed the discussion about interventions aimed to foster rational decision making.
References


Bayes, T., & Price, R. (1763). An essay toward solving a problem in the doctrine of


Huang, J., & Kisgen, D. J. (2013). Gender and corporate finance: Are male executives


Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter,


Magnus, J. R., & Peresetsky, A. A. (2018). Grade Expectations: Rationality and


negativity reflects detection of negative reward prediction error. *Neuroreport, 15*(16), 2561–2565.


# A. Appendix

## A.1. Pretest

<table>
<thead>
<tr>
<th>Item</th>
<th>$T$</th>
<th>$df$</th>
<th>$M$</th>
<th>$\Delta M$</th>
<th>Item</th>
<th>$T$</th>
<th>$df$</th>
<th>$M$</th>
<th>$\Delta M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item_1</td>
<td>4.78</td>
<td>66</td>
<td>.51</td>
<td>.51***</td>
<td>Item_43</td>
<td>-1.14</td>
<td>68</td>
<td>.52</td>
<td>.14</td>
</tr>
<tr>
<td>Item_2</td>
<td>4.20</td>
<td>66</td>
<td>.43</td>
<td>.46***</td>
<td>Item_44</td>
<td>.50</td>
<td>65</td>
<td>.81</td>
<td>.05</td>
</tr>
<tr>
<td>Item_3</td>
<td>3.68</td>
<td>67.80</td>
<td>.66</td>
<td>.37***</td>
<td>Item_45</td>
<td>-.21</td>
<td>66</td>
<td>.97</td>
<td>.01</td>
</tr>
<tr>
<td>Item_4</td>
<td>.62</td>
<td>68</td>
<td>0.57</td>
<td>.07</td>
<td>Item_46</td>
<td>.38</td>
<td>2.51</td>
<td>55.62</td>
<td>.30**</td>
</tr>
<tr>
<td>Item_5</td>
<td>.70</td>
<td>68</td>
<td>0.39</td>
<td>.08</td>
<td>Item_47</td>
<td>1.81</td>
<td>60.65</td>
<td>.85</td>
<td>.15</td>
</tr>
<tr>
<td>Item_6</td>
<td>1.28</td>
<td>67.63</td>
<td>.84</td>
<td>.11</td>
<td>Item_48</td>
<td>.19</td>
<td>66</td>
<td>.69</td>
<td>.02</td>
</tr>
<tr>
<td>Item_7</td>
<td>1.95</td>
<td>65.22</td>
<td>.84</td>
<td>.16*</td>
<td>Item_49</td>
<td>2.20</td>
<td>69</td>
<td>.44</td>
<td>.26**</td>
</tr>
<tr>
<td>Item_8</td>
<td>2.93</td>
<td>59.89</td>
<td>.40</td>
<td>.37***</td>
<td>Item_50</td>
<td>.81</td>
<td>69</td>
<td>.96</td>
<td>.04</td>
</tr>
<tr>
<td>Item_9</td>
<td>2.70</td>
<td>65.73</td>
<td>.44</td>
<td>.31***</td>
<td>Item_51</td>
<td>.11</td>
<td>66</td>
<td>.44</td>
<td>.01</td>
</tr>
<tr>
<td>Item_10</td>
<td>1.09</td>
<td>68</td>
<td>.44</td>
<td>.13</td>
<td>Item_52</td>
<td>1.72</td>
<td>68</td>
<td>.77</td>
<td>.17</td>
</tr>
<tr>
<td>Item_11</td>
<td>2.21</td>
<td>67</td>
<td>.42</td>
<td>.26**</td>
<td>Item_53</td>
<td>1.38</td>
<td>55.64</td>
<td>.25</td>
<td>.15</td>
</tr>
<tr>
<td>Item_12</td>
<td>.52</td>
<td>69</td>
<td>.85</td>
<td>.05</td>
<td>Item_54</td>
<td>.61</td>
<td>69</td>
<td>.30</td>
<td>.07</td>
</tr>
<tr>
<td>Item_13</td>
<td>1.93</td>
<td>67</td>
<td>.54</td>
<td>.23*</td>
<td>Item_55</td>
<td>.60</td>
<td>69</td>
<td>.62</td>
<td>.07</td>
</tr>
<tr>
<td>Item_14</td>
<td>1.75</td>
<td>66</td>
<td>.51</td>
<td>.21</td>
<td>Item_56</td>
<td>1.74</td>
<td>68</td>
<td>.51</td>
<td>.21</td>
</tr>
<tr>
<td>Item_15</td>
<td>.53</td>
<td>69</td>
<td>.92</td>
<td>.04</td>
<td>Item_57</td>
<td>2.88</td>
<td>39</td>
<td>.90</td>
<td>.18***</td>
</tr>
<tr>
<td>Item_16</td>
<td>1.89</td>
<td>59.79</td>
<td>.41</td>
<td>.22</td>
<td>Item_58</td>
<td>.16</td>
<td>7691</td>
<td>.59</td>
<td>.02</td>
</tr>
<tr>
<td>Item_17</td>
<td>.92</td>
<td>65</td>
<td>.25</td>
<td>.10</td>
<td>Item_59</td>
<td>.14</td>
<td>68</td>
<td>.43</td>
<td>.02</td>
</tr>
<tr>
<td>Item_18</td>
<td>.48</td>
<td>65</td>
<td>.75</td>
<td>.05</td>
<td>Item_60</td>
<td>.84</td>
<td>67</td>
<td>.62</td>
<td>.10</td>
</tr>
<tr>
<td>Item_19</td>
<td>.69</td>
<td>69</td>
<td>.27</td>
<td>.07</td>
<td>Item_61</td>
<td>-1.44</td>
<td>30</td>
<td>.97</td>
<td>-.06</td>
</tr>
<tr>
<td>Item_20</td>
<td>1.64</td>
<td>56.67</td>
<td>.26</td>
<td>.18</td>
<td>Item_62</td>
<td>1.50</td>
<td>65.38</td>
<td>.87</td>
<td>.11</td>
</tr>
<tr>
<td>Item_21</td>
<td>1.94</td>
<td>67</td>
<td>.55</td>
<td>.23*</td>
<td>Item_63</td>
<td>2.57</td>
<td>60.93</td>
<td>.62</td>
<td>.3**</td>
</tr>
<tr>
<td>Item</td>
<td>Value</td>
<td>Value</td>
<td>Value</td>
<td>Value</td>
<td>Item</td>
<td>Value</td>
<td>Value</td>
<td>Value</td>
<td>Value</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Item_22</td>
<td>.57</td>
<td>68</td>
<td>.84</td>
<td>.05</td>
<td>Item_64</td>
<td>1.32</td>
<td>61.15</td>
<td>.88</td>
<td>.10</td>
</tr>
<tr>
<td>Item_23</td>
<td>.36</td>
<td>69</td>
<td>.96</td>
<td>.02</td>
<td>Item_65</td>
<td>2.03</td>
<td>64.74</td>
<td>.57</td>
<td>.23**</td>
</tr>
<tr>
<td>Item_24</td>
<td>.88</td>
<td>67</td>
<td>.99</td>
<td>.03</td>
<td>Item_66</td>
<td>1.37</td>
<td>60.40</td>
<td>.58</td>
<td>.17</td>
</tr>
<tr>
<td>Item_25</td>
<td>-.29</td>
<td>64</td>
<td>.61</td>
<td>.04</td>
<td>Item_67</td>
<td>1.69</td>
<td>56.03</td>
<td>.25</td>
<td>.18</td>
</tr>
<tr>
<td>Item_26</td>
<td>.23</td>
<td>66</td>
<td>.78</td>
<td>.02</td>
<td>Item_68</td>
<td>-1.19</td>
<td>43.33</td>
<td>.94</td>
<td>.07</td>
</tr>
<tr>
<td>Item_27</td>
<td>1.20</td>
<td>63</td>
<td>.83</td>
<td>.11</td>
<td>Item_69</td>
<td>1.43</td>
<td>39</td>
<td>.97</td>
<td>.05</td>
</tr>
<tr>
<td>Item_28</td>
<td>.40</td>
<td>68</td>
<td>.54</td>
<td>.05</td>
<td>Item_70</td>
<td>1.67</td>
<td>67.63</td>
<td>.67</td>
<td>.18</td>
</tr>
<tr>
<td>Item_29</td>
<td>.29</td>
<td>64</td>
<td>.68</td>
<td>0.03</td>
<td>Item_71</td>
<td>.39</td>
<td>68</td>
<td>.96</td>
<td>.02</td>
</tr>
<tr>
<td>Item_30</td>
<td>3.19</td>
<td>65.96</td>
<td>.60</td>
<td>.35***</td>
<td>Item_72</td>
<td>.83</td>
<td>68</td>
<td>.83</td>
<td>.08</td>
</tr>
<tr>
<td>Item_31</td>
<td>2.00</td>
<td>60.98</td>
<td>.39</td>
<td>.23**</td>
<td>Item_73</td>
<td>1.74</td>
<td>55.37</td>
<td>.33</td>
<td>.21</td>
</tr>
<tr>
<td>Item_32</td>
<td>1.78</td>
<td>37</td>
<td>.04</td>
<td>.08</td>
<td>Item_74</td>
<td>1.13</td>
<td>66.86</td>
<td>.67</td>
<td>.13</td>
</tr>
<tr>
<td>Item_33</td>
<td>.02</td>
<td>65</td>
<td>.87</td>
<td>.001</td>
<td>Item_75</td>
<td>.03</td>
<td>68</td>
<td>.51</td>
<td>.003</td>
</tr>
<tr>
<td>Item_34</td>
<td>1.54</td>
<td>65.54</td>
<td>.74</td>
<td>.16</td>
<td>Item_76</td>
<td>.11</td>
<td>67</td>
<td>.90</td>
<td>.01</td>
</tr>
<tr>
<td>Item_35</td>
<td>.18</td>
<td>66</td>
<td>.74</td>
<td>.02</td>
<td>Item_77</td>
<td>.38</td>
<td>69</td>
<td>.62</td>
<td>.05</td>
</tr>
<tr>
<td>Item_36</td>
<td>.03</td>
<td>68</td>
<td>.51</td>
<td>.003</td>
<td>Item_78</td>
<td>2.84</td>
<td>58.34</td>
<td>.80</td>
<td>.25***</td>
</tr>
<tr>
<td>Item_37</td>
<td>1.53</td>
<td>60.74</td>
<td>.35</td>
<td>.18</td>
<td>Item_79</td>
<td>.22</td>
<td>68</td>
<td>.73</td>
<td>.02</td>
</tr>
<tr>
<td>Item_38</td>
<td>2.40</td>
<td>52.64</td>
<td>.87</td>
<td>.18**</td>
<td>Item_80</td>
<td>.23</td>
<td>67</td>
<td>.93</td>
<td>.01</td>
</tr>
<tr>
<td>Item_39</td>
<td>.16</td>
<td>66</td>
<td>.56</td>
<td>.01</td>
<td>Item_81</td>
<td>-.15</td>
<td>68</td>
<td>.81</td>
<td>.01</td>
</tr>
<tr>
<td>Item_40</td>
<td>2.35</td>
<td>68</td>
<td>.43</td>
<td>.28**</td>
<td>Item_82</td>
<td>.56</td>
<td>67</td>
<td>.36</td>
<td>.07</td>
</tr>
<tr>
<td>Item_41</td>
<td>1.84</td>
<td>64.85</td>
<td>.81</td>
<td>.17</td>
<td>Item_83</td>
<td>.89</td>
<td>68</td>
<td>.99</td>
<td>.03</td>
</tr>
<tr>
<td>Item_42</td>
<td>4.28</td>
<td>64.86</td>
<td>.58</td>
<td>.45***</td>
<td>Item_84</td>
<td>.54</td>
<td>66</td>
<td>.40</td>
<td>.06</td>
</tr>
</tbody>
</table>

Notes. *p < .10, **p < .05, ***p < .01.
A.2. Material: Confidence Judgments

Selected Knowledge Questions by Difficulty Level

easy

Item_12: Womit stand das folgende Symbol ursprünglich in Verbindung? 
  a) Psychedelische Drogen
  b) ökologisches Denken
  c) eine östliche Religion
  d) Astrologie
  e) die Hippiebewegung

Item_22: Mit welchem Symbol stellt man elektrische Lichter dar?

Item_23: Welches Symbol bedeutet "Pause"?

Item_33: Wie viele Bundesländer hat die Bundesrepublik Deutschland?
  a) 21
  b) 18
  c) 11
  d) 16
  e) 13

Item_44: Wer erfand die Glühbirne?
  a) Bell & Reis
  b) Faraday
  c) Hertz
  d) Edison
  e) Watt

Item_45: Welches Ergebnis ist die richtige Lösung für die folgende Aufgabe? 16 - (-9)
  a) 25
  b) 7
  c) -7
  d) -25
  e) 5
Item_71: Wie hoch ist die erlaubte Höchstgeschwindigkeit in geschlossenen Ortschaften?
   a) 30
   b) 40
   c) 50
   d) 60
   e) 70

Item_76: Wie viele Geißlein gibt es im entsprechenden Märchen der Gebrüder Grimm?
   a) 5
   b) 7
   c) 8
   d) 12
   e) 13

Item_80: Was heißt ABS?
   a) Automatisches Betriebssystem
   b) Allgemeiner Betriebsschutz
   c) Auto-Brems-System
   d) Anti-Beschlag-System
   e) Anti-Blockier-System

Item_81: Welches grafische Zeichen kommt im Standard-Notensatz der Musik nicht vor?

   a) [ ]
   b) [ ]
   c) [ ]
   d) [ ]
   e) [ ]

medium

Item_25: In welchem Jahr sank die Titanic?
   a) 1908
   b) 1905
   c) 1914
   d) 1920
   e) 1912
Item_26: Welches Land ist auf der schematischen Darstellung Südamerikas dargestellt (schwarz)?

Item_29: 'Der Steppenwolf' ist ein Roman von...
   a) Stefan Zweig
   b) Hermann Hesse
   c) Anna Seghers
   d) Günter Grass
   e) Rainer Maria Rilke

Item_36: Wie viele Teile hat Goethes Theaterstück 'Faust'?
   a) 1
   b) 2
   c) 3
   d) 4
   e) 5

Item_39: Wie viele Sinfonien hat Beethoven komponiert?
   a) 5
   b) 7
   c) 8
   d) 9
   e) 12

Item_48: Was ergibt folgende Berechnung? 1/\frac{3}{4}
   a) 0.125
   b) 4.5
c) 2

d) $\frac{1}{2}$
e) $\frac{1}{4}$

Item_58: Wie verlaufen die Feldlinien eines elektrischen Dipols?

Item_75: In der SCHUFA sind zum Zwecke der Kreditsicherung organisiert:

a) Verbraucherverbände

b) alle Gläubiger der Bundesverwaltungsstelle

c) die Interessenvertreter der Privatgläubiger Deutschlands

d) die fünf größten Konzerne Deutschlands

e) Unternehmen und Kreditinstitute

Item_77: Aus welcher Zeit stammt diese Zeichnung?

a) Renaissance

b) Mittelalter

c) Neuzeit

d) Steinzeit

e) Kreidezeit
Item_79: Wie viele Nächte hindurch erzählte Scheherezade ihre Geschichten?
a) 7  
b) 13  
c) 500  
d) 1000  
e) 1001

difficult

Item_5: Wie viele Wasserstoffatome enthält Alkohol (Ethanol)?
a) 0  
b) 4  
c) 6  
d) 1  
e) 3

Item_10: Welche Oper stammt nicht von Mozart?
a) Don Giovanni  
b) Zauberflöte  
c) Die Hochzeit des Figaro  
d) La Bohème  
e) Cosi fan tutte

Item_19: Wofür steht das folgende Symbol? \(\Omega\)
a) „Schmied“ in der Handwerkssymbolik  
b) „Widerstand“ in Schaltplänen  
c) „Ende der Reihe“ in der Mathematik  
d) „Waage“ in der Astrologie  
e) „Wasserspeicher“ in Bauplänen

Item_20: Wer ist als Gründer der Olympischen Spiele der Neuzeit bekannt?
a) A. Onassis  
b) P. de Coubertin  
c) K. Kavaphis  
d) Sadi-Carnot  
e) Roosevelt

Item_37: Wie nennt man folgende Figur?

XIII
a) Pyramide
b) Trapez
c) Drachen
d) Raute
e) Tetraeder

Item_53: Die Formel für die kinetische Energie $E$ eines Körpers mit der Masse $m$ lautet:
a) $E = mc^2$
b) $E = 2mv$
c) $E = 3v/m$
d) $E = 0.5mv^2$
e) $E = m^2v$

Item_54: Wie viele Töne enthält Schönbergs neue Tontechnik?
a) 5
b) 7
c) 8
d) 9
e) 12
Item_59: In welcher Theorie spielt das "Gesetz der großen Zahl" eine entscheidende Rolle?
   a) Wahrscheinlichkeitstheorie
   b) Spezielle Relativitätstheorie
   c) Algebra
   d) Zahlentheorie
   e) Logik

Item_82: Der typische Grundriss welchen Gebäudes ist hier dargestellt?

   a) ägyptischer Tempel
   b) römisches Privathaus
   c) romanische Basilika
   d) gotische Kirche
   e) Renaissancepalast

Item_84: Welche Behörde ist für die Rundfunkgebührenbefreiung zuständig?
   a) Telekom
   b) Jugendamt
   c) Kulturamt
   d) ARD
   e) Sozialamt
A.3. Material: Intentions & Rationality

Questionnaire Control Condition (blue balls rewarded)

Abschlussfragebogen

Bitte beantworten Sie die folgenden Fragen zu Ihrer Person spontan, ohne lange nachzudenken!
IHRE DATEN WERDEN ABSOLUT ANONYM BEHANDELT!

Falls Sie Fragen haben, wenden Sie sich bitte an die Versuchsleiterin bzw. den Versuchsleiter!

Erstmal ein paar kurze Fragen zu dem Experiment an sich

1. Wie verständlich fanden Sie die Instruktionen des Entscheidungsspiels?

2. Wie sehr bemühten Sie sich darum, das Entscheidungsspiel präzise durchzuführen?

3. Wie schwierig fanden Sie das Entscheidungsspiel?

4. Wie wichtig war es Ihnen im Entscheidungsspiel gut abzuschneiden?

5. Sehen Sie sich generell in der Lage, schwierige Aufgaben gut meistern zu können?

Wenn Sie an das Ziel „Ich will die gezogenen Kugeln genau analysieren!“ denken …

6. Wie wichtig war Ihnen dieses Ziel?

7. Wie sehr haben Sie sich vorgenommen, dieses Ziel zu erreichen?

8. Wie sehr haben Sie sich verpflichtet gefühlt, dieses Ziel zu erreichen?
9. Wie sehr hat Ihnen dieses Ziel bei den Entscheidungen geholfen?

   mehr | sehr
   --- | ---
   gar nicht | sehr

10. Wie wichtig war es Ihnen, die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

   mehr | sehr
   --- | ---
   gar nicht | sehr

11. Wie sehr haben Sie sich vorgenommen die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

   mehr | sehr
   --- | ---
   gar nicht | sehr

12. Wie sehr haben Sie sich verpflichtet gefühlt, die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

   mehr | sehr
   --- | ---
   gar nicht | sehr

13. Wie sehr hat Ihnen dieser Satz bei den Entscheidungen geholfen?

   mehr | sehr
   --- | ---
   gar nicht | sehr

Bitte beantworten Sie die folgenden Fragen spontan, ohne lange nachzudenken!

14. Stellen Sie sich vor, Sie hätten 150 € im Lotto gewonnen. Wie sehr würden Sie sich über den Gewinn freuen?

   mehr | sehr
   --- | ---
   gar nicht | sehr

15. Stellen Sie sich vor, Sie hätten 30 € verloren. Wie sehr würden Sie den Verlust bedauern?

   mehr | sehr
   --- | ---
   gar nicht | sehr

16. Stellen Sie sich bitte Folgendes vor: Sie haben gerade ein Fernsehgerät für 500 € gekauft. Ein Bekannter erzählt Ihnen, dass er nur 450 € für genau das gleiche Gerät bezahlt hat. Wie sehr würden Sie sich deswegen ärgern?

   mehr | sehr
   --- | ---
   gar nicht | sehr

17. Wie viel Geld brauchen Sie in einem durchschnittlichen Monat unbedingt zum Leben (mit Miete)?

   _________ €
18. Wie sehr hat Sie die Möglichkeit, durch geschickte Entscheidungen Geld hinzu zu gewinnen, zur Teilnahme an diesem Experiment motiviert?

19. Wie sehr hat Sie die Möglichkeit, Geld hinzu zu gewinnen, dazu angespornt, im Experiment möglichst gut abzuschneiden?

20. Wie schätzen Sie Ihre allgemeinen statistischen Vorkenntnisse ein?

Wodurch haben Sie diese ggf. erworben?

21. Wie schätzen Sie Ihre Fähigkeiten in der Berechnung von Wahrscheinlichkeiten ein?

22. Stellen Sie sich vor, ein fairer, sechseitiger Würfel würde 1000-mal geworfen. Was denken Sie, in wie vielen von diesen 1000 Würfen eine gerade Zahl (2, 4 oder 6) geworfen würde?

23. In einer Lotterie betragen die Chancen auf einen Gewinn von 10 Euro 1%. Was denken Sie, wie viele Leute einen Gewinn von 10 Euro erhalten würden, wenn 1000 Menschen jeweils ein Los für die Lotterie kaufen würden?

24. Bei einer Verlosung besteht eine Chance von 1 zu 1000, ein Auto zu gewinnen. Welcher Prozentsatz an Losen gewinnt ein Auto?

25. Wie gut sind Sie im Bruchrechnen?

26. Wie gut sind Sie im Prozentrechnen?
27. Wie gut sind Sie im Berechnen eines Trinkgelds von 15%?

überhaupt 

nicht gut 

extrem gut

28. Wie gut sind Sie im Berechnen des Preises eines Shirts, das 25% reduziert ist?

überhaupt 

nicht gut 

extrem gut

29. Wie hilfreich finden Sie Tabellen und Diagramme als Zusatzinformation beim Lesen von Zeitungsartikeln?

überhaupt 

nicht 

extrem

30. Wenn Ihnen jemand die Wahrscheinlichkeit eines möglichen Ereignisses mitteilt, bevorzugen Sie diese in Worten ausgedrückt (z.B. „es passiert selten“) oder in Zahlen ausgedrückt (z.B. „die Wahrscheinlichkeit beträgt 1%“)?

bevorzuge 

immer 

Wörter 

bevorzuge 

immer 

Zahlen

31. Wenn Sie sich die Wettervorhersage anhören, bevorzugen Sie die Prognosen in Prozent ausgedrückt (z.B. „die Regenwahrscheinlichkeit beträgt heute 20%“) oder in Worten ausgedrückt (z.B. „die Regenwahrscheinlichkeit ist heute gering“)?

bevorzuge 

immer 

Prozent 

bevorzuge 

immer 

Wörter

32. Wie oft finden Sie in Zahlen ausgedrückte Informationen nützlich?

die 

sehr oft

33. Haben Sie bereits Erfahrung mit dieser Art von Entscheidungssituationen/Szenarien?

Ja

Nein

Falls ja, in welchem Zusammenhang haben Sie diese gemacht?
34. Erinnern Sie sich noch, welcher Betrag Ihnen im Experiment für jede blaue Kugel gutgeschrieben wurde? __________ Cent

35. Auf welchen Überlegungen beruhten Ihre Entscheidungen während der Studie?

Bitte stellen Sie uns für die wissenschaftliche Auswertung noch einige Informationen über Ihre Person zur Verfügung.

1. Ihr Alter: ___________________________

2. Ihr Geschlecht: □ weiblich □ männlich

3. Muttersprache: _______________________

4. Hauptfach: _______________________

5. Nebenfach: _______________________

6. Anzahl der Fachsemester: _______________________

7. Ihre Leistungskurse im Abi? _______________________

8. Ihr Notendurchschnitt im Abi? _______________________

9. Das Bundesland, in dem Sie Abi gemacht haben? _______________________

10. Ihre letzte Mathematiknote? _______________________

11. Ihr Berufswunsch? _______________________

Ihre Daten werden absolut ANONYM behandelt!

Vielen Dank für Ihre Mitarbeit!
Abschlussfragebogen

Bitte beantworten Sie die folgenden Fragen zu Ihrer Person spontan, ohne lange nachzudenken! IHRE DATEN WERDEN ABSOLUT ANONYM BEHANDELT!

Falls Sie Fragen haben, wenden Sie sich bitte an die Versuchsleiterin bzw. den Versuchsleiter!

Erstmal ein paar kurze Fragen zu dem Experiment an sich

1. Wie verständlich fanden Sie die Instruktionen des Entscheidungsspiels?
   - gar nicht
   - sehr

2. Wie sehr bemühten Sie sich darum, das Entscheidungsspiel präzise durchzuführen?
   - gar nicht
   - sehr

3. Wie schwierig fanden Sie das Entscheidungsspiel?
   - gar nicht
   - sehr

4. Wie wichtig war es Ihnen im Entscheidungsspiel gut abzuschneiden?
   - gar nicht
   - sehr

5. Sehen Sie sich generell in der Lage, schwierige Aufgaben gut meistern zu können?
   - gar nicht
   - sehr

Wenn Sie an das Ziel „Ich will die gezogenen Kugeln genau analysieren!“ denken …

6. Wie wichtig war Ihnen dieses Ziel?
   - gar nicht
   - sehr

7. Wie sehr haben Sie sich vorgenommen, dieses Ziel zu erreichen?
   - gar nicht
   - sehr

8. Wie sehr haben Sie sich verpflichtet gefühlt, dieses Ziel zu erreichen?
   - gar nicht
   - sehr
9. Wie sehr hat Ihnen dieses Ziel bei den Entscheidungen geholfen?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

10. Wie wichtig war es Ihnen, die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

11. Wie sehr haben Sie sich vorgenommen die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

12. Wie sehr haben Sie sich verpflichtet gefühlt, die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

13. Wie sehr hat Ihnen dieser Satz bei den Entscheidungen geholfen?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

Bitte beantworten Sie die folgenden Fragen spontan, ohne lange nachzudenken!

14. Stellen Sie sich vor, Sie hätten 150 € im Lotto gewonnen. Wie sehr würden Sie sich über den Gewinn freuen?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

15. Stellen Sie sich vor, Sie hätten 30 € verloren. Wie sehr würden Sie den Verlust bedauern?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

16. Stellen Sie sich bitte Folgendes vor: Sie haben gerade ein Fernsehgerät für 500 € gekauft. Ein Bekannter erzählt Ihnen, dass er nur 450 € für genau das gleiche Gerät bezahlt hat. Wie sehr würden Sie sich deswegen ärgern?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
</table>

17. Wie viel Geld brauchen Sie in einem durchschnittlichen Monat unbedingt zum Leben (mit Miete)?

| ________€ |

- 2 -
18. Wie sehr hat Sie die Möglichkeit, durch geschickte Entscheidungen Geld hinzu zu gewinnen, zur Teilnahme an diesem Experiment motiviert?

19. Wie sehr hat Sie die Möglichkeit, Geld hinzu zu gewinnen, dazu angesprochen, im Experiment möglichst gut abzuschneiden?

20. Wie schätzen Sie Ihre allgemeinen statistischen Vorkenntnisse ein?

21. Wie schätzen Sie Ihre Fähigkeiten in der Berechnung von Wahrscheinlichkeiten ein?

22. Stellen Sie sich vor, ein fairer, sechseitiger Würfel würde 1000-mal geworfen. Was denken Sie, in wie vielen von diesen 1000 Würfen eine gerade Zahl (2, 4 oder 6) geworfen würde?

23. In einer Lotterie betragen die Chancen auf einen Gewinn von 10 Euro 1%. Was denken Sie, wie viele Leute einen Gewinn von 10 Euro erhalten würden, wenn 1000 Menschen jeweils ein Los für die Lotterie kaufen würden?

24. Bei einer Verlosung besteht eine Chance von 1 zu 1000, ein Auto zu gewinnen. Welcher Prozentsatz an Losen gewinnt ein Auto?

25. Wie gut sind Sie im Bruchrechnen?

26. Wie gut sind Sie im Prozentrechnen?
27. Wie gut sind Sie im Berechnen eines Trinkgelds von 15%?

überhaupt nicht gut  extreme gut

28. Wie gut sind Sie im Berechnen des Preises eines Shirts, das 25% reduziert ist?

überhaupt nicht gut  extreme gut

29. Wie hilfreich finden Sie Tabellen und Diagramme als Zusatzinformation beim Lesen von Zeitungsartikeln?

überhaupt nicht  extreme

30. Wenn Ihnen jemand die Wahrscheinlichkeit eines möglichen Ereignisses mitteilt, bevorzugen Sie diese in Worten ausgedrückt (z.B. „es passiert selten“) oder in Zahlen ausgedrückt (z.B. „die Wahrscheinlichkeit beträgt 1%“)?

bevorzuge immer  bevorzuge immer

Wörter   Zahlen

31. Wenn Sie sich die Wetterprognose anhören, bevorzugen Sie die Prognosen in Prozent ausgedrückt (z.B. „die Regenwahrscheinlichkeit beträgt heute 20%“) oder in Worten ausgedrückt (z.B. „die Regenwahrscheinlichkeit ist heute gering“)?

bevorzuge immer  bevorzuge immer

Prozent  Wörter

32. Wie oft finden Sie in Zahlen ausgedrückte Informationen nützlich?

sehr oft  nie

33. Haben Sie bereits Erfahrung mit dieser Art von Entscheidungssituationen/Szenarien?

Ja  Nein

Falls ja, in welchem Zusammenhang haben Sie diese gemacht?
34. Erinnern Sie sich noch, welcher Betrag Ihnen im Experiment für jede blaue Kugel gutgeschrieben wurde?

__________ Cent

35. Auf welchen Überlegungen beruhten Ihre Entscheidungen während der Studie?


Bitte stellen Sie uns für die wissenschaftliche Auswertung noch einige Informationen über Ihre Person zur Verfügung.

1. Ihr Alter: ____________________________

2. Ihr Geschlecht: □ weiblich □ männlich

3. Muttersprache: ____________________________

4. Hauptfach: ____________________________

5. Nebenfach: ____________________________

6. Anzahl der Fachsemester: ____________________________

7. Ihre Leistungskurse im Abi?: ____________________________

8. Ihr Notendurchschnitt im Abi?: ____________________________

9. Das Bundesland, in dem Sie Abi gemacht haben?: ____________________________

10. Ihre letzte Mathematiknote?: ____________________________

11. Ihr Berufswunsch?: ____________________________

Ihre Daten werden absolut ANONYM behandelt!

Vielen Dank für Ihre Mitarbeit!

- 5 -

XXV
Abschlussfragebogen

Bitte beantworten Sie die folgenden Fragen zu Ihrer Person spontan, ohne lange nachzudenken!
IHRE DATEN WERDEN ABSOLUT ANONYM BEHANDELT!

Falls Sie Fragen haben, wenden Sie sich bitte an die Versuchsleiterin bzw. den Versuchsleiter!

Erstmal ein paar kurze Fragen zu dem Experiment an sich

1. Wie verständlich fanden Sie die Instruktionen des Entscheidungsspiels?

   gar nicht  sehr

2. Wie sehr bemühten Sie sich darum, das Entscheidungsspiel präzise durchzuführen?

   gar nicht  sehr

3. Wie schwierig fanden Sie das Entscheidungsspiel?

   gar nicht  sehr

4. Wie wichtig war es Ihnen im Entscheidungsspiel gut abzuschneiden?

   gar nicht  sehr

5. Sehen Sie sich generell in der Lage, schwierige Aufgaben gut meistern zu können?

   gar nicht  sehr

Wenn Sie an das Ziel „Ich will die gezogenen Kugeln genau analysieren!“ denken …

6. Wie wichtig war Ihnen dieses Ziel?

   gar nicht  sehr

7. Wie sehr haben Sie sich vorgenommen, dieses Ziel zu erreichen?

   gar nicht  sehr

8. Wie sehr haben Sie sich verpflichtet gefühlt, dieses Ziel zu erreichen?

   gar nicht  sehr
9. Wie sehr hat Ihnen dieses Ziel bei den Entscheidungen geholfen?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10. Wie wichtig war es Ihnen, die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11. Wie sehr haben Sie sich vorgenommen die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

12. Wie sehr haben Sie sich verpflichtet gefühlt, die gezogenen Kugeln zu analysieren und sorgfältig nachzudenken?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

13. Wie sehr hat Ihnen dieser Satz bei den Entscheidungen geholfen?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Wenn Sie an den Plan „Immer wenn eine gezogene Kugel auftaucht, dann analysiere ich diese genau!“ denken …

9. Wie wichtig war Ihnen dieser Plan?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10. Wie sehr haben Sie sich vorgenommen, nach diesem Plan zu arbeiten?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11. Wie sehr haben Sie sich verpflichtet gefühlt, diesen Plan anzuwenden?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

14. Wie sehr hat Ihnen dieser Plan bei den Entscheidungen geholfen?

<table>
<thead>
<tr>
<th>gar nicht</th>
<th>sehr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

XXVII
Bitte beantworten Sie die folgenden Fragen spontan, ohne lange nachzudenken!

14. Stellen Sie sich vor, Sie hätten 150 € im Lotto gewonnen. Wie sehr würden Sie sich über den Gewinn freuen?
   - gar nicht
   - sehr

15. Stellen Sie sich vor, Sie hätten 30 € verloren. Wie sehr würden Sie den Verlust bedauern?
   - gar nicht
   - sehr

16. Stellen Sie sich bitte Folgendes vor: Sie haben gerade ein Fernsehgerät für 500 € gekauft. Ein Bekannter erzählt Ihnen, dass er nur 450 € für genau das gleiche Gerät gezahlt hat. Wie sehr würden Sie sich deswegen ärgern?
   - gar nicht
   - sehr

17. Wie viel Geld brauchen Sie in einem durchschnittlichen Monat unbedingt zum Leben (mit Miete)?
   ________ €

18. Wie sehr hat Sie die Möglichkeit, durch geschickte Entscheidungen Geld hinzu zu gewinnen, zur Teilnahme an diesem Experiment motiviert?
   - gar nicht
   - sehr

19. Wie sehr hat Sie die Möglichkeit, Geld hinzu zu gewinnen, dazu angespornt, im Experiment möglichst gut abzuschneiden?
   - gar nicht
   - sehr

20. Wie schätzen Sie Ihre allgemeinen statistischen Vorkenntnisse ein?
    - sehr
    - sehr gut
    - schlecht
    - sehr gut

   Wodurch haben Sie diese ggf. erworben?

21. Wie schätzen Sie Ihre Fähigkeiten in der Berechnung von Wahrscheinlichkeiten ein?
    - sehr
    - schlecht
    - sehr gut
22. Stellen Sie sich vor, ein fairer, sechsseitiger Würfel würde 1000-mal geworfen. Was denken Sie, in wie vielen von diesen 1000 Würfen eine gerade Zahl (2, 4 oder 6) geworfen würde?

23. In einer Lotterie betragen die Chancen auf einen Gewinn von 10 Euro 1%. Was denken Sie, wie viele Leute einen Gewinn von 10 Euro erhalten würden, wenn 1000 Menschen jeweils ein Los für die Lotterie kaufen würden?

24. Bei einer Verlosung besteht eine Chance von 1 zu 1000, ein Auto zu gewinnen. Welcher Prozentsatz an Losen gewinnt ein Auto?

25. Wie gut sind Sie im Bruchrechnen?

<table>
<thead>
<tr>
<th></th>
<th>extrem gut</th>
<th>nicht gut</th>
<th>überhaupt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

26. Wie gut sind Sie im Prozentrechnen?

<table>
<thead>
<tr>
<th></th>
<th>extrem gut</th>
<th>nicht gut</th>
<th>überhaupt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

27. Wie gut sind Sie im Berechnen eines Trinkgelds von 15%?

<table>
<thead>
<tr>
<th></th>
<th>extrem gut</th>
<th>nicht gut</th>
<th>überhaupt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

28. Wie gut sind Sie im Berechnen des Preises eines Shirts, das 25% reduziert ist?

<table>
<thead>
<tr>
<th></th>
<th>extrem gut</th>
<th>nicht gut</th>
<th>überhaupt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

29. Wie hilfreich finden Sie Tabellen und Diagramme als Zusatzinformation beim Lesen von Zeitungsartikeln?

<table>
<thead>
<tr>
<th></th>
<th>extrem</th>
<th>nicht</th>
<th>überhaupt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

30. Wenn Ihnen jemand die Wahrscheinlichkeit eines möglichen Ereignisses mitteilt, bevorzugen Sie diese in Worten ausgedrückt (z.B. „es passiert selten“) oder in Zahlen ausgedrückt (z.B. „das Wahrscheinlichkeit beträgt 1%“)?

<table>
<thead>
<tr>
<th>bevorzuge</th>
<th>immer</th>
<th>bevorzuge</th>
<th>immer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wörter</td>
<td></td>
<td>Zahlen</td>
<td></td>
</tr>
</tbody>
</table>
31. Wenn Sie sich die Wettervorhersage anhören, bevorzugen Sie die Prognosen in Prozent ausgedrückt (z.B. „die Regenwahrscheinlichkeit beträgt heute 20%“) oder in Worten ausgedrückt (z.B. „die Regenwahrscheinlichkeit ist heute gering“)?

<table>
<thead>
<tr>
<th>bevorzuge immer</th>
<th>bevorzuge immer</th>
<th>Prozent</th>
<th>Wörter</th>
</tr>
</thead>
</table>

32. Wie oft finden Sie in Zahlen ausgedrückte Informationen nützlich?

nie | sehr oft

33. Haben Sie bereits Erfahrung mit dieser Art von Entscheidungssituationen/Szenarien?

Ja | Nein

Falls ja, in welchem Zusammenhang haben Sie diese gemacht?

__________________________

34. Erinnern Sie sich noch, welcher Betrag Ihnen im Experiment für jede blaue Kugel gutgeschrieben wurde?

__________ Cent

35. Auf welchen Überlegungen beruhten Ihre Entscheidungen während der Studie?

__________________________

__________________________

__________________________

XXX
Höchstpersönlich unterzeichnete
Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und
ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus
anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter
Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im
Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

--------------------------------------------------------------------------------
Ort, Datum ...................................... Julia Felfeli