Metacognition in Decision Making

Annika Boldt

University College

Thesis presented for the degree Doctor of Philosophy in Experimental Psychology at the University of Oxford

Metacognition in Decision Making

Annika Boldt

University College

Thesis presented for the degree Doctor of Philosophy in Experimental Psychology at the University of Oxford

Submitted Hilary 2015

Humans effortlessly and accurately judge their subjective probability of being correct in a given decision, leading to the view that metacognition is integral to decision making. This thesis reports a series of experiments assessing people's confidence and error-detection judgements. These different types of metacognitive judgements are highly similar with regard to their methodology, but have been studied largely separately. I provide data indicating that these judgements are fundamentally linked and that they rely on shared cognitive and neural mechanisms. As a first step towards such a joint account of confidence and error detection, I present simulations from a computational model that is based on the notion these judgements are based on the same underlying processes.

I next focus on how metacognitive signals are utilised to enhance cognitive control by means of a modulation of information seeking. I report data from a study in which participants received performance feedback, testing the hypothesis that participants will focus more on feedback when they are uncertain whether they were correct in the current trial, whilst ignoring feedback when they are certain regarding their accuracy.

A final question addressed in this thesis asks which information contributes internally to the formation of metacognitive judgements, given that it remains a challenge for most models of confidence to explain the precise mechanisms by which confidence reflects accuracy, under which circumstances this correlation is reduced, and the role other influences might have, such as the inherent reliability of a source of evidence. The results reported here suggest that multiple variables – such as response time and reliability of evidence – play a role in the generation of metacognitive judgements. Inter-individual differences with regard to the utilisation of these cues to confidence are tested. Taken together, my results suggest that metacognition is crucially involved in decision making and cognitive control.

Acknowledgements

First and foremost, I would like to thank my supervisor Nick Yeung, for his unwavering support, encouragement, humour, and patience during the last years. His guidance made this project a rewarding journey, and I could not have wished for a better DPhil supervisor.

I would also like to thank all past and present members of the ACC Lab. The warm and welcoming atmosphere in the group was one of the main reasons why I enjoyed my time in Oxford so much. I will miss discussing research questions with you – discussions that often ended in the pub together with the equally-fun Summerfield Lab. Special thanks to my past-and-now-again-present colleague Franka for all the help and for answering my countless questions before I even arrived at Oxford; and to Lizzie (team squircle!), Marike, and Lucie, who spent many hours proofreading and provided me with valuable feedback.

I would furthermore like to thank my collaborators: Chris Summerfield and Vincent De Gardelle for all the inspiring discussions about our project, Stanislas Dehaene for welcoming me at Neurospin, and Robert Rogers. I am also grateful to Mihaela Duta for teaching me the hidden secrets of Psychtoolbox. I would also like to express my gratitude to Benedetto De Martino, for welcoming me to the BdM Lab, for being so patient and supportive during the past months, and for all those inspiring discussions about confidence and value-based decision making. I am looking forward to working with you over the course of the next years.

Over the past years I have furthermore received support from a great number of individuals 'outside' of academia: First, I want to express special thanks to Timo for all his help, encouragement, and humour. Thank you for reading every single page of this thesis and for making me laugh whenever I needed it most. This thesis would not have been possible without your support. Moreover, I thank my family, especially my parents for being supportive in every imaginable way and for encouraging me to be curious and critical in thinking, as long as I can remember. I would also like to thank my brother Niklas, for being the best brother in the world (but not for installing 2048 on my phone, which took days out of my writing time). Finally, I would like to thank all of my friends – both in the UK and back in Germany – for all their moral support, advice, and understanding, especially during the last months.

The research reported was supported by a studentship from the ESRC.

Publications arising from this thesis

Portions of this thesis appear in the following publications:

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience, 35*(8), 3478-3484. doi:10.1523/JNEUROSCI.0797-14.2015

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences, 18*(4), 186-193. doi:10.1016/j.tics.2014.01.006

Contents

Chapter 1

Introduction: What is metacognition?

In everyday life, we constantly make decisions. Some of these decisions are complex, high-level choices, such as deciding which of two different car models to buy. Decisions of this type usually require a considerable amount of effort and careful deliberation, and gathering of information can span days, weeks, or even months. On the other hand, there are low-level decisions, such as driving a car towards a set of traffic lights and perceiving the state it is in, classifying its signal as either green or red. Such perceptual decisions are usually formed without us noticing, that is non-consciously, automatically, and effortlessly.

Despite surface differences, it is assumed that these decisions share the fundamental characteristic that an effective way to reach a decision is to accumulate evidence (favouring one or another decision) until evidence in favour of one of the decisions reaches a predefined criterion, which marks the point at which a choice is settled upon. There are many different versions of such evidence sampling models, Figure 1 shows the above described example of deciding between two cars for such a model: Over time, which is represented on the x-axis, the participant accumulates evidence in two separate counters, one for each car model. These evidence counters 'race' towards a fixed threshold and the counter that reaches the threshold first determines the outcome of

the choice. For quite some time, both counters might develop fairly similarly, but then the participant watches a history documentary on the Mini Cooper, which leads to a larger increase in evidence in the blue counter. However, one day, the participant walks past the local car dealer who offers the Vauxhall Corsa with a 5% discount. This final, positive piece of evidence in favour of buying the Corsa leads to crossing the decision threshold: The participant has decided to buy the Vauxhall Corsa.

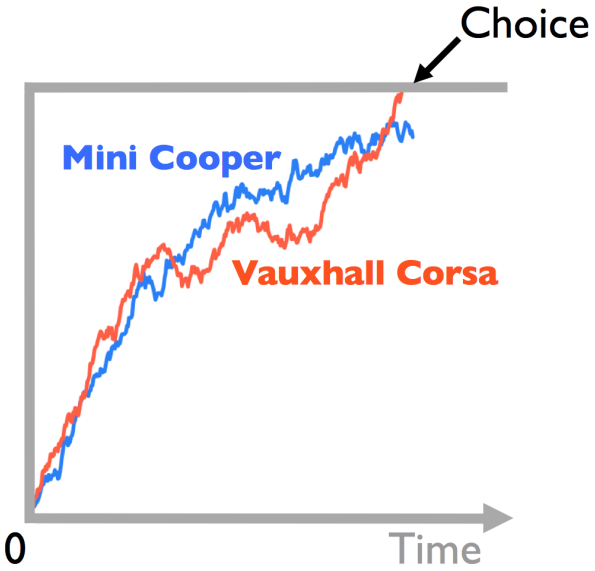

Figure 1: Schematic example of a sequential-sampling model for which two non-competing counters accumulate evidence over time, as presented on the x-axis, until one of the counters reaches the decision threshold (upper horizontal line), at which time a choice is elicited and accumulation stops. The blue counter presents evidence in favour of the Mini Cooper car model; the red counter evidence in favour of the Vauxhall Corsa model.

This example already illustrates two of the dependent variables that are usually studied in the context of decision making. One of them is the choice itself: In the case of value-based decision making, such as the car example just described, there is a preferred and a non-preferred option. In the low-level tasks often used in laboratory research on perceptual decision making, such

as the random dot-motion task (e.g., Kiani, Corthell & Shadlen, 2014), there is usually a correct and an incorrect response option. Averaging the choices made by a participant across decisions leads to error or accuracy rates. The other dependent variable is decision time, represented on the x-axis in Figure 1. Tracking both the speed and the actual choice of a decision makes sense from an economic perspective – there are situations in which we have to put a stress on the accuracy of a choice, whereas in other situations we need to be as fast as possible. There are also many situations in which we have to balance both factors, given that there is usually a tradeoff between the two, that is decisions that are slow are usually more accurate and vice versa (Heitz & Schall, 2012; Bogacz, Wagenmakers, Forstmann & Nieuwenhuis, 2010).

There is a third variable that has increasingly gained interest over the past decades in the decision-making literature, which is the sense of whether or not a decision just made is a good or a bad one. Human observers have been found to be capable of finely calibrated evaluations of their task performance. In perceptual-decision tasks, for example, participants reliably detect their errors produced under speed pressure (Rabbitt, Cumming & Vyas, 1978), and report graded judgements of confidence that correlate closely with their objective performance (Audley, 1960; Baranski & Petrusic, 1994; Vickers & Packer, 1982).

This sense of correctness or incorrectness is often defined as an instance of *metacognition*. As a prefix, the Greek *meta* denotes "something of a higher or second-order kind" (*meta-*, n.d.). *Cognition* comes from the Latin word *cognoscere* which means "get to know" (*cognition*, n.d.) and can be translated with "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses". Putting both terms together, metacognition therefore "refers to one's knowledge concerning one's own cog-

nitive processes or anything related to them" (Flavell, 1976, p. 232). According to such a broad definition, a whole range of cognitive processes must be classified as metacognitive, such as reward prediction errors (Shea, 2012). For the purpose of clarity, I will, however, use a more narrow, operationalised definition of metacognition, understanding it to be "behaviour about behaviour" (Fleming, Dolan & Frith, 2012). Such a definition includes, for example, performance judgements in which the participant is asked to rate how accurate the previously made decision was.

In the present thesis, I focus on metacognition in decision making. I ask how studies regarding such signals should best be designed, how different lines of research on metacognition in decision making can be linked, what the uses of metacognition in decision making are, as well as how metacognitive judgements are formed. Accumulation models have been useful in understanding decision making, and might be equally useful as a framework for understanding metacognition. I therefore assume sequential sampling as an underlying mechanism when asking these questions. In the following sections, I review findings on metacognition in decision making and derive my research questions from these findings.

1.1 Fields of metacognition research

Metacognition has raised interest in many different lines of research, some of which I briefly review here to give an overview on how research in this field has developed over the past decades. The idea of metacognition reaches back to the roots of experimental psychology: William James distinguished the self as knower ("I") and known ("me") and used introspection to explore mechanisms of the human mind (James, 1890). Introspection can be defined as the act

of reading out internal, mental states. Such states can be metacognitive (detection of an error), but also *object-level* representations (feeling hungry). It should therefore be stressed that while metacognition is a case of introspection, not all introspection is necessarily metacognitive. Introspective judgements – both metacognitive and non-metacognitive – were one of the main methods of early experimental psychology (James, 1890). Introspection proved to be a suitable psychological method to study metacognitive insight as reflected in confidence ratings. The first systematic experiments focusing on confidence of being correct with regard to a just-made decision were conducted in the context of psychophysics paradigms. The standard paradigm to assess confidence is highly similar to the one used nowadays: Immediately after making a perceptual decision, such as judging which of two lines is longer (Henmon, 1911) or which of two weights is heavier (Peirce & Jastrow, 1884), participants rate their confidence, often using a discrete confidence scale with verbal categories. Henmon (1911), for example, used a 4-point scale ranging from "perfectly confident", "fairly confident", and "with little confidence", to "doubtful".

Most importantly, these studies revealed that confidence judgements covaried reliably with participants' objective accuracy (Peirce & Jastrow, 1884; Fullerton & Cattell, 1892; Henmon, 1911). In other words, when participants stated that they had made an error, there was a high likelihood that they had indeed committed a mistake in the previous trial. This finding, no doubt, fuelled further interest in confidence as a measure *per se*. It was moreover found that participants vary to a large degree in how they assigned the confidence categories to their internal feelings of confidence or certainty (Fullerton & Cattell, 1892), with their introspective ability being far from perfect and with signs of systematic over- and underconfidence for some of the participants. Moreover, these early studies already established that confidence depends on

task difficulty (Griffing, 1895), that faster responses are often classified as more confident (Henmon, 1911), and that distributions of confidence responses for correct and error trials are largely overlapping (Griffing, 1895). Furthermore, Peirce and Jastrow (1884) explicitly defined confidence as a metacognitive, second-order mental process: "The quantity which we have called the degree of confidence was probably the secondary sensation of a difference between the primary sensations compared."

Taken together, work on confidence in decision making reaches back to the early beginnings of experimental psychology. The general approach of letting participants make a choice and then asking them how certain they are that their response was correct is still used by most studies focusing on perceptual decision making (Pleskac & Busemeyer, 2010; Fleming, Weil, Nagy, Dolan & Rees, 2010; Koriat, 2011; Baranski & Petrusic, 1998), as well as value-based choice (De Martino, Fleming, Garrett & Dolan, 2013). The focus of this thesis is on confidence in being correct, studied in the context of perceptual decision-making studies. Other lines of research in which metacognitive judgements have been studied will furthermore be reviewed in the present section of this chapter.

During behaviourism, introspective methods were heavily criticised due to their subjective nature and hence measuring confidence lost its appeal for most researchers. It was not until the late 1960s and early 1970s that researches became interested in it again (Dunlosky & Metcalfe, 2009), especially in the context of metacognition in memory, or *metamemory*. Both retrospective confidence judgements, for example confidence that a memory that had just been retrieved is accurate, similar to the ones used in early confidence studies, and prospective metacognitive judgements were used (for a review see Dunlosky & Metcalfe, 2009): For example, judgements of learning (JOLs; Nel-

son & Dunlosky, 1991; Vesonder & Voss, 1985) require participants to judge how likely they are to remember a to-be-studied item when tested later. Relatedly, ease-of-learning judgements (EOL; Richardson & Erlebacher, 1958) ask participants to predict which items are hard or easy to memorise, therefore measuring metacognition before acquisition of memory material. Other retrospective judgements include feeling-of-knowing judgements (FOK; Hart, 1965), made when participants failed to remember an item and are then asked to judge how confident they are that they would recognise the item if presented with it alongside other items. Related to those judgements is the *tip-of-tongue* phenomenon, which describes the participant's introspective judgement of being close to recalling an answer without actually being able to remember the item. Metamemory judgements have also been studied in the context of educational psychology, where it has been found, for example, that ease-of-learning judgements predict how much study time is allocated to each item in self-paced designs, that is more study time for items for which participants claim to be less confident (Nelson & Leonesio, 1988). Taken together, research on metamemory – both within the memory literature itself and in educational psychology – constitutes an important line of research on metacognition, focusing on how "knowledge about knowledge" is formed during acquisition, retention, retrieval, and recognition, acknowledging the crucial role metamemory plays in guiding behaviour. I return to this latter notion in Section 1.3.

Yet another line of research in which metacognitive processes have been studied intensively is the literature on error monitoring. This line of research started with the seminal research by Rabbitt and colleagues (Rabbitt, 1966, 2002; Rabbitt et al., 1978; Rabbitt & Rodgers, 1977), who investigated the conditions under which people can detect and correct errors. Participants are usually required to quickly respond to a visual stimulus and directly afterwards

have to signal whether they believe that their response was correct or incorrect. Prominent theories of error detection assume that errors are detected by means of a comparison of what would have been the correct response with the actually performed one (Falkenstein, Hohnsbein, Hoormann & Blanke, 1990; Coles, Scheffers & Holroyd, 2001). In the early 1990s, neural patterns characteristic for errors and error awareness were observed in the human electroencephalogram (EEG) for the first time (Falkenstein, Hohnsbein, Hoormann & Blanke, 1991; Gehring, Goss, Coles, Meyer & Donchin, 1993) and have further shed light on the underlying mechanisms of metacognition: Steinhauser and Yeung (2010), for example, have found that one of these EEG correlates, the error positivity (Pe), reflects the internal evidence accumulated in favour of an error having occurred. Nieuwenhuis, Ridderinkhof, Blom, Band and Kok (2001) had previously found that this component is more enhanced following detected errors compared with errors than went undetected. Taken together, apart from research focusing on psychophysics and memory, the error monitoring literature constitutes yet another line of research in which metacognitive judgements – in this case usually binary error detection – have been studied extensively. This line of research has furthermore focused on the underlying neurophysiological mechanisms of the error detection process.

In addition to the already described lines of research, researchers focusing on the underlying mechanism of consciousness and awareness also became interested in metacognitive judgements. Confidence judgements were often used as a means to decide between different theories of consciousness (Lau & Rosenthal, 2011). The above-described retrospective confidence ratings have been used in this line of research, as well as other methods, for example *post-decision wagering* (PDW). PDW requires participants to first make a primary decision and afterwards put a wager on the correctness of that decision. It is

assumed that in the case of non-advantageous wagering (placing a low bet on a correct answer), participants had no conscious representation of the correctness of the primary answer. PDW is an indirect measure to assess confidence and is often considered as a better option than confidence judgements because it overcomes certain weaknesses (Persaud, McLeod & Cowey, 2007). Others have argued, however, that this method is hardly more objective than a confidence judgement, given that wagers are placed using a subjective response criterion (Seth, Dienes, Cleeremans, Overgaard & Pessoa, 2008). In summary, researchers interested in consciousness have used metacognitive judgements as a tool to study visual awareness, using both explicit (confidence judgements) and implicit (PDW) measures of confidence. This constitutes yet another line of research on metacognition.

Similarly, researchers within the field of comparative psychology have become interested in metacognition, asking whether animals experience metacognitive states of mind similar to those humans experience (Kornell, 2009; Terrace & Son, 2009; J. D. Smith, 2009; J. D. Smith, Shields & Washburn, 2003). Studies have found behaviour supporting the hypothesis of such metacognition in rhesus monkeys (J. D. Smith, Shields, Schull & Washburn, 1997), bottlenose dolphins (J. D. Smith et al., 1995), and rats (Foote & Crystal, 2007; Kepecs & Mainen, 2012). Many of these studies used *opt-out* or *uncertainty* paradigms: The animals usually perform a simple decision task and are rewarded for correct responses. Instead of making a choice, however, they can also choose to opt out of the current trial and wait for the next one. Usually, such an opt-out, or uncertainty response, leads to a small, less desirable reward. These studies show that the animals opted out more frequently on difficult trials – trials which have a higher probability of being incorrect. These results led to the interpretation that the animals chose the less desirable reward rather than

risking an error, therefore being able to process their error likelihood, which would constitute an example of metacognitive processing. In general, this approach assumes that animal behaviour that mimics the behaviour of humans who give metacognitive reports of their actions should be interpreted as metacognitive. However, paradigms that follow this approach have been heavily criticised, mainly because the results elicited with such tasks can alternatively be explained in terms of stimulus-response (S-R) connections between difficult stimuli and opting out as an alternative response option (Carruthers, 2008; J. D. Smith, Beran, Couchman & Coutinho, 2008; Le Pelley, 2012). In other words, animals might learn to associate a certain type of stimulus (more difficult stimuli) with the uncertainty response – a strategy that maximises reward. Critically, this explanation does not rely on the assumption that they possess a subjective feeling of uncertainty similar to humans. However, this critique does not apply to all comparative studies of metacognition (Hampton, 2001; for a review see J. D. Smith, 2009). For example, Kepecs and Mainen (2012) have argued for an alternative approach to studying metacognition in animals: Rather than classifying animal behaviour as metacognitive whenever it mimics the behaviour of humans who have reported to act metacognitively, they argued that instead we should search for evidence of neural signatures that mimic the expected properties of a confidence signal. They found that activity in the rat orbitofrontal cortex (OFC) provided precisely such a confidence signal (Kepecs, Uchida, Zariwala & Mainen, 2008). Taken together, metacognition in decision making has been studied in animals with mixed results as to whether or not they are capable of metacognitive processing. In recent years, a new approach to study animal metacognition has been suggested by which neural correlates of confidence signals are studied rather than behaviour that mimics that of humans. This approach, as well as more refined

versions of the previous, behaviour-based studies could arguably circumvent these problems. Whether or not animals are capable of metacognition remains highly debated.

In conclusion, metacognition has been studied since the early days of experimental psychology. The general method of letting participants rate their confidence after making a perceptual choice is still used nowadays and will also form the main method used in this thesis. Graded confidence judgements have also been used in the memory literature. Moreover, several other metamemory judgements, such as judgements of learning and feeling-of-knowing judgements have been studied, each focusing on different stages in memory processing. Yet another context in which metacognition has been studied is error monitoring. In this line of research, participants are usually asked whether or not they think they made an error, often using a binary scale. Researchers studying error monitoring have furthermore used neurophysiological recordings to investigate the underlying neural mechanisms that give rise to such error detection processes. Moreover, another line of research originated in the consciousness literature. In this context, metacognitive judgements have been used more as a tool – rather than an object of study – to investigate visual awareness. All these different strands of research have in common the conclusion that people are capable of judging their own performance, often with astonishing accuracy. This finding raises the question as to whether animals possess such insights as well. Two different approaches have been used to study this question. The first assumes that if an animal behaves as if it monitors its performance metacognitively, we should assume that it possesses such metacognitive insights. More recently, another approach has been used to study animal metacognition, searching instead for neural correlates of uncertainty or confidence. The question as to whether animals engage in metacognitive processing is still highly debated,

though. Taken together, metacognition has been studied in many different contexts. In this thesis, I focus on metacognition in perceptual decision making, whilst capitalising on what has been discovered in the research on error monitoring. More on how these two lines of research can be linked will be outlined in the following section.

1.2 Linking research on decision confidence and error detection

The previous section highlighted the fact that metacognition has been studied in many different lines of research. Here, I argue that two of these lines – focusing on confidence in perceptual decision making and error monitoring – should be linked. These lines of research use highly similar methods: In both cases, participants first make a perceptual decision and are asked for a metacognitive rating after every choice they make. In case of confidence judgements, participants are usually required to judge the likelihood that their previous choice was correct, using a graded scale. In case of error detection, however, participants usually have to judge whether or not they responded incorrectly in their previous choice, using a binary scale. Despite these striking similarities, error monitoring and confidence judgements have rarely been linked (Yeung & Summerfield, 2012, 2014; but see also Fernandez-Duque, Baird & Posner, 2000).

Despite the many similarities the two lines share in their methodological approaches, there is little compatibility between theories proposed in the respective research literatures. For example, theories of error monitoring can explain why people sometimes say with high certainty that a just-given response was incorrect. These types of trials cannot be explained by most models

of decision confidence, which always calculate confidence relative to the choice made, therefore not being able to account for such changes of mind (Vickers & Packer, 1982; Kiani et al., 2014). On the other hand, error detection is usually looked upon as being all-or-none (Charles, Van Opstal, Marti & Dehaene, 2013), while confidence is treated as a continuous variable (Fetsch, Kiani, Newsome & Shadlen, 2014). Taken together, the two lines of research have rarely been linked theoretically, even though such a link could be worthwhile given their different strengths and weaknesses.

Moreover, these lines of research are similarly discrepant with regard to the neural systems that have been studied. Error monitoring research has mainly studied the involvement of the medial prefrontal cortex (PFC), commonly focussing on the anterior cingulate cortex (ACC). It has often been assumed that the medial PFC monitors for errors and cognitive conflict, passing this information on to lateral PFC, which is then in turn able to execute cognitive control (Yeung, 2013; Egner & Hirsch, 2005). In this context, two components of error processing have been studied using EEG and magnetoencephalography (MEG) recordings (Falkenstein et al., 1991; Gehring et al., 1993; Steinhauser & Yeung, 2010; Nieuwenhuis et al., 2001; Van Veen & Carter, 2002; Botvinick, Cohen & Carter, 2004) – the error-related negativity (ERN) and the Pe, the latter of which has already been mentioned above.

Research on confidence, on the other hand, has focused on more anterior and lateral areas of the PFC (Chua, Pergolizzi & Weintraub, 2014; Fleming & Dolan, 2014). Yokoyama et al. (2010), for example, compared a confidence rating task to a brightness discrimination task that was constructed to be as similar as possible to the confidence rating condition. Using neuroimaging techniques, the authors identified the right frontopolar cortex in the PFC (area BA10) as crucially involved in confidence judgements. Re-

latedly, in a study by De Martino et al. (2013), participants decided which of two snack items they preferred, as well as rating their confidence in this choice. Ventromedial PFC tracked both the difference in subjective value between the snack items, as well as confidence. The results furthermore suggested that activity in the right rostrolateral PFC was also modulated by confidence, but not by value. This result was interpreted by the authors as confidence being 'read out' by the right rostrolateral PFC. This interpretation also matches findings from a previous study by Fleming, Huijgen and Dolan (2012), who suggested that the right rostrolateral PFC receives input from other brain regions to signal uncertainty associated with the decision. These uncertainty signals, they argue, can then be integrated in the right rostrolateral PFC to form metacognitive reports. The role of the right rostrolateral PFC was also highlighted in a structural neuroimaging study by Fleming et al. (2010), who found that grey matter volume in BA10 correlated positively with how well participants' confidence ratings reflected their performance in a psychophysical task. Furthermore, Rounis, Maniscalco, Rothwell, Passingham and Lau (2010) conducted a study in which activity in the dorsolateral PFC was depressed using transcranial magnetic stimulation (TMS), which impaired metacognitive performance whilst not affecting first-order performance.

Taken together, despite using similar methodological approaches, there is little compatibility between current theories of confidence and error detection. Empirical findings with regard to the neural bases of these judgements are similarly discrepant: Studies focusing on the neural underpinnings of error detection judgements have usually investigated the role of the medial PFC, especially the ACC. Confidence judgements, on the other hand, have been linked to the more anterior and lateral areas of the PFC. Here, I test whether those different forms of judgements constitute different forms of the same underly-

ing metacognitive process. It is beyond the scope of this thesis to resolve the question of the precise anatomical networks involved in the formation of these different metacognitive judgements. Instead, I aim to compare the different neurophysiological patterns associated with these types of judgements using multivariate pattern classification techniques, testing the hypothesis that error detection and confidence give rise to highly similar patterns of activity in the EEG. A first research question, which I address with this thesis therefore becomes

> **Conceptual Question 1**: Are error detection and confidence judgements two sides of the same coin?

1.3 Interaction of metacognition and behaviour

In the previous section, I have already mentioned that research on error detection has often suggested that information about the likelihood of an error is communicated by the medial PFC to lateral PFC, which is then in turn able to execute cognitive control (Yeung, 2013; Egner & Hirsch, 2005). This hypothesis suggests that metacognition might play a crucial role in cognitive control, as has previously been suggested (Yeung & Summerfield, 2012, 2014; Fernandez-Duque et al., 2000). Consistent with this notion, Shea et al. (2014) have recently defined metacognition as "control processes that make use of one or more metacognitive representations" (p. 187). This suggests that one of the key reasons metacognition is worthwhile studying is the fact that such self-directed evaluations not only reflect information regarding the previous decisions but can also influence future decisions. Indeed, Koriat (2011) proposed a list of key questions on metacognition, including the question "what are these [measurable effects on cognition and behaviour] effects, and how do

they influence actual performance" (p. 117).

Little work has been done on this question in the context of perceptual decision confidence, with most studies being conducted on the role of confidence in group decision making. Bahrami et al. (2010), for example, tested the hypothesis that when participants make joint perceptual decisions they will discuss how confident they each think they are with regard to their own opinion, which in turn makes the joint judgement more accurate. This strategy allows them to weight the influence of each member of the group by the level of confidence expressed, discounting those who report low confidence. The role of confidence has furthermore been highlighted in social interactions in which one member of the group or dyad has an advantage in knowledge, so-called Judge-Advisor Systems (JAS). In a study by Sniezek and Van Swol (2001), advisors made recommendations to judges while also expressing their confidence in their statements. Judges trusted this advice more if made with high confidence. This finding makes sense if we consider that highly confident choices are also more likely to be accurate. Moreover, shared confidence also plays an important role in the context of eyewitness data. Research has shown that juries often determine whether or not witness testimony ought to be trusted by taking the witness's confidence into account, judging more confident witnesses as more credible (Wells, Lindsay & Ferguson, 1979; Wells, Ferguson & Lindsay, 1981; Tenney, MacCoun, Spellman & Hastie, 2007). Taken together, all of these findings suggest that people are sensitive to the level of confidence expressed by others and that they tend to weight the opinions of others by this evidence, thereby discounting the influence of others who express to be uncertain.

In a related field, error monitoring, the use of this metacognitive information has also been a critical point. One instance of metacognitive control

that has been studied extensively is *post-error slowing* (PES): It has often been observed that people tend to be slower after making an error (Dutilh et al., 2012; Rabbitt & Rodgers, 1977; Rabbitt, 1966; Laming, 1979). Different interpretations for this effect have been proposed (e.g., Jentzsch & Dudschig, 2009; Danielmeier & Ullsperger, 2011). Notebaert et al. (2009), for instance, have suggested that participants are slower after committing an error because errors are infrequent and therefore trigger an orienting response. Dutilh, Forstmann, Vandekerckhove and Wagenmakers (2013), on the other hand, explained post-error slowing from the perspective of an evidence accumulation model, suggesting that participants increase their decision threshold after responding incorrectly, therefore requiring more absolute accumulated information for triggering a response, which in turn leads to increased response, or reaction times (RTs). Post-error slowing has been observed even when participants were not aware of their errors: Logan and Crump (2010), for example, conducted a study in which participants had to type words presented on screen. On a proportion of trials, visual feedback was manipulated so that the word they typed presented on screen was either corrected if they had made an error, or errors were introduced to make participants believe that they had committed a mistake. Critically, even though participants accepted those introduced errors as their own, post-error slowing was not found after these trials. Instead, people slowed down after 'real' errors, as well as errors that had been corrected by the computer. The authors explain these effects in terms of two hierarchical control loops with different feedback mechanisms: The internal control loop detects errors and produces post-error slowing, while the outer control loop produces explicit reports of errors. Inserted errors are never classified as mistakes in the inner loop, therefore they do not trigger post-error slowing.

Furthermore, also relevant in this context are studies on the impact

of object-level uncertainty on cognitive processing. Object-level uncertainty refers to an imprecision that can occur at any level of cognitive processing (Bach & Dolan, 2012), in other words noise inherent in the cognitive processes or representations rather than evaluative, re-interpreted *meta-level* uncertainty. For example, risk or expected uncertainty describe any imprecision in the representation of a probabilistic outcome. Such outcome uncertainty has been found to affect participants' likelihood to explore rather than exploit known choice alternatives: In recent studies by Frank, Doll, Oas-Terpstra and Moreno (2009) and Badre, Doll, Long and Frank (2012), participants have been found to track the uncertainty of an option and to then use this information to guide exploration. More precisely, they decide to explore the environment whenever a large amount of information could be gained by exploring that particular option. Uncertainty in an environment also influences learning, as shown by Behrens, Woolrich, Walton and Rushworth (2007). The authors let participants complete a one-armed bandit task in which outcome probabilities of two choice options had to be tracked over time. Participants were found to take into account uncertainty in the environment when making their choices, discounting new evidence that was produced during volatile periods as opposed to stable periods, as could be expected from a Bayesian optimal observer. Taken together, object-level uncertainty has been shown to affect future behaviour in ways that might be relevant for the study of metacognitive uncertainty and guide future research on this topic.

In metamemory, too, the use of metacognitive judgements seems to be critical. For example, Nelson and Leonesio (1988) conducted a study in which participants gave both ease-of-learning (EOL) and feeling-of-knowing judgements in a self-paced learning task. The lower their judgements were – particularly their ease-of-learning judgements – the more study time was

allocated to respective items, consistent with the "monitoring-affects control hypothesis". If metacognitive insight is beneficial in educational contexts then intervention methods should aim at training students to develop such skills. Indeed, such attempts have been made in the educational literature. One such method is the *joker-word* method applied in spelling and writing training (Spitta, 2011), where pupils are encouraged to choose and highlight one word in a dictation test when they are uncertain how it is spelled. If this word is indeed misspelled, the teacher will not count the mistake towards the final grade, which presumably trains students to monitor their own likelihood of misspelling words. Yet another example in which students' metacognitive insight is encouraged can be found in the context of choir singing. It is common practice that singers upon noticing that they sang a wrong note, raise their hands to signal to the director that they are aware of their error (De Quadros, 2012). The choir director can then use this error-detection signal, either choosing to continue the rehearsal without interruption – especially when the error was signalled by a particularly strong singer who can be trusted to correct his or her mistake without help – or intervene and repeat the part in question. Together, these examples suggest that metacognitive signals play an important role in learning by guiding students' attention to areas the have not yet fully grasped and by improving student-teacher interactions.

Taken together, in the present section, I reviewed findings on how confidence judgements are utilised. Most of these findings focus on the role of confidence in joint decision making or learning, while this question has been largely understudied in the context of cognitive psychology – with the exception of how (binary) error detection leads to post-error slowing. In this thesis, I therefore ask how graded confidence judgements are utilised to enhance cognitive control, more precisely by asking if they modulate information seeking.

The general notion would be that metacognitive signals can serve as an internal proxy of feedback to the participant, especially when external feedback is not present or unreliable. Thus, whenever participants are uncertain whether or not they committed an error in the previous response, and would presumably state that they were merely guessing if asked to rate their confidence, external feedback should be the most valuable and participants should be expected to pay close attention to it.

To my knowledge, this question has not been studied in a decision-making context, but interesting findings exist in the literature on educational psychology. Kulhavy and Stock (1989), for example, found that participants spent more time studying feedback when there was a mismatch between their perceived accuracy (confidence) and their objective accuracy, as conveyed by feedback. This findings provides support for their *certitude model*, according to which time spend studying *elaborated feedback* in a learning task depends on confidence or rather a matching process between metacognitive insight and external feedback. The authors did not study the case of *sure errors*, though, as these were unlikely to occur in their design. I therefore designed a task in which participants would use the entire confidence scale reaching from *sure errors* to *sure corrects*. The second research question addressed in this thesis therefore becomes

> **Conceptual Question 2**: How does confidence affect attention to feedback?

1.4 How metacognitive judgements are generated

In the previous sections, I have discussed both how metacognitive judgements are used to exert cognitive control, and how different types of judgements arise

from the same metacognitive processes – without making strong assumptions as to how confidence judgements are formed. How metacognitive judgements are generated will therefore be the focus of the present section.

Within research on metamemory, the question of whether metacognition relies on *direct access* to the memory trace or cues and heuristics has generated significant debate (Schwartz, 1994; Koriat, 2012). For instance, different theories have been proposed as to how feeling-of-knowing judgements are formed: Hart (1965) suggested that they arise from an internal monitor, which can tell whether or not a memory can be retrieved, similar to an inventory list. This theoretic account of feeling-of-knowing judgements constitutes a direct-access model because people base their judgement directly on the presence (or strength) of the memory trace. Most other theories of this phenomenon, however, instead assume a *heuristics-*, *cue-based*, or *inferential* approach. Koriat (1993), for example, suggested that feeling-of-knowing judgements are based on how much information is accessible during the retrieval process. This assumption differs from Hart's model in the sense that it does not rely on an internal monitor with privileged access to the memory trace itself, but is instead based on information that arises as a by-product from the retrieval process itself. Accessibility does not necessarily refer to correctly retrieved information. In fact, Koriat (1993) found that when participants studied strings of random letters, feeling-of-knowing judgements were based on how many letters participants recalled, regardless of whether they were correctly or incorrectly recalled.

Another heuristics-based model was proposed by Reder and Ritter (1992). The authors found that participants's feeling-of-knowing judgements were affected by the familiarity of a question: Participants had to solve simple arithmetic tasks (e.g., $18 * 37 =?$), choosing whether they would retrieve the answer from memory or calculate it. Retrieval of the answer yielded much

higher rewards but much more time was granted for calculation. Participants' judgements as to whether or not they would have to calculate the answer (i.e., their feeling-of-knowing judgement) predicted retrieval performance, suggesting that they had good metacognitive insight. As a critical manipulation, the authors included tasks in the test set that had not been studied previously but were very similar (e.g., $18 + 37 =?$). The more frequently the original operand pair that formed the new task had been shown previously, the more likely participants were to choose to retrieve their answer rather than calculate it. It can therefore be argued that familiarity influences the strength of participants' feeling-of-knowing judgements.

In conclusion, different theories to explain the formation of metamemory judgements have been proposed. Some of these theories assume direct access to memory contents, whereas most theories have proposed a heuristics-based view according to which metamemory judgements are influenced by cues such as familiarity or accessibility of the studied material. Many such theories suggested that multiple cues and situational circumstances contribute to metacognition. For example, the time point at which the confidence judgements were triggered (e.g., during a study or test phase in a memory task) can change which cues contribute towards the final confidence judgement in a specific task (Schwartz, 1994).

In the decision-making literature, there has been less debate regarding those two classes of models. Most models assume a version of direct access, suggesting that confidence judgements rely on the same information as the decision itself. The *balance-of-evidence* model (Vickers & Packer, 1982; Van Zandt & Maldonado-Molina, 2004; Kiani et al., 2014; De Martino et al., 2013), for example, assumes that confidence is a function of the final states of the evidence counters at the time of the decision. Two examples are shown in

Figure 2, continuing the previous example of choosing a car model. If the losing counter had accumulated almost as much evidence as the winning one, then the decision would be made with low confidence (upper panel of Figure 2). If however the winning counter had accumulated significantly more information at the time of the decision, then the decision is made with high confidence, as presented in the lower panel of Figure 2.

While most research has assumed that decision confidence relies on direct access to the internally accumulated information of the primary decision, there has still been a lively debate about the time point at which this information is read out (Gherman & Philiastides, 2014; Baranski & Petrusic, 1998; Yeung & Summerfield, 2012, 2014). The original balance-of-evidence model, for instance, assumes that confidence is based on the difference between the two counters at the time of the decision (Vickers & Packer, 1982). These models therefore assume a *decisional locus* of confidence. In contrast, a balance-of-evidence model proposed by Van Zandt and Maldonado-Molina (2004) assumes that evidence continues to accumulate after the decision threshold has been reached, up until the time at which the confidence judgement is made. This model therefore constitutes an example of a *post-decisional locus* model. These two classes of models will furthermore be reviewed in the next section of this chapter.

However, there have been some competing views to the assumption that decision confidence is based on direct access. One prominent example of a heuristics-based model that has been studied in detail is the *time heuristic*. According to this model, faster responses should be judged as more confident (Audley, 1960; Moreno-Bote, 2010; Zylberberg, Barttfeld & Sigman, 2012). Hanks, Mazurek, Kiani, Hopp and Shadlen (2011) have recently suggested that the elapsed time might serve as a useful proxy for how reliable the ac-

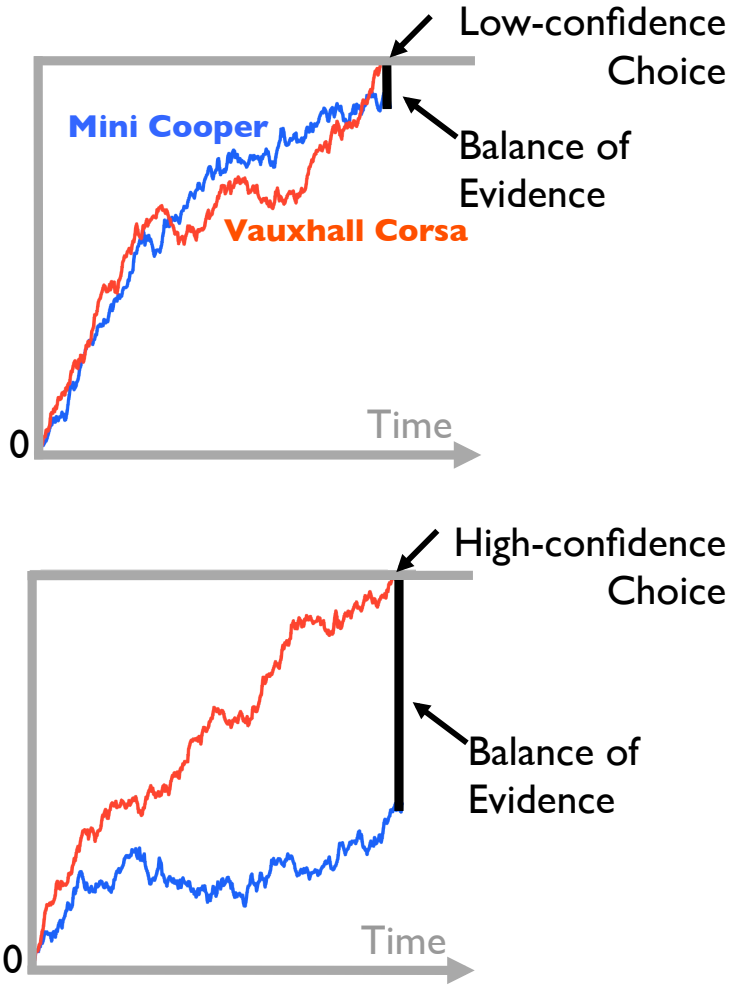

Figure 2: Two example trials in which confidence is based on a balance-of-evidence mechanism, that is the difference in accumulated evidence between the two counters. The upper panel depicts a low-confidence trial in which the participant chose the Vauxhall Corsa (red counter), but the counter for the Mini Cooper (blue counter) had accumulated almost as much evidence, leading to a low balance of evidence. The lower panel shows an example in which there was significantly less evidence in favour of the Mini Cooper, leading to a large balance of evidence and therefore high confidence that the decision was the correct one.

cumulated evidence is. The time heuristic exploits the regularity of slower errors compared to faster correct responses – a pattern that is found in many situations and tasks. This is due to the fact that such errors are caused by a low rate in the accumulation of evidence (unreliable evidence), meaning that the random fluctuations in the accumulation process might govern the decision, leading to a substantial number of errors. In such situations, the time heuristic would lead to well-calibrated confidence judgements. However, many tasks – especially when speed is stressed over accuracy – exhibit faster errors compared to correct responses (Scheffers & Coles, 2000). The time heuristic does not provide a valid cue for metacognition in those situations. In addition to RT, I consider whether reliability of evidence also contributes towards how confident participants judge their decisions, as has been previously suggested by Yeung and Summerfield (2014, 2012), as well as Irwin, Smith and Mayfield (1956) and Zylberberg, Roelfsema and Sigman (2014). This makes sense if we consider decision making from a Bayesian perspective, according to which the decision maker should be able to take into account the reliability of a source before integrating information collected from that source.

Evidence reliability as a cue to metacognition has been considered previously, especially by researchers who assumed a drift diffusion model (DDM) framework, as shown in Figure 3. In this class of sequential-sampling models, participants are assumed to accumulate evidence in one single counter that tracks the difference in evidence for two choice options (Ratcliff, 1978; P. L. Smith & Ratcliff, 2004) – not in two separate counters as in all previously discussed examples. Evidence accumulation stops when the amount of evidence exceeds one of the two absorbing boundaries that represent the choice options, arranged above and below a starting point. Evidence reliability or *information quality* in such a model would be represented by the rate at

which evidence is accumulated. A highly reliable source of evidence – such as a road sign perceived in clear weather conditions – is associated with a high mean accumulation rate, or *drift rate*, whereas reading the same road sign in a heavy snow storm results in a lower drift rate and the evidence accumulation process being more governed by noise. *Information quantity*, on the other hand, is the total amount of evidence calculated by the evidence counter. A model which assumes that decision confidence is based on this information quantity would per definition predict that confidence is precisely the same on every trial, because on each trial evidence is accumulated to the same fixed threshold (Yeung & Summerfield, 2012). Instead, combinations of information quantity and information quality have been suggested to form the basis of confidence. Pleskac and Busemeyer (2010), who refer to this combination as *Peirce's model* (Peirce & Jastrow, 1884), proposed a computational model in which decision confidence relies on this combination of parameters. Their two-stage dynamic signal detection (2DSD) model extends the standard diffusion model framework by a post-decision phase during which participants continue to accumulate evidence until their confidence judgement. The model provided a good fit to empirical first-order decisions (both accuracy and RTs), as well as confidence judgements for different tasks. Taken together, there is a great deal of research to suggest that evidence reliability affects confidence judgements. This parameter will therefore be considered in this thesis as one of the cues that determine confidence judgements, using a colour-judgement paradigm by De Gardelle and Summerfield (2011) that allows me to directly manipulate evidence reliability orthogonally to other proposed confidence cues[1].

In conclusion, whereas there has been a lively debate in the meta-memory literature as to whether confidence is based on direct access or heur-

[1]Here, I will understand cues as internal sources on which metacognitive judgements are based, which can be both direct-access or heuristics-based.

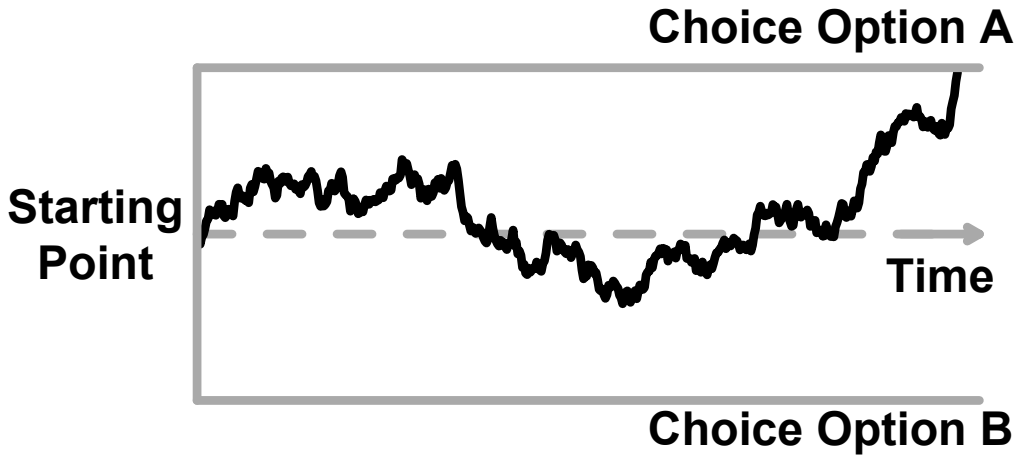

Figure 3: Schematic, simplified example of a trial in a drift-diffusion model (DDM). Noisy evidence is accumulated over time, shown on the x-axis. Two absorbing boundaries lie equidistant to an unbiased starting point, representing the two options. The evidence counter therefore tracks the difference in evidence for the two choice options. The decision is terminated when the counter reaches one of the boundaries, in this case the upper boundary, therefore choosing option A.

istics, this question has not received much discussion in the context of decision confidence. Most commonly, a version of a direct-access model is implicitly assumed, with the exception of the time heuristic – a heuristics-based account of decision confidence. I am therefore going to propose and test different cues – both direct-access and heuristics-based – on which decision confidence could be based internally, for example evidence reliability, which I aim to manipulate independently of evidence strength in my experiments. In the context of this question, I furthermore ask whether individual differences exist in the use of such different confidence cues, for instance in the context of clinical populations: Patients diagnosed with depression have previously been observed to be better at ignoring irrelevant information (Ahveninen et al., 2002; Schmitt et al., 2000). Highly unreliable stimuli contain much of such irrelevant information that should be ignored to efficiently and effectively process the stimulus. Such findings raise the question whether patients with depression also show differences in their use of evidence reliability as a cue to decision confidence. I test this question in this thesis using reductions in the levels of serotonin – a neurotransmitter that has been shown to play a key role in clinical depression – to model depression in healthy participants. Taken together, the third research question of this thesis therefore becomes

> **Conceptual Question 3**: Which cues contribute to the formation of metacognitive judgements and are there inter-individual differences in the use of such cues?

1.5 Potential difficulties arise when measuring metacognition

This section will discuss the fourth and final research question of this thesis. While the previous three questions have focused mainly on the theoretical background of metacognition in decision making, the focus is now shifted towards more methodological issues that arise when decision confidence is measured. To address the research questions previously outlined in this chapter, the development of a suitable confidence paradigm is necessary. This question is by no means trivial. For instance, as mentioned above, metacognition in decision making has been studied in different lines of research, which have used somewhat different ways of measuring metacognitive judgements. While in the context of error monitoring studies, error detection has commonly been assessed using binary scales (*error* versus *correct*), decision confidence is usually measured on graded scales ranging from *guessing* to *high confidence*. For example, the scale used by Wierzchon, Paulewicz, Asanowicz, Timmermans and Cleeremans (2014) included four categories: "guessing", "not confident", "quite confident", and "very confident". The question therefore arises as to which scale is optimal in the sense that it provides the researcher with fine-grained confidence ratings while not imposing unnecessary processing load on the participant.

While – to my knowledge – no explicit guidelines exist as to how to measure decision confidence, there have been several attempts at formulating such guidelines elsewhere. For instance, in the context of consciousness research, Overgaard and Sandberg (2014) have discussed and contrasted different scales (see also Seth et al., 2008; Reingold & Merikle, 1988; Sandberg,

Timmermans, Overgaard & Cleeremans, 2010), often comparing them to other measures of awareness. Using an artificial grammar learning paradigm, Tunney and colleagues (Tunney, 2012; Tunney & Shanks, 2003) suggested that only binary scales were sensitive enough to detect differences in participants' awareness of grammatical rules. On the other hand, evidence exists that suggests that participants are capable of making such finer-grained judgements – critically without additional processing costs. For instance, findings from a study by Dienes (2008) – also using an artificial grammar learning task – suggest that participants were just as accurate when using binary as when using continuous scales.

Another facet of the question as to how metacognitive judgements should best be collected concerns the precise timing of metacognitive judgements. In the previous section, I have already introduced two competing models – the decisional-locus model and the post-decisional locus model (Yeung & Summerfield, 2014, 2012). While the former assumes that decision confidence is based evidence at the time of the primary decision, the latter one suggests that evidence accumulation continues until the confidence judgement is made, including therefore also evidence that did not contribute to the primary decision. For example, the 2DSD model by Pleskac and Busemeyer (2010) constitutes a post-decisional locus model: First- and second-order processes are simulated by the same evidence accumulation process, but within two consecutive processing stages: a first decision stage, and a second post-decision stage, which gives rise to the metacognitive judgement. Many influential models of error monitoring also assume that the intended response is compared to the actually performed one and an error is signalled when there is a mismatch between the two (Falkenstein et al., 1991; Gehring et al., 1993; Coles et al., 2001). These models are capable of simulating situations in which

the participant changed their mind, which is one of the main advantage of post-decisional locus models. Such changes of mind are especially common in decision-making paradigms in which speed pressure is imposed and where participants become aware of a large proportion of their mistakes. Such mistakes have been labelled *premature responses* by Scheffers and Coles (2000), who contrasted them with *errors due to data limitation*, that is errors that would have happened even if participants had been given unlimited time to respond due to the difficulty of the task.

The decisional-locus model and the post-decisional locus model are by no means mutually exclusive. In fact, Baranski and Petrusic (1998) suggested that participants form their confidence judgements at the time of the decision if they were granted enough time to make this decision. However, if primary-task responses are speeded, the locus of metacognitive judgements is shifted more towards post-decision processing modes. The latter finding is particularly crucial for the measurement of decision confidence: To permit planned contrasts of various dependent variables on correct versus error trials, participants need to commit a sufficient numbers of errors. This can be achieved by imposing speed stress, as it was done for the experiments reported in the present thesis. Baranski and Petrusic (1998) would therefore predict that decision confidence will mainly be formed at a post-decisional locus here. This prediction affects the optimal timing of confident measurements: Increasing the time window between primary response and confidence judgement should give participants a higher chance of correcting initial mistakes. This follows from the fact that evaluating more as opposed to fewer evidence samples reduces the noise in the averages on which both choice and confidence are based. To my knowledge, nobody has ever addressed the question as to where the optimal time window for collecting confidence judgements lies, but attempts to specify windows of

high metacognitive insight have been made in the metamemory literature. Judgements of learning, for instance, have been found to more accurately predict memory performance when they are made a few seconds or minutes after having learned the pair association compared to judgements made immediately after having learned the association (the "delayed-JOL effect"; Nelson & Dunlosky, 1991). I therefore address the question of the precise timing of confidence judgements in this thesis.

In addition to the two discussed methodological issues – regarding the metacognitive scale and the precise timing of the measurement of confidence ratings – there is a third methodological question that I address: whether metacognitive monitoring is distinct from first-order decision processes. In other words, metacognition could depend on the same internal processes as the primary decision or could resort to entirely different processes. Some neuroimaging evidence exists in support of the claim that they rely on the same processes. For example, Kiani and Shadlen (2009) concluded from an opt-out task with monkeys that decision confidence together with the choice itself are represented by neurons in the lateral intraparietal cortex (LIP). This notion was furthermore supported by Fetsch et al. (2014), who reported that microstimulation to motion-sensitive neurons in visual cortex changed both motion perception and opt-out behaviour regarding these perceptions in a consistent way, suggesting that both responses arise from the same underlying processes. Moreover, Gherman and Philiastides (2014) reported evidence from an opt-out study in support of the hypothesis that decision confidence relies on the same internal processes as the decision itself: EEG activity related to decision confidence arose around the time at which the decision emerged, and was related to the same neural sources. Similarly, De Martino et al. (2013) – as previously discussed – have provided evidence that the ventromedial PFC

represents both value and confidence. However, at the same time, the authors also found confidence to be represented in right rostrolateral PFC, interpreting this finding as a metacognitive 'read-out' of first-order uncertainty with regard to the calculation of value. Such a 'read-out' would mean that even though decisions and uncertainty related to these decisions might be represented by the same neural processes, what we measure with confidence judgements in humans is a re-interpreted version of this uncertainty, also taking other sources of uncertainties and confidence cues such as familiarity or fluency into account (see previous Section 1.4).

Assuming such a 'read-out' would also explain why many studies have found evidence for dissociable mechanisms: Rounis et al. (2010), for instance as previously mentioned, depressed dorsolateral PFC activity using TMS, leading to impairments in metacognitive performance whilst not affecting first-order performance. Rahnev, Maniscalco, Luber, Lau and Lisanby (2012) also used TMS to depress activity in the visual cortex, leading to impaired first-order performance, whilst counterintuitively resulting in overall increases in confidence. Moreover, Overgaard, Koivisto, Sørensen, Vangkilde and Revonsuo (2006) suggested that whether or not participants had to judge a visual stimulus with regard to its identity (first-order judgement) or visual awareness (metacognitive judgement) was reflected in several event-related potentials (ERPs) – a finding which led the authors to conclude that first-order representations are qualitatively different from metacognitive processes (also discussed in Overgaard & Sandberg, 2014, 2012). Chua et al. (2014) furthermore recently discussed neuropsychological studies which found a dissociation between metamemory judgements and primary memory processes. Taken together, there is ongoing debate as to whether metacognitive monitoring is distinct from first-order decision processes, including 'hybrid models' that assume that confidence is

based at least to some extent on 'read-out' first-order uncertainty.

Whether metacognitive monitoring arises from the decision process itself or not can be expected to have critical effects on the measurement of confidence: If making a choice and judging one's confidence are fundamentally different processes, then we should expect to find that rating confidence impairs task performance, similar to switch costs in the domain of task-switching studies (Monsell, 2003). In this thesis, I therefore test whether such switch-costs can be found in confidence designs and qualify their severity.

In conclusion, for the precise study of metacognitive processes, design of an optimal decision paradigm for measuring confidence is necessary. To my knowledge, no carefully constructed methodological guidelines as to how to study decision confidence have been proposed. I therefore report findings from several experiments in this thesis – focusing on confidence scales, timing specifics, and potential task-switching effects – that could form a first step towards such guidelines. Such explicit guidelines could potentially advise future researchers on how to design studies so that they minimise possible confounds and create optimal conditions in which the participants are able to demonstrate maximal metacognitive accuracy given a certain level of performance. The fourth research question studied in this thesis is therefore

> **Methodological Research Question**: How should paradigms to study metacognition be designed to guarantee the highest level of experimental control?

1.6 Thesis outline

Reflecting on one's own thoughts is an essential ability. Without this ability, learning and adapting to an ever-changing environment would be difficult. In

this thesis, I investigate metacognition in decision making, asking three conceptual research questions together with a more methodologically motivated question, which have been outlined in the previous sections. Given that the methodological research question is an important precursor to all other questions, I first focus on this topic before addressing more conceptual issues.

In Chapter 2, I first focus on my **Methodological Research Question**, asking how best to design an experiment to study such metacognitive judgements. Experiment 1 tests the suitability of a perceptual decision-making task and compares different types of confidence scales. Experiment 2 then focuses on the precise timing of the assessment of confidence, that is how quickly after a response participants should be asked to rate their confidence. Experiment 3 tests whether asking participants to evaluate their performance leads to impaired first-order performance, similar to the idea of switch costs caused by alternating between the perceptual decision and the confidence judgement. As well as providing preliminary evidence on some key theoretical questions – such as the relationship between confidence judgments and error monitoring – the results from this chapter serve to provide methodological guidelines for the remaining chapters.

Chapter 3 addresses **Conceptual Question 1**: How can two lines of research – focused on graded confidence judgements and binary error detection respectively – be linked? This question is particularly interesting given that they have been studied largely separately. Experiment 4 focuses on the question of whether well-characterised error-related EEG activity also varies with confidence, which would speak in favour of the hypothesis that error detection (as studied with these ERPs) and confidence are really two sides of the same coin. To foreshadow the results, such an overlap in the underlying neural mechanisms was indeed found and in this thesis, I therefore use the

terms metacognition, confidence, error detection, error awareness, subjectively judged accuracy, or *type-II* judgements interchangeably[2]. The second part of Chapter 3 then introduces a computational model which is used to explore the hypothesis that confidence and error detection arise from the same internal metacognitive processes. This model can be regarded as an exemplification of what my previously discussed results have suggested regarding the formation of metacognition – a question, which will furthermore be studied in Chapter 5.

In Chapter 4, I then focus on **Conceptual Question 2**: How confidence judgements are used internally once they are formed. More specifically, I investigate how confidence and certainty affect attention to feedback. In Experiment 5, I test the hypothesis that participants will pay closest attention to feedback when they are uncertain about the correctness of their decisions, that is when they can only guess whether their given response was correct or incorrect. A challenge of this research is to assess confidence in a way that does not disrupt task performance and assessment of feedback. I address this issue using EEG methods to measure confidence without requiring explicit (and potentially disruptive) confidence judgments on each trial.

The final empirical chapter is Chapter 5, where I address **Conceptual Question 3**: Which cues contribute to the formation of metacognitive judgements. As previously discussed, I expect evidence reliability to be a prominent cue to how confident participants judge their responses, but I also test the influence of other cues, such as RT and evidence strength, suggesting that multiple cues and signals contribute to metacognitive ratings that are expressed by participants in these types of tasks. Experiment 6 tests this general hypothesis using a different perceptual decision-making paradigm compared to the previ-

[2]I define *low confidence* as a *sure error*. I furthermore refer to certainty as a dimension enclosed within the confidence or error detection dimension, with *low certainty* referring to states of guessing, that is the midpoint of the confidence scale, and *high certainty* referring to either *sure corrects* or *sure errors*.

ous experiments, one that allows independent manipulation of both evidence strength and evidence reliability of the stimulus. I therefore used a paradigm similar to the one developed by De Gardelle and Summerfield (2011), in which participants had to judge whether an array of coloured shapes was on average more red or more blue. Evidence reliability in this task is manipulated by changing the variance in the coloured shapes of which the stimuli are composed. Experiment 7 then tests this hypothesis in a clinical context, using acute tryptophan depletion (ATD) as a means to lower serotonin levels – and therefore induce depression-like symptoms – in healthy participants. I test the prediction that tryptophan-depleted participants will show differences in the influence of evidence reliability on their confidence judgements. Finally, Experiment 8 focuses on the neurophysiological underpinnings of the different influences of evidence reliability and evidence strength on confidence.

Chapter 6 presents a general discussion of the experiments of this thesis. In this chapter, I draw together conclusions and implications from the previous empirical chapters, discuss limitations and propose future avenues to explore in the context of metacognition in decision making.

1.7 Measurement of metacognitive judgements

For the purpose of clarity, when referring to confidence judgements in this thesis, only second-order confidence is meant as opposed to first-order confidence. Such first-order confidence is often used in memory studies, meaning that confidence is assessed together with the first-order decision (e.g.; Ratcliff, McKoon & Tindall, 1994). This means instead of first making a binary decision and then a confidence judgement, participants make both ratings together, often on a scale ranging from *sure old* to *sure new* in recognition

memory paradigms. However, according to the operationalised definition of metacognition ("behaviour about behaviour") the former type of judgements (*type-I* confidence judgements; see Galvin, Podd, Drga & Whitmore, 2003, for a detailed explanation of the type-I/II terminology) does not fall under the category metacognition, and possibly even relies on different neural mechanisms (Overgaard & Sandberg, 2014; Overgaard et al., 2006). I therefore only study type-II confidence judgements in this thesis. Another argument supporting my decision is that type-II confidence judgements have the advantage that they can be used in combination with most psychological paradigms (Overgaard & Sandberg, 2014), and they might therefore be more transferrable across tasks (De Gardelle & Mamassian, 2014) or even domains.

A critical question in metacognition research is how metacognitive accuracy – for example, the correlation between confidence and objective performance – is assessed. The relative accuracy (or *resolution*) of metamemory judgements is usually measured through Goodman-Kruskal gamma correlations (Goodman & Kruskal, 1954), which measure the relationship between confidence judgements and objective test performance. Gamma correlations have been considered as one of the best ways to measure resolution in feeling-of-knowing judgements (Nelson, 1984), but they can be easily distorted when one of the rating conditions contains only a small number of trials (Yokoyama et al., 2010). Given that we can expect participants to choose some of the confidence categories less often than others – for instance *probably correct* more often then *certainly wrong* – this measure is not suitable for the paradigms used in this thesis.

A related measure often used in the research on metamemory is *calibration*, or absolute accuracy. This measure expresses the correspondence between the rated and the actual probability of being correct in a given condi-

tion (Dunlosky & Metcalfe, 2009). Calibration also expresses whether people are over- or underconfident, meaning that their subjectively-rated accuracy is higher (overconfidence) or lower (underconfidence) than their actual, objective accuracy. Calibration is usually measured by calculating a component of a *Brier score* (Brier, 1950), which is an averaged probability score, that is a score expressing correspondence between a probability rating and the actual probability. Calibration values lie between 0 (perfect calibration) and 1 (lowest possible calibration), but in most studies, calibration scores are lower than 0.1 (Baranski & Petrusic, 1994). However, to estimate calibration, confidence ratings have to be measured using a numerical confidence scale, given that subjective probabilities of being correct are compared to objectively measured error rates. In this thesis, I use a verbal confidence scale. The reason for this decision is discussed in detail in Chapter 2. I will therefore not use calibration as a measure[3].

Metacognitive judgements are often analysed using type-II signal detection theory (SDT) measures, as presented in Figure 4. Such models assume that participants decide whether a just-made response was correct or incorrect by judging evidence accumulated in favour of the chosen response option against a decision criterion (Galvin et al., 2003; Higham, Perfect & Bruno, 2009). Just as we can use SDT to quantify participants' ability to categorise correctly the presence or absence of a stimulus, we can use second-order SDT to quantify participants' ability to categorise their responses as correct or incorrect. In this context, a *hit* is a correct response classified as correct; a *miss* is an error classified as a correct trial; a *false alarm* (FA) is a correct trial classified as an error; and a *correct rejection* (CR) is an error classified

[3]Of course, one could circumvent this problem by using only categories of a verbal confidence scale that refer to the extreme points of the confidence scale (e.g., *sure correct* – 100% of being correct) or guessing (50% of being correct), but this would still mean that a large proportion of the confidence data cannot be used.

as such. Similar to the type-I SDT model, it is assumed that there are two Normal distributions representing the stimulus categories, which overlap to some extent. The distance between the means of these distributions expressed in units of standard deviation is first-order sensitivity, d', the ability to discriminate between response alternatives. Furthermore, the participant sets a first-order criterion, c, to distinguish between these distributions. According to the position of a stimulus relative to this response criterion, a decision is made. An unbiased criterion would be placed dividing the two distributions at their intersection, but often participants respond with a certain bias to one or the other response alternative. Moreover, a type-II criterion is assumed for metacognitive judgements, which is placed on either side of the first-order, as also illustrated in Figure 4: If evidence in favour of the chosen response alternative lies between the first- and the second-order criterion, a low-confidence trial is indicated.

SDT parameters calculated from these secondary responses have interesting characteristics for metacognitive analyses. First of all, type-II sensitivity is the ability to differentiate between one's own correct and incorrect answers. It can therefore be seen as a general metacognitive accuracy or metacognitive efficacy. The type-II criterion on the other hand expresses whether participants are biased towards accepting their own answers as correct or more often judging them to be errors. Metacognitive ability is furthermore sometimes operationalised as the area under the Receiver Operating Characteristic (ROC) type-II curve – that is plotting type-II hit rates against type-II false alarm rates for each of the confidence categories (Fleming & Dolan, 2010). An example of an ROC curve is given in Figure 5 (grey curve).

Caution is advised when estimating type-II SDT parameters, however, given that the distributional assumptions, such as normally distributed noise,

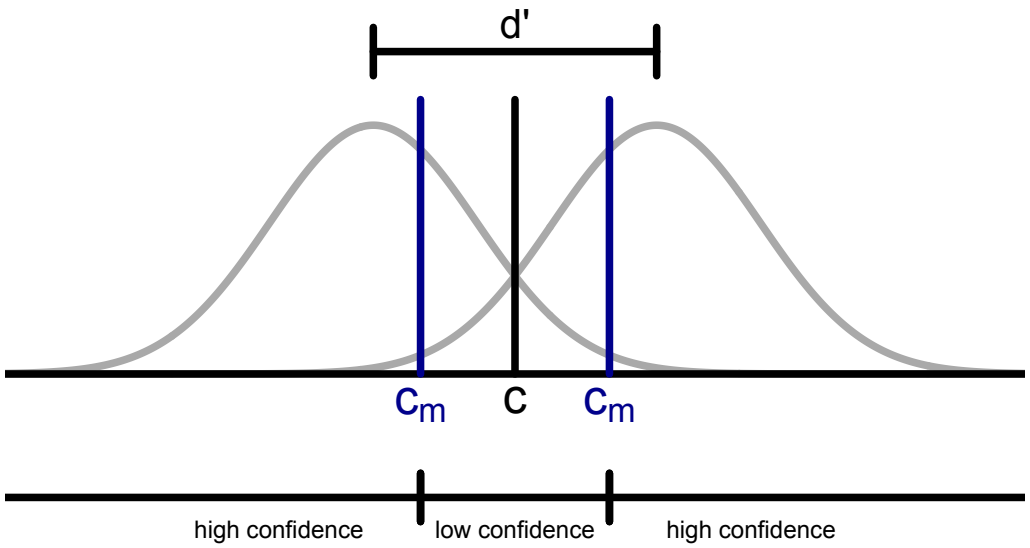

Figure 4: Signal detection theory assumes two overlapping Normal distributions representing the two choice alternatives (depicted in grey). The amount of overlap of these distributions determines the participant's overall sensitivity in distinguishing the two evidence sources (d'). The two distributions are divided by a criterion (c; black vertical line), and samples smaller than this criterion are classified as belonging to one class and samples larger than it to the other class. Type-II signal detection theory furthermore assumes the placement of two confidence criteria (c_m; blue vertical lines), symmetrically around the first-order criterion.

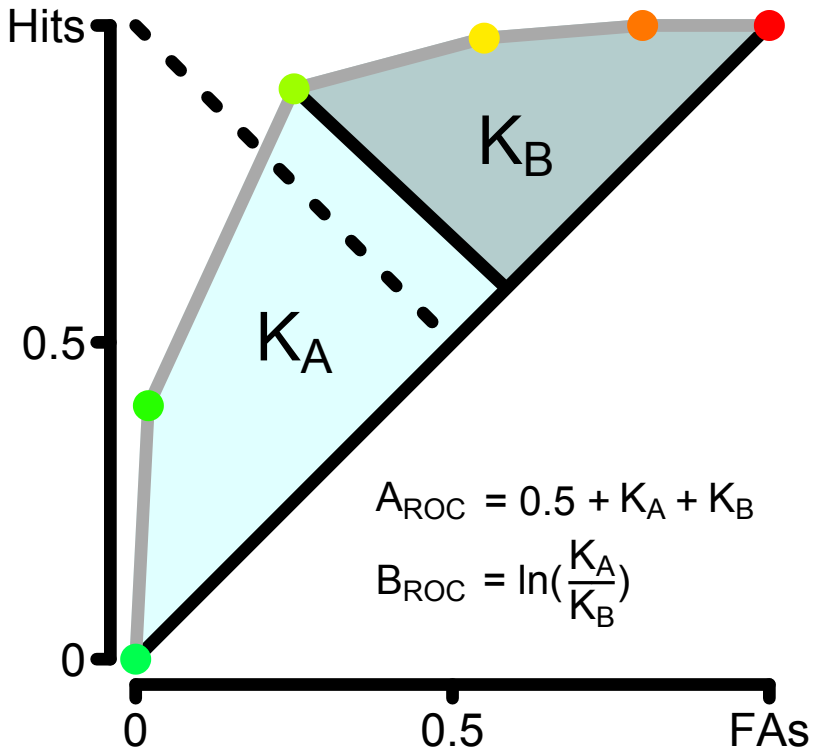

Figure 5: A Receiver Operating Characteristic (ROC) type-II curve plots type-II false alarms (FAs; correct responses, classified as errors) against type-II hits (correct responses, classified as correct) cumulatively for every confidence level. For example, the second data point (light green) from the origin plots FAs against hits for all trials on which the participant chose the highest level of confidence, whereas the next data point (yellow-green) includes trials from both the highest and second-highest levels of confidence. The area enclosed by the ROC curve and the main diagonal can be divided into two sub-areas, K_A (light blue) and K_B (dark blue), from which non-parametric estimates of metacognitive sensitivity, A_{ROC}, and metacognitive bias, B_{ROC}, can be calculated (Kornbrot, 2006). The minor diagonal is shown as a dashed black line (see main text).

are often violated on the second-order level (Barrett, Dienes & Seth, 2013; Evans & Azzopardi, 2007; see also Fleming & Lau, 2014, for a review). A distribution-free approach, which circumvents these problems, has been proposed by Kornbrot (2006). This approach estimates non-parametric metacognitive sensitivity (A_{ROC}) and bias (B_{ROC}) from type-II ROC curves, as presented in Figure 5. To calculate these distribution-free SDT measures, the area between the ROC curve and the main diagonal first has to be divided into two parts (see light and dark blue shades in Figure 5), K_A and K_B, the areas of which can then be calculated using simple geometry given that those areas can be subdivided into several triangles. Metacognitive sensitivity A_{ROC} is the sum of both of these areas plus 0.5, to account for the area below the main diagonal. A_{ROC} can take values up to 1, which would express perfect metacognitive sensitivity. Similarly, metacognitive bias, B_{ROC}, can be calculated as the natural logarithm of the ratio of the two areas K_A and K_B. For instance, if participants are overconfident (as in Figure 5) then they are relatively more likely to use the higher confidence levels of the scale. Their ROC curve – which plots high-confidence levels closer to the origin ('more green' dots) and low-confidence levels further away from the origin ('more red' dots) – will thus rise faster and have a larger amount of data points above the minor diagonal (dotted line), resulting in a smaller area of K_B and therefore a larger B_{ROC}.

However, there is one issue concerning the interpretation of metacognitive sensitivity measures: Metacognitive sensitivity will always – at least to some extent – depend on first-order sensitivity, d' (Fleming & Lau, 2014). This makes it very difficult to compare metacognitive judgements across conditions that differ in basic task performance, which is something that will conceivably happen very often. Maniscalco and Lau (2012) have therefore proposed a

method to estimate type-II SDT sensitivity relative to their underlying type-I parameters (see also Barrett et al., 2013). Such corrected type-II sensitivity measures (*meta-d′*) can then be compared to type-I sensitivity (d') by forming a ratio from the two measures: Metacognitive efficiency (*M-ratio*) expresses how well participants differentiated between their own correct responses and errors while taking first-order performance into account. This parameter can be interpreted as follows: A value of 1.0 means that participants used all information from the primary decision for their metacognitive judgement. A value of 0.7 means that they only used 70%. Values larger than 1.0 express that participants used more information for their second-order judgement, presumably because of ongoing evidence sampling (as often seen in speeded tasks where people make many premature responses). *M-ratio* is therefore an index of the degree to which metacognitive sensitivity (*meta-d′*) falls below or exceeds what would be expected given primary task sensitivity (d').

In this thesis, I report metacognitive efficiency, *M-ratio*, whenever I am interested in how well participants were capable of distinguishing between their own correct and error responses. Rather than reporting the raw parameter, I calculate the common logarithm of it to correct for non-normality (Fleming & Lau, 2014). Values larger than zero therefore signal that participants used more information for their second-order judgement. I use this measure rather than, for example, A_{ROC}, as I will study and compare conditions that differ in basic task performance. In terms of metacognitive bias parameters, I use the non-parametric approach originally proposed by Kornbrot (2006), and report B_{ROC} whenever I am interested in how participants mapped their metacognitive judgements to the rating scale used. As stated above, this approach is preferable to 'standard' SDT ways of calculating metacognitive bias, as it does not make any distributional assumptions, which have often been found to be

violated for metacognitive data. Moreover, I prefer to use B_{ROC} as a measure of relative under- and overconfidence as opposed to calibration, given that the latter parameter would dictate the use of a numerical confidence scale rather than a scale with verbal labels that I use in the experiments reported in this thesis (see Chapter 2, for a more detailed discussion regarding this decision).

1.8 Statistical methods used in this thesis

Most statistical tests reported in this thesis were conducted using the statistical programming language *R* (R Core Team, 2013), version 3.0.2. Within this programming language, several additional packages were used: *Hmisc* for correlation analyses (Harrell Jr & Dupont, 2014), and *outliers* for the outlier analyses reported in Chapter 5 (Komsta, 2011).

Whenever within-subject linear trends are reported for an analysis of variance (ANOVA) model, they were calculated with SPSS, release 22.0.0.0 (IBM, 2013). In case of violations of sphericity according to Mauchly's sphericity test (Mauchly, 1940) for repeated-measures ANOVAs, Greenhouse-Geisser corrected F-values are reported (Greenhouse & Geisser, 1959). Furthermore, all ANOVAs are based on type-II sums of squares (SS). However, given that the choice of SS is a somewhat controversial issue (Langsrud, 2003), all tests of unbalanced designs (applies only to Experiment 7) were also conducted with type-III SSs (using the *R* package *ez*; Lawrence, 2013) and any substantial differences in results are reported. Moreover, Welch's t-tests were used whenever the means of two samples were compared to account for unequal variances. Approximated, non-integer degrees of freedom will be reported wherever this applied.

Whenever a null effect is of specific interest, this effect will be also

analysed with Bayesian statistical methods using the *R* package *BayesFactor* (Morey & Rouder, 2014; Rouder, Speckman, Sun, Morey & Iverson, 2009). The reason for this is that these methods permit estimating the probability with which the null hypothesis is true given the data – something that cannot be achieved using conventional null hypothesis significance testing (NHST). For the ANOVAs and *t*-tests in question, *Bayes Factors* (BF) will be reported. Bayesian statistical approaches do not differentiate between the alternative and null hypothesis, but I still refer to a BF in favour of the null hypothesis as BF_{NULL} for shorthand and BF will be used whenever the Bayes Factor was calculated with regard to the alternative hypothesis. As cutoff values for all BFs reported in this thesis, values proposed by Kass and Raftery (1995) will be used. According to these guidelines, values below the cutoff value between 1 and 3 fall into the category of "not worth more than a bare mention", whereas values between 3 and 20 are considered to be "positive" evidence in favour of the hypothesis in question. Values between 20 and 150 are "strong" evidence, and above 150 are classified as "very strong" (Kass & Raftery, 1995, p. 777).

If not stated otherwise, all error bars in the figures presented in this thesis are based on within-subject confidence intervals, calculated according to the method proposed by Loftus and Masson (1994; see also Masson, 2003). These confidence intervals are therefore based on the error term of the respective within-subject ANOVA of the experimental factor in question.

A subset of the analyses reported in this thesis was calculated using MATLAB, version R2012a (7.14.0.739; MathWorks, 2012). Amongst these analyses are the multivariate pattern classification approaches used in Experiments 4 and 5 (based on code used by Macdonald, Mathan & Yeung, 2011; Parra et al., 2002; Steinhauser & Yeung, 2010), as well as diffusion model fits that will be discussed in Chapter 5, which were fitted using the toolbox DMAT,

version 0.4 (Vandekerckhove & Tuerlinckx, 2008). Most of the EEG analyses (EXPERIMENTS 4, 5, and 8) that will be reported in this thesis were carried out using the MATLAB toolbox EEGLAB, version 11.0.5.4b (Delorme & Makeig, 2004). Recording and preprocessing of the EEG data was carried out using SCAN software (versions 4.3 and 4.4; Compumedics Neuroscan, 2003, 2007), with ocular artefacts corrected using a regression-based approach proposed by Semlitsch, Anderer, Schuster and Presslich (1986).

Finally, for type-II SDT analyses reported above, I used both a parametric and a non-parametric approach. First-order sensitivity, d', metacognitive sensitivity, $meta\text{-}d'$, and metacognitive efficiency, $log(M\text{-}ratio)$, were calculated using the MATLAB code provided by Maniscalco and Lau (2012). Metacognitive bias (B_{ROC}), on the other hand, was calculated with the non-parametric procedure proposed by Kornbrot (2006), using custom-written code.

1.9 Statement of authorship

A manuscript reporting the results of EXPERIMENT 4 (Chapter 3) has recently been published in The Journal of Neuroscience. The manuscript was written by me and my supervisor, Professor Nick Yeung. However, the text included in the thesis has been re-written (except for some paragraphs regarding research methods) and more detail is given compared to the submitted manuscript.

As disclosed in Chapter 3, the model reported there was designed and implemented by me, but further developed in collaboration with Professor Stanislas Dehaene and Dr Lucie Charles, during a one-month visit at Neurospin, INSERM CEA (Paris).

EXPERIMENT 7 (Chapter 5) was a collaboration with Professor Robert Rogers (University of Oxford, Department of Psychiatry at the time, now

at Bangor University). As stated clearly in the research methods for this experiment, I programmed and installed the experimental software but was not involved in the data collection or administration of the amino acid drinks because this was a double-blind design and I was involved in the assignment of participants to conditions to ensure balanced groups. I then processed and analysed the data.

Experiment 8 (Chapter 5) was a collaboration with another graduate student, Ms Elizabeth Michael. Ms Michael programmed the adaptive staircase algorithm as well as her part of the experiment (first four blocks), while I programmed my part of the experiment (last three blocks). All data collection was conducted together, with two experimenters present for every participant. The analyses in my thesis were all carried out by myself and they all refer to only my part of the experiment, with the exception of one diffusion model analysis for which a large number of trials was needed. All of these details are also highlighted in the respective thesis chapters. I hereby confirm that everything else in the thesis is wholly my own work.

Chapter 2

Methodological issues arising when studying metacognition

The goal of the experiments in this chapter was to establish a robust paradigm suitable for measuring metacognition in decision making. As discussed in Chapter 1, guidelines on how best to test metacognition have already been developed in the literature focusing on memory (Nelson & Dunlosky, 1991; Dunlosky & Metcalfe, 2009) and consciousness (Seth et al., 2008; Overgaard & Sandberg, 2014; Reingold & Merikle, 1988). These guidelines have contrasted, for example, direct and indirect ways of assessing metacognition (Reingold & Merikle, 1988). In this context, it has moreover been suggested that more time between learning the correct answer to a question and rating the probability of correctly remembering these answers will increase participants' accuracy in the latter judgement (the "delayed-JOL effect"; Nelson & Dunlosky, 1991). However, such guidelines are currently missing in the context of metacognition in decision making. In this chapter, I therefore review findings related to such guidelines, highlight several outstanding questions, and then address these questions in three behavioural experiments.

Metacognitive judgements cannot be measured on their own. They are always embedded into another task, for instance a decision making task. It therefore has to be decided which paradigm is best suited to be such a primary task. In perceptual decision-making paradigms, participants are usually asked to categorise a stimulus, which is assumed to be achieved by integrating noisy sensory information until a threshold level of evidence is reached (Ratcliff & Smith, 2004). Value-based decision making, on the other hand, requires the participant to make a preference choice, which depends on the different values assigned to the choice options by the participant. The choice itself therefore relies heavily on internal value representations, which introduces an additional source of noise that is difficult to control experimentally and increases the number of processing steps, as reflected in the framework by Sugrue, Corrado and Newsome (2005). In perceptual decision making, however, the choice is driven by external stimulus properties to a large extent (Gold & Shadlen, 2007). The experimenter thus has precise experimental control over the noisy evidence that has to be integrated and can manipulate the difficulty of a task by independently manipulating the different dimensions of a stimulus (for example stimulus intensity and reliability; De Gardelle & Summerfield, 2011). I therefore used a perceptual decision-making task to investigate decision confidence for the experiments reported in this thesis.

Figure 6 presents an example stimulus for such a simple perceptual decision-making task: Participants are presented with two rectangular fields, each containing a number of dots. They then have to quickly decide which of the fields contained more dots. The task used here is similar to one used by Ratcliff, Van Zandt and McKoon (1999), who used a dot count task in which participants had to quickly determine whether the number of asterisks contained in a field was small or large. The number of asterisks presented

mapped closely onto the reported evidence accumulation rates (Ratcliff et al., 1999, Figure 4). The rationale behind using a comparison between two fields rather than a categorisation was that no internal representation has to be formed of what 'many' or 'few' dots means. Arguably, the task is therefore less influenced by response biases towards one of the two response options, caused by fluctuations in the internal representation of the two response categories.

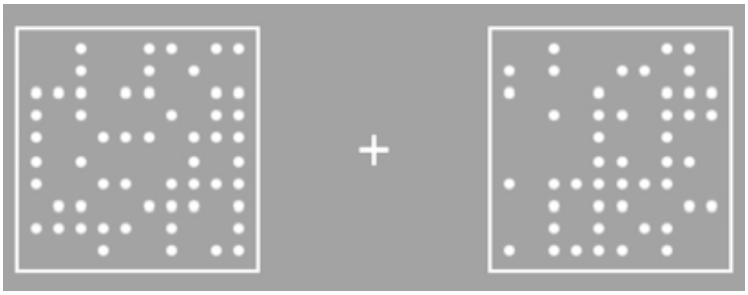

Figure 6: Example of a stimulus in the dot task. Participants had to quickly judge which field contained more dots and signal their decision by pressing a corresponding button.

Basic task parameters of this paradigm could then be explored. Difficulty of the first-order judgement has a substantial effect on second-oder judgements – the amount of evidence available for the first-order decision is thought to affect the second-order decision process (Maniscalco & Lau, 2012). For example, if participants are guessing in the first-order judgement, they are also likely to guess in the confidence judgement (but see Charles et al., 2013). However, that does not mean that the task has to be made as easy as possible. In fact, the absolute number of errors is just as important and has to be kept high enough so that error detection can be measured. In the present task, difficulty was manipulated in small discrete steps by increasing or decreasing the difference in dots between the two fields. In Experiment 1, I tested ten levels of difficulty to determine which of these are best-suited for the subsequent experiments of my thesis. For most studies in this thesis, the goal was to find

a difficulty condition in which participants would commit 15-20% errors under moderate speed pressure. Some of these errors will have been caused by the speed pressure and are therefore likely to be detected. Other errors were caused by the small difference in dots on each side, which can sometimes be misrepresented internally due to processing noise, leading to undetected errors (Scheffers & Coles, 2000; Rabbitt & Vyas, 1981).

Experiment 1 also focused on the question whether it is better to ask participants to rate their confidence on a graded scale, or whether binary error detection yields better metacognitive accuracy. I therefore designed a task in which, after each trial, a metacognitive judgement was required but it randomly varied whether a 2- or 6-point scale was used. Clearly, the 6-point scale provides the experimenter with a more fine-grained measure of confidence. The question arises as to whether the increased resolution comes at a cost imposed by the additional demand of making finer judgements.

My second experiment focused on the temporal dynamics of metacognitive judgements. More precisely, I aimed to identify the ideal time at which metacognitive judgements should be probed. On the one hand, metacognitive judgements are made with regard to another mental representation – the primary decision – such that we might expect metacognitive accuracy to decrease as the judgement is delayed relative to that decision. Such memory decay might play an important role that needs to be considered in this context, independent of the locus assumption, that is whether metacognitive judgements are formed at the same time as the decision and then stored until a judgement is required, or whether they are formed at a later point in time based on whatever is remembered from the primary choice. This could, for example, play a role for both the time heuristic (Audley, 1960; Moreno-Bote, 2010; Zylberberg et al., 2012), where the response speed of the primary de-

cision has to be remembered, as well as for any theory that assumes a balance-of-evidence mechanism (Kepecs & Mainen, 2012; Van Zandt & Maldonado-Molina, 2004; Vickers & Packer, 1982; De Martino et al., 2013), where the final amounts of evidence accumulated in the two counters have to be stored until the metacognitive judgement is made. One could therefore expect metacognitive accuracy to decrease over time.

On the other hand, a recent study by Resulaj, Kiani, Wolpert and Shadlen (2009) provided evidence suggesting that the integration of post-decision evidence is crucial to explain *changes of mind*, using a task in which participants were asked to judge the direction of a cloud of moving dots. Participants signalled their decisions by moving a handle either to the left or to the right. Precise motion trajectories of this movement were measured. The results indicated that on a proportion of trials, participants changed the direction of their hand movement midways through the trajectory. The authors interpret these trials as evidence for changes of mind and provided a quantitative model of their behaviour, according to which people continue to accumulate evidence even after they make their decision. This additional information sampled from the "processing pipeline" (p. 264) can then cause them to change their mind. These changes of mind are most likely to happen after the participant committed a mistake. Such self-correcting mechanisms have been studied extensively by Rabbitt and colleagues (Rabbitt & Vyas, 1981; Rabbitt, 1966), who found that these correcting responses happen even if participants were explicitly instructed not to correct their mistakes (Rabbitt, 2002). These findings suggest that the brain processes stimuli even after a response to those stimuli was made, thereby matching findings from Ding and Gold (2012), who found in nonhuman primates that stimulus evidence affects activity in frontal eye field (FEF) neurons even after a decision was made. All these studies show that

even in the period after a choice has been made decisional information is processed and can even cause the initial choice to be reversed, thereby supporting the idea that there is post-decision processing. Recent studies have argued that confidence might be formed during this time (Pleskac & Busemeyer, 2010; for a review see Yeung & Summerfield, 2014, 2012). Following from this model, the more time participants are allowed to think about their just-made decision the higher their metacognitive accuracy.

The second experiment therefore aimed to investigate and quantify these possible effects on metacognitive accuracy. To do so, I systematically varied the response-stimulus interval (RSI) between the primary (dot-decision) response and the onset of the confidence scale, thus influencing the amount of time participants spend on internally developing confidence decisions.

Finally, Experiment 3 asked whether metacognitive judgements are fundamentally different from basic task processes such that making them represents a different, potentially interfering task. In other words, metacognition could depend on internal processes that are responsible for the original decision versus it could resort to entirely different processes. I therefore tested whether decision-making tasks with metacognitive judgements can cause task-switching effects, such as switch costs. Switch costs can be found when participants start a new task, and they are reflected in higher RTs and error rates (Monsell, 2003; Allport, Styles & Hsieh, 1994; Jersild, 1927; Meiran, 1996). Switch costs arise from the necessity to reconfigure the current task-set to match that of the new task, thus reorganising mental resources. In the latter case, there is reason to believe that task-sets need to be reconfigured prior to making a confidence judgement, which should cause switch costs if compared to blocks in which only the decision task was required.

Taken together, four key questions were addressed in this chapter.

The first focused on the nature and difficulty of the primary task in which metacognition can be measured. The second and third questions focused on the scale with which metacognition is assessed and when is the best point in time to measure confidence. Finally, the possible interaction between the primary and secondary judgement tasks were addressed by asking whether they rely on the same or different task processes, arguing that switch costs could possibly arise from such a design. These results from these experiments will then be used to create a paradigm with which the questions in this thesis can best be addressed.

2.1 Experiment 1: Influence of metacognitive rating scale

The present experiment focused on two questions. The first concerns the suitability of the perceptual decision-making paradigm, as described in the introduction to this chapter. With the second question, two different confidence scales were directly compared to assess which of them is a more suitable tool for measuring confidence.

A range of different confidence scales has been used in the literature, such as binary, discrete graded, and continuous scales. Those scales have previously been compared, both with regard to their effects on the accuracy of metamemory judgements (Dunlosky & Metcalfe, 2009), as well as comparing them to other measures of awareness (Sandberg et al., 2010; Seth et al., 2008; Overgaard & Sandberg, 2014; Dienes, 2008). Despite commonly concluding that confidence scales are not suitable for measuring awareness to external stimuli, these studies are still very informative for the purpose of the present experiment, given their direct comparison of different types of scales.

One reason for favouring a binary scale is that it is easy and straightforward to interpret the two response options, which are usually *correct* and *error*. Responding on such a scale should therefore be simple. Moreover, participants should be able to make judgements very quickly on such a scale, assuming that Hick's law of a growing RTs with the number of response options also holds for confidence judgements (Hick, 1952). Binary scales are often used in research on error monitoring (Charles et al., 2013; Endrass, Franke & Kathmann, 2005; Wessel, Danielmeier & Ullsperger, 2011); and evidence exists to suggest that they lead to more accurate judgements compared to other scales: Studies by Tunney and colleagues (Tunney, 2012; Tunney & Shanks, 2003) suggested that only binary scales can detect differences between conscious and unconscious states. The authors had participants complete an artificial grammar learning task. Participants rated their confidence after a response on either a binary ("more confident" versus "less confident"; Experiment 1B in Tunney & Shanks, 2003) or a continuous scale (any number between 50 and 100; Experiment 3 in Tunney & Shanks, 2003). Confidence and objective accuracy covaried only in case of the binary scale, suggesting that although participants might possess correct metacognitive knowledge, they were unable to express it on the continuous scale.

Graded confidence scales are common in research on consciousness (e.g., Wierzchon et al., 2014; Li, Hill & He, 2014) and recognition memory (e.g., Busey, Tunnicliff, Loftus & Loftus, 2000; Pannu, Kaszniak & Rapcsak, 2005). This type of graded confidence scale has also been adopted in the decision-making domain (Fleming, Huijgen & Dolan, 2012; Zylberberg et al., 2012; Bahrami et al., 2012; Fleming et al., 2010; Baranski & Petrusic, 1994; Pleskac & Busemeyer, 2010; Baranski & Petrusic, 1998; De Martino et al., 2013; Petrusic & Baranski, 2003, just to name a few examples). If we assume

that participants can make unlimited fine-grained judgements, then using a graded scale is clearly advantageous for the experimenter because of its higher measurement resolution. The question remains as to whether this increment in resolution comes at the cost of decreased metacognitive accuracy. Participants could, for example, be confused as to what the different categories mean. This confusion could introduce noise into the judgement process and increase confidence RTs because participants have to continuously re-read the labels of the categories or because participants vary in how they interpret or apply the different labels. However, evidence exists that people are able to make such finer-grained judgements without additional processing costs. Findings from a study by Dienes (2008), for example, suggest that participants were just as accurate when using binary as when using continuous scales. The paradigm used in this experiment was similar to the task described above by Tunney and Shanks (2003).

There exists one particular type of categorical graded scale, which will be called quasi-continuous scale here. Such scales are fine-grained enough to appear continuous to the participant. The participant is then asked to click on any position on this scale or type in a perceived probability of being correct (Koriat, 2011; Baranski & Petrusic, 1994; Macdonald et al., 2011; De Martino et al., 2013). Using a scale with such fine gradations has one crucial consequence: Measuring confidence RTs becomes increasingly difficult, that is the time it takes to give a confidence rating following the onset of the scale. The importance of measuring those RTs has previously been highlighted. Pleskac and Busemeyer (2010), for example, implemented a model which assumes confidence judgements are formed through post-decision accumulation of decision evidence. Their 2DSD model provided a good fit to RTs, error rates, confidence judgements and confidence RTs, which they regard as equally important

measures, reflecting the same underlying judgement process (see also Moran, Teodorescu & Usher, 2015). Furthermore, Baranski and Petrusic (1998) suggested that confidence judgements are formed post-decisionally under speed pressure. The length of the time to make a confidence judgement thus reflects some important properties of the underlying processes and therefore quasi-continuous scales will be excluded as a possibility.

In this thesis I therefore use a confidence scale with discrete, monotonic categories to facilitate precise measurement of the confidence RTs. Whether this scale should be binary or graded, however, must be tested empirically, given that existing findings are inconclusive as to whether graded scales come at an additional processing cost for the participant. The present study therefore addresses this question to compare a 6-point scale with a binary scale.

Furthermore, the scale used in all subsequent experiments of this thesis had verbal as opposed to numerical labels, marking the different regions and categories of the scale. Such categories can for example be "Not confident at all", "Slightly confident", "Quite confident", and "Very confident" (Sandberg et al., 2010). The choice in favour of a verbal category was supported by a number of empirical findings, which will subsequently be discussed.

Confidence scales usually have either verbal (*certainly wrong*, *probably wrong*, etc.) or numerical categories, the latter often in the form of probabilities that the last response was correct. Mixed findings exist with regard to the question as to which of these scales yield more accurate confidence judgements. Some findings suggest that probabilities are the more precise way of describing confidence categories, given that verbal descriptions of probabilities can be somewhat noisy, participant- and context-specific (see Budescu & Wallsten, 1995, for a review). This position is also supported by the previously mentioned fact that only with a probability scale can calibration be calculated,

which is the correspondence between the rated and the actual probability of being correct in a given condition (Dunlosky & Metcalfe, 2009). There are other findings, however, which suggest that verbal scales are more suitable tools for measuring confidence. Zimmer (1983), for example, suggested that verbal scales lead to more accurate measures, arguing that this was the case because they could be used much more intuitively than numerical ones. Other studies concluded that participants display similar levels of sensitivity in their confidence responses no matter which type of scale was used (Dienes, 2008; Wallsten & Budescu, 2009; Wallsten, Budescu & Zwick, 1993). From these findings, one can conclude that numerical ratings are not necessarily better tools for measuring confidence. In fact, an argument which speaks clearly in favour of verbal categories is that little instruction is needed, meaning that participants can use the scale without much practice. In the course of this thesis, no calibration curves will be analysed and this disadvantage of verbal scales can therefore be neglected.

Moreover, I chose to use a confidence scale that includes *error* judgements as lowest confidence level, similar to studies on confidence and error monitoring (Baranski & Petrusic, 1994; Koriat, 2011; Wallsten et al., 1993; Buratti & Allwood, 2012, just to name a few; see Wessel, 2012, for a review). With such a scale, it is possible to measure whether participants detected an error, and therefore changed their mind. Such changes of mind, or response reversals, have previously been studied (Van Zandt & Maldonado-Molina, 2004; Rabbitt, 1966), and found to occur even when stimulus input was no longer given (Resulaj et al., 2009). This type of scale stands in contrast to many studies, which instead explicitly define *guessing* as the lowest level of confidence (e.g., Dunlosky & Metcalfe, 2009; Graziano, Parra & Sigman, 2010; Selmeczy & Dobbins, 2013; Graziano & Sigman, 2009; Koriat, 2008; Edelson, Dudai,

Dolan & Sharot, 2014). The reason why many studies choose to ignore error detection links to the notion that two qualitatively different types of errors exist. As mentioned previously, Scheffers and Coles (2000) distinguish "errors due to premature responding" from "errors due to data-limited processing". The former are usually fast and detected in most cases, while the latter ones remain often undetected, participants report to have guessed the answer on those trials, and they often form the slow tail of the RT distribution. Many confidence studies do not impose speed stress and *sure errors* are therefore rare enough to be neglected. Error detection constitutes an extreme case of confidence, however, and to study such error detection, such premature responses have to be included as well. It therefore has to be concluded that if we want to design a paradigm that tests participants' metacognition, we should aim for a speed instruction. Indeed, findings reported by Pleskac and Busemeyer (2010) suggest that the speed-accuracy regime of the primary task has a crucial effect on confidence judgements. With sufficient speed pressure, confidence judgements have been found to vary more and there is an increased difference in mean confidence ratings for errors and correct trials, that is an increase in resolution or relative metacognitive accuracy (see also Baranski & Petrusic, 1994; Moran et al., 2015).

To sum up, findings from the consciousness, metamemory and confidence literature were reviewed and a verbal, categorical confidence scale was then chosen for all subsequent experiments in this thesis. This scale ranges from *certainly wrong* to *certainly correct* and the primary decision paradigm will impose speed stress on the participants, so that more premature responses occur, which are errors that are usually easier to detect than errors arising from the difficulty of the stimulus material. In the present experiment, two versions of such a scale were tested against each other – one with six response

categories and one with two response categories. To foreshadow, participants were found to be able to make more fine-grained judgements with the 6-point scale and these judgements did not occur at a cost of lower metacognitive performance. The primary decision-making task was also evaluated and found to be a suitable framework for conducting studies on metacognition.

In Section 2.1.2.1, I first assess whether the current paradigm allows a precise manipulation of task difficulty. The effect of difficulty on correct RTs and error rates is analysed, as well as the effect of difficulty on raw confidence. The analyses in the second part of the results section (Section 2.1.2.2), are focused on effects of the different confidence scales. Error rates are analysed as a function of subjective error probability, as well as mean confidence and the distribution of confidence judgements. The main question is of course whether the scale has an effect on type-II accuracy. The focus is therefore on metacognitive efficiency as a measure that allows one to directly compare the two scales. Moreover, data from the 6-options scale are analysed with a non-parametric SDT approach based on ROC curves, which is not possible for a binary scale. Yet another question addressed is whether using a more fine-grained scale increases the time participants take to rate their confidence, as could be predicted by Hick's law (Hick, 1952).

Finally, additional analyses are addressed in Section 2.1.2.3. Here, I test whether participants' confidence changed over blocks. They could presumably get better at judging their confidence, which would result in higher confidence for correct responses and lower confidence for incorrect responses in late as opposed to early blocks. This should also be reflected in measures of metacognitive efficiency.

2.1.1 Methods

2.1.1.1 Participants

Twenty-five participants were tested, 15 of whom were female, and 3 were left-handed. The participants' ages ranged from 18 to 24 (M = 19.3). All had normal or corrected-to normal vision. One participant had to be excluded because they did not classify a single trial as incorrect in the case of several difficulty conditions. This meant that no SDT measures could be calculated for these conditions. The final sample therefore included 24 participants. All participants gave informed consent and received course credit for their time. Each session lasted approximately 45 minutes, including instruction and debriefing. All procedures were approved by the local ethics committee.

2.1.1.2 Task and procedure

The experiment comprised a series of trials on which participants first performed a perceptual decision task under time pressure, and were then asked to rate their subjective confidence in the decision just made (Figure 7). The perceptual decision task required participants to judge which of two briefly flashed (160 ms) fields contained more dots by pressing a left- or right-hand button. The differences of dots varied in steps of 2 from 2 to 20, resulting in 10 levels of difficulty. For example, for a difficulty of 10, one field contained 45 dots arrayed in a 10-by-10 matrix, the other one 55. There was a 1,520 ms deadline for this decision, and participants were encouraged verbally and through feedback to respond quickly. Speed was stressed so that participants made sufficient numbers of errors to permit planned error detection analyses.

After participants' responses, the screen cleared for 600 ms at which

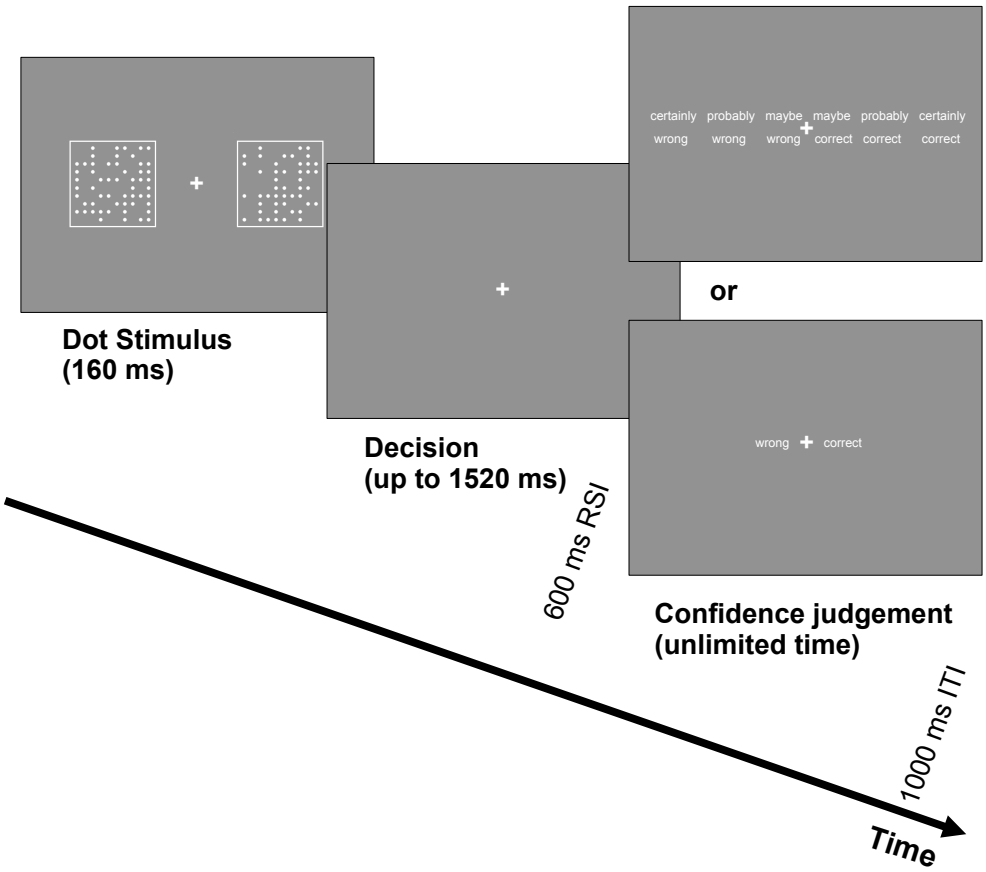

Figure 7: Methods of the dot task. Participants first had to say which of two fields contained more dots by pressing the left or right key. After their response, one of the two confidence scales was presented on screen and participants were given unlimited time to choose how confident they were that their last response was correct. RSI: response-stimulus interval; ITI: inter-trial interval; ms: millisecond.

point either a 6-point confidence scale appeared, with values of *certainly wrong*, *probably wrong*, *maybe wrong*, *maybe correct*, *probably correct*, and *certainly correct*, or a 2-point scale with *wrong* and *correct* as response options. These scales were randomly interleaved across trials. Participants were given unlimited time to indicate how confident they were about the preceding perceptual decision, but were instructed to react according to their first impression. They indicated their confidence by pressing one of six or two keys respectively. After the confidence response, the screen cleared for a 1000 ms inter-trial interval (ITI). Participants were instructed to keep their eyes on the fixation cross between the two fields and they were instructed to react "as quickly and accurately as possible".

Stimuli were presented on a 20″ CRT monitor with a 75 Hz refresh rate and a visual angel of 10.0° by 3.8° using the MATLAB toolbox Psychtoolbox3 (Brainard, 1997; Pelli, 1997; Kleiner, Brainard & Pelli, 2007). All responses were made with an RB-830 Cedrus response pad (San Pedro, CA). This response pad has a lower row of two and an upper row of six buttons. All buttons are arranged ergonomically in semi-circles, so that they can be operated comfortably with the thumb, index, middle and ring finger of the respective hand. Participants responded with their thumbs for the perceptual decision. The fingers used for the 2-point rating scale (index versus middle versus ring finger) varied across blocks to reduce the possibility that the corresponding fingers would be preferentially used in the 6-point scale condition. Half of the participants saw the confidence scale ranging from responses classified as *wrong* on the left to *correct* on the right, and the other half of participants seeing the reverse orientation. This hand assignment was fixed for each participant.

In the first block, participants practised the task without confidence judgements. The RSI before the onset of a new stimulus was 600 ms long. After

errors, an auditory feedback tone was played and the RSI was 1000 ms long, therefore penalising the participants for the incorrect response. In the second and third block, no more error tones were played as they then practised the use of either the 2- or 6-point confidence scale. The RSIs following a response to the dot stimulus were always 600 ms, independent of whether or not an error had been made. Which scale was practised first was counterbalanced across participants. The use of the full scale was encouraged and frequencies of the confidence categories were presented on screen after these blocks so that the experimenter could discuss and remind the participant again to use the entire scale. In the fourth practice block both scale types were mixed. Throughout the whole experiment, participants could not predict which of the two confidence scales would be used, while making their decision. They completed 12 experimental blocks. Each block was 40 trials long, meaning that four stimuli from each of the ten difficulty levels were shown per block, that is one for every side the larger stimulus was displayed on and for each of the two scales. After each block, participants received feedback with regard to their mean correct RT and mean error rate.

2.1.1.3 Data analysis

Behavioural data analysis focused on the resolution of participants' confidence ratings; that is, the degree to which subjective confidence ratings were predictive of objective accuracy. These ratings were treated as varying on an ordinal 6-point scale, with "1" reflecting the least confident response (*certainly wrong*) and "6" reflecting the most confident response (*certainly correct*) for the 6-point scale, and "0" (*wrong*) and "7" (*correct*) for the 2-point scale.

2.1.2 Results

2.1.2.1 Difficulty manipulation

A first set of analyses addressed the question of whether the difficulty manipulation worked as anticipated. The first analysis focuses entirely on primary task performance, that is responses regarding the dot stimulus. The left panel of Figure 8 shows condition averages for correct RTs ranging from most difficult (red bar) to easiest condition (magenta bar). A difference of 2 dots is the most difficult condition and should therefore have the slowest RT, whereas 20 dots difference should be expected to have the fastest RTs. The main effect of difficulty was indeed reliable, $F(2.9,\ 66.7) = 23.3,\ p < 0.001,\ \eta_p^2 = 0.50$. There was also a reliable linear trend, $F(1,\ 23) = 45.9,\ p < 0.001,\ \eta_p^2 = 0.67$. Moreover, the right panel of Figure 8 shows that a pattern similar to the one for correct RTs holds for error rates, where smaller dot differences lead to larger error rates, $F(9,\ 207) = 184.5,\ p < 0.001,\ \eta_p^2 = 0.89$. There was a reliable linear trend, $F(1,\ 23) = 1033.6,\ p < 0.001,\ \eta_p^2 = 0.98$. These analyses indicate that the difficulty manipulation worked as anticipated, with more difficult conditions leading to slower correct RTs and higher error rates.

The second analysis focused on second-order performance, that is the effect of difficulty on average confidence. This analysis assumes an interval scale, as previously used in the confidence literature (see, for example, Baranski & Petrusic, 1998). Figure 9 shows how confidence scaled with difficulty, for both the 6-point (upper panels) and the 2-point scale (lower panels), separately for correct (left panels) and error trials (right panels). Given the arbitrary numerical coding of confidence levels, no direct comparison of the two scales is possible and the analyses therefore had to be conducted separately. To avoid missing data due to a low number of errors especially in the easier conditions,

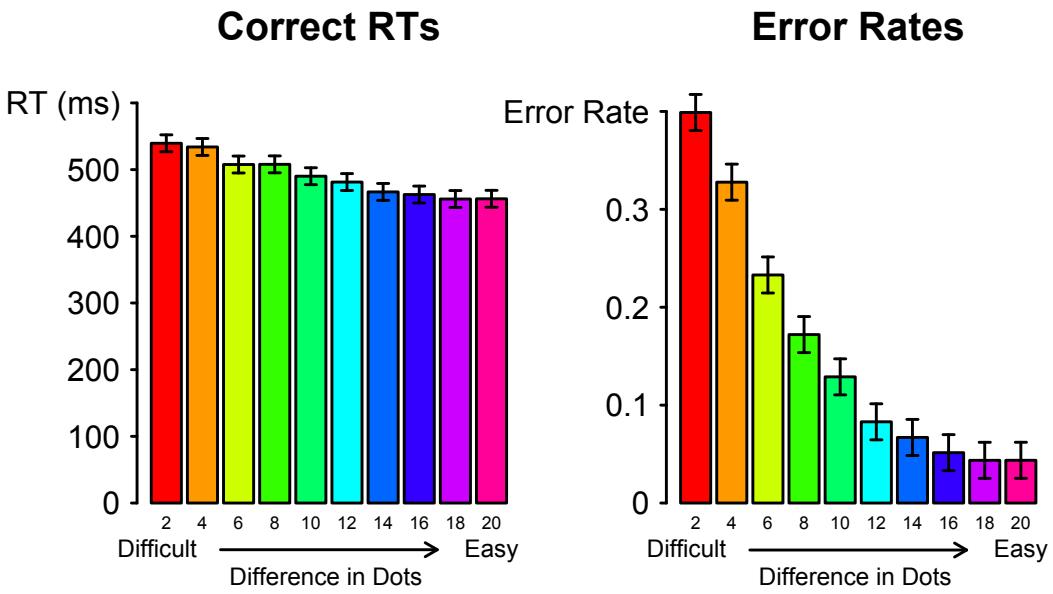

Figure 8: First-order performance as a function of difficulty, that is difference in dots. Correct response times (RTs, left panel) and error rates (right panel) ranging from the most difficult (red bar) to the easiest (magenta bar) condition; ms: millisecond.

the data were aggregated across successive pairs of difficulty level (i.e., 2 and 4 formed the most difficult condition, 6 and 8 the second most difficult one, etc.). Accuracy had a reliable effect on confidence, $F(1,\ 16) = 174.3,\ p < 0.001,\ \eta_p^2 = 0.92$, with confidence significantly lower on error than on correct trials, $M_{cor} = 4.9$, $M_{err} = 3.0$. For correct trials rated on the 6-point scale, difficulty had a significant influence on confidence, $F(4.2,\ 95.6) = 72.2,\ p < 0.001,\ \eta_p^2 = 0.76$, with higher confidence in easier conditions. There was a similar effect for error trials, $F(1.9,\ 29.9) = 13.4,\ p < 0.001,\ \eta_p^2 = 0.46$, but in the opposite direction with lower confidence for easier conditions. This was also reflected in a reliable interaction between difficulty and accuracy, when analysing the data from correct and error trials together, $F(2.1,\ 33.2) = 29.6,\ p < 0.001,\ \eta_p^2 = 0.65$. For the analysis of error trials, as well as the combined analysis, 7 participants had to be excluded because they did not commit any errors in at least one of the conditions.

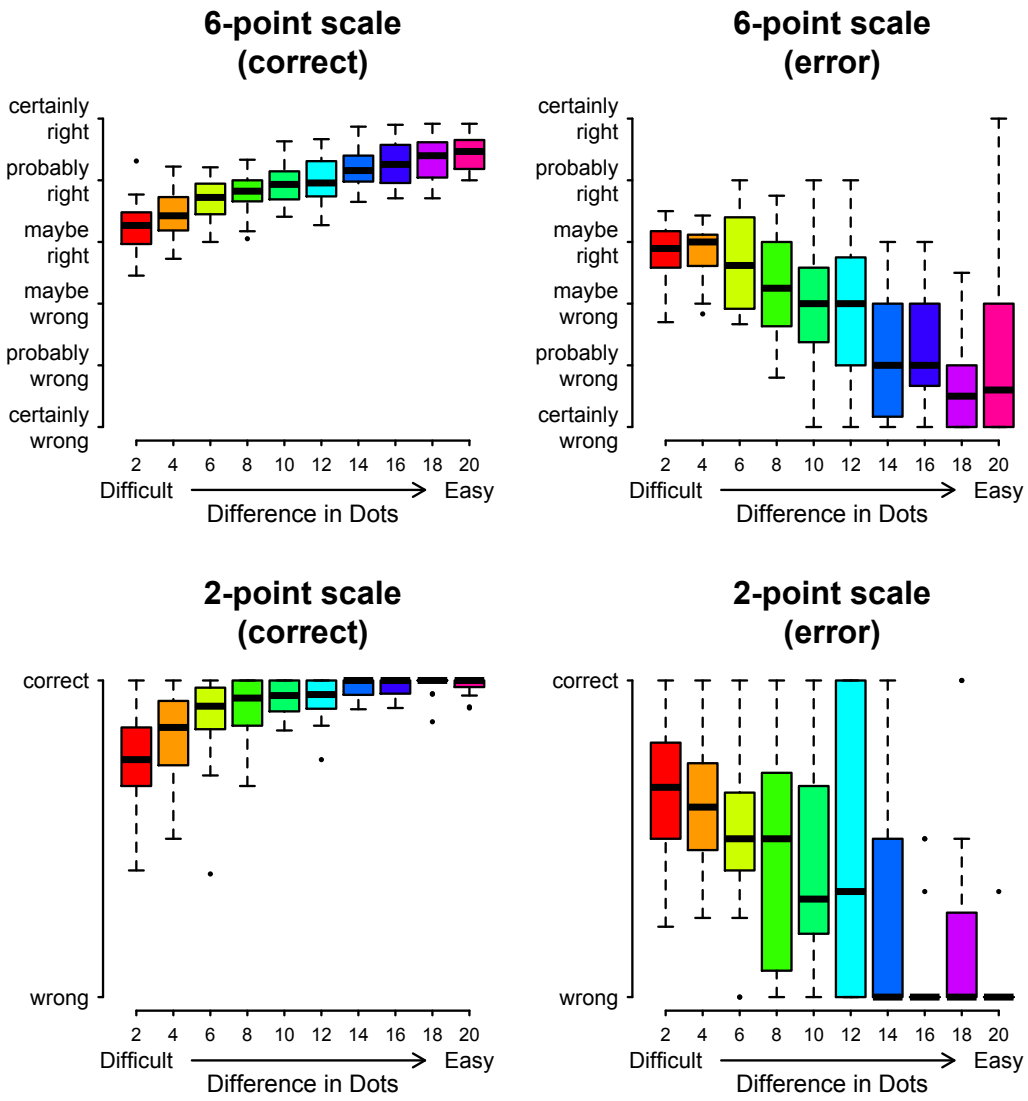

Figure 9: Confidence for the ten difficulty conditions (difference in dots) as box plots; 50% of the data lie within the box, the thick black horizontal line within each box indicates the median, the whiskers extend to 1.5 times the interquartile range (IQR); outliers are displayed as dots; upper panels: 6-point scale; lower panels: 2-point scale; left panels: correct trials; right panels: error trials.

The findings for the 2-options scale were similar to those of the 6-options scale: Both correct, $F(3.5,\ 80.6) = 28.8,\ p < 0.001,\ \eta_p^2 = 0.56$, and error trials, $F(4,\ 52) = 10.5,\ p < 0.001,\ \eta_p^2 = 0.45$, had a reliable effect of difficulty, again in opposite directions, as expressed by a reliable interaction between difficulty and accuracy in a combined repeated-measures ANOVA, $F(4,\ 52) = 18.2,\ p < 0.001,\ \eta_p^2 = 0.58$. This analysis replicated the main effect of difficulty, $F(4,\ 52) = 4.7,\ p < 0.01,\ \eta_p^2 = 0.26$, as well as the effect of accuracy, $F(1,\ 13) = 152.7,\ p < 0.001,\ \eta_p^2 = 0.92$, with correct trials being rated as more confident than incorrect trials, $M_{cor} = 6.4$, $M_{err} = 2.9$. Ten participants were excluded from the analysis of error trials, as well as the combined analysis, due to the fact that they did not commit any errors in some of the difficulty conditions.

These findings suggest that the difficulty manipulation used in this decision paradigm affects how confident participants judged their choices, and that this effect followed opposing patterns for correct and error trials. For correct trials, easier conditions are associated with a higher level confidence, presumably because more evidence in favour of the chosen response is available. For errors, on the other hand, easier conditions are associated with lower mean confidence and this is driven largely by increases in the number of detected errors. This makes sense if we consider that most errors in the easier conditions have been caused by premature responding. Such errors are more likely to be detected (Scheffers & Coles, 2000), which means they are more likely to be rated as *certainly wrong*.

2.1.2.2 Comparison of confidence scales

The following section focuses on the effects of the different confidence scales on a range of dependent variables. Given that participants could not predict

which scale would be shown after their response, these analyses focused entirely on second-order performance.

All further analyses in this thesis assume that metacognition is worthwhile studying based on – amongst others – the assumption that those introspective judgements of accuracy indeed covary with objective accuracy. I first tested this assumption. The left panel of Figure 10 shows the error rates for the six levels of confidence reported by participants. Participants show impressive resolution and calibration of confidence: When they reported that they were *certainly wrong* on a given trial, they had high error rates ($M = 93.6\%$), whereas when they said they were *certainly correct* they had almost never committed an error ($M = 2.3\%$). Some participants did not make use of the entire confidence scale, which would have resulted in the need to exclude them from a repeated-measures ANOVA. These data were therefore analysed using a rank correlation. Across participants, confidence correlated significantly with error rate. The estimates of Spearman's rank correlation ranged from $r = -1.00$ to $r = -0.83$, $ps <= 0.04$. A paired-samples t-test was used to compare the error rates for the two confidence categories for the 2-point scale, for which the data can be seen in the right panel of Figure 10. There was a significant difference in error rates between trials that were reported as correct versus trials reported as incorrect, $t(23) = 12.6,\ p < 0.001$. Both scales were then analysed together, that is for the 6-point scale, the first three levels were collapsed into a *wrong* category and the second three levels of confidence into a *correct* category. Scale was then submitted as a second factor into a repeated-measures ANOVA. This analysis replicated the effect of subjective rating with trials rated as errors having higher error rates, $F(1,\ 23) = 222.7,\ p < 0.001,\ \eta_p^2 = 0.91$. There was no effect of scale, however, $F < 1$, and no interaction between the two factors, $F < 1$.

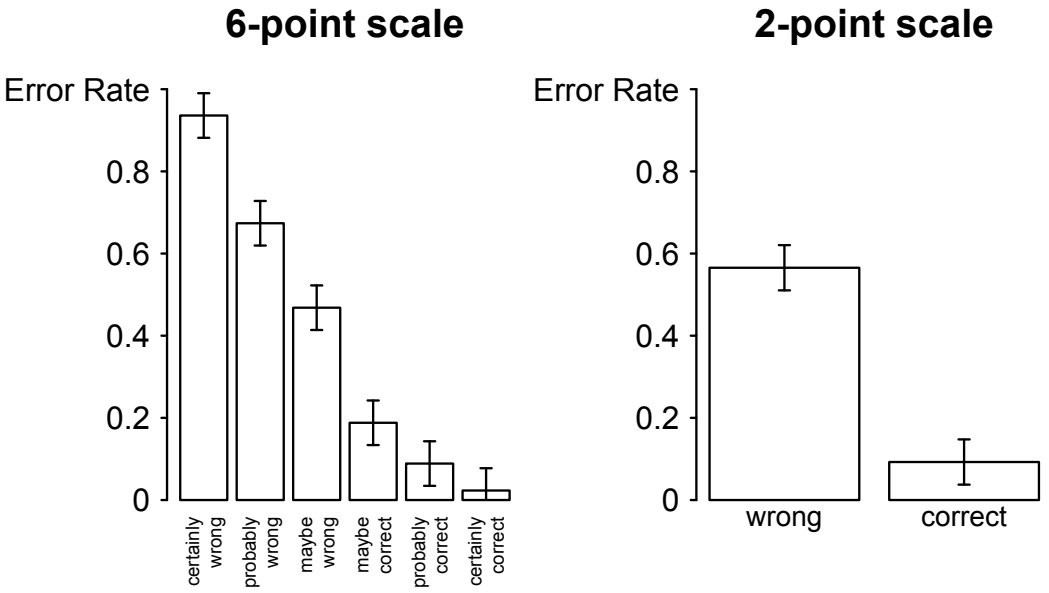

Figure 10: Error rates as a function of subjective confidence ratings for the 6-point scale (left panel) versus the 2-point scale (right panel).

These analyses indicate that this paradigm is a suitable tool to measure metacognition: Confidence was diagnostic of task performance for both the graded scale as well as the binary scale. However, any possible differences that might exist between the two scales could not be detected with this analysis and will be analysed in the course of this chapter with more detailed analyses.

Differences in error rates for different levels of confidence are reflective of changes in confidence distributions for correct and error trials. A related question is therefore whether the scale changes these distribution of confidence judgements. Figure 11 shows confidence distributions for the different scales, levels of difficulty and objective accuracy. From visual inspection, it is apparent that the distributions for the correct versus error trials overlap more for the difficult conditions (lower dot difference) compared to the easy conditions (higher dot difference). This seems to be the case for both scales. To compare the different scales statistically, data from the 6-point scale had to be aggregated with the first three categories forming the category *wrong* and the last

three the category *correct*. These collapsed data are also presented in Figure 11. These data were then submitted to a repeated-measures ANOVA with objective accuracy, confidence scale and level of difficulty as factors. Proportions of trials classified as correct versus incorrect are of course complementary. The statistical tests were therefore just calculated over the trials classified as correct. Whether the current trial was objectively correct or incorrect had a significant effect on whether the trial was classified as a correct response or an error, $F(1,\ 23) = 354.2,\ p < 0.001,\ \eta_p^2 = 0.94$. There was also a significant effect of difficulty, that is more trials were classified as errors in the more difficult conditions, $F(1,\ 23) = 39.7,\ p < 0.001,\ \eta_p^2 = 0.63$. Moreover, there was a significant interaction of objective accuracy with difficulty, that is people discriminated between correct and error trials more clearly in the easier conditions, $F(1,\ 23) = 179.4,\ p < 0.001,\ \eta_p^2 = 0.89$. However, confidence scale had no significant effect on metacognitive rating, $F < 1$. There was no interaction of objective accuracy with confidence scale, $F < 1$, or difficulty with confidence scale, $F(1,\ 23) = 1.3,\ p = 0.27,\ \eta_p^2 = 0.05$, and also no three-way interaction, $F < 1$. For this analysis, the data were collapsed to form two difficulty conditions: an *easy* (2, 4, 6, 8, and 10 dots difference) and a *difficult* (12, 14, 16, 18, and 20 dots difference) condition, to avoid missing data in the cells. Taken together, these results suggest that while difficulty shifted the distribution of confidence levels reported by the participants, there was no such effect of the scale with which this judgement of confidence was reported.

The previous analysis focused on the degree to which participants' judgements distinguished correct and error trials (resolution). SDT measures capture this ability to distinguish events coming from different classes – in this case correct and error trials – in the sensitivity parameter. In the present section, I focus on metacognitive sensitivity together with metacognitive bias;

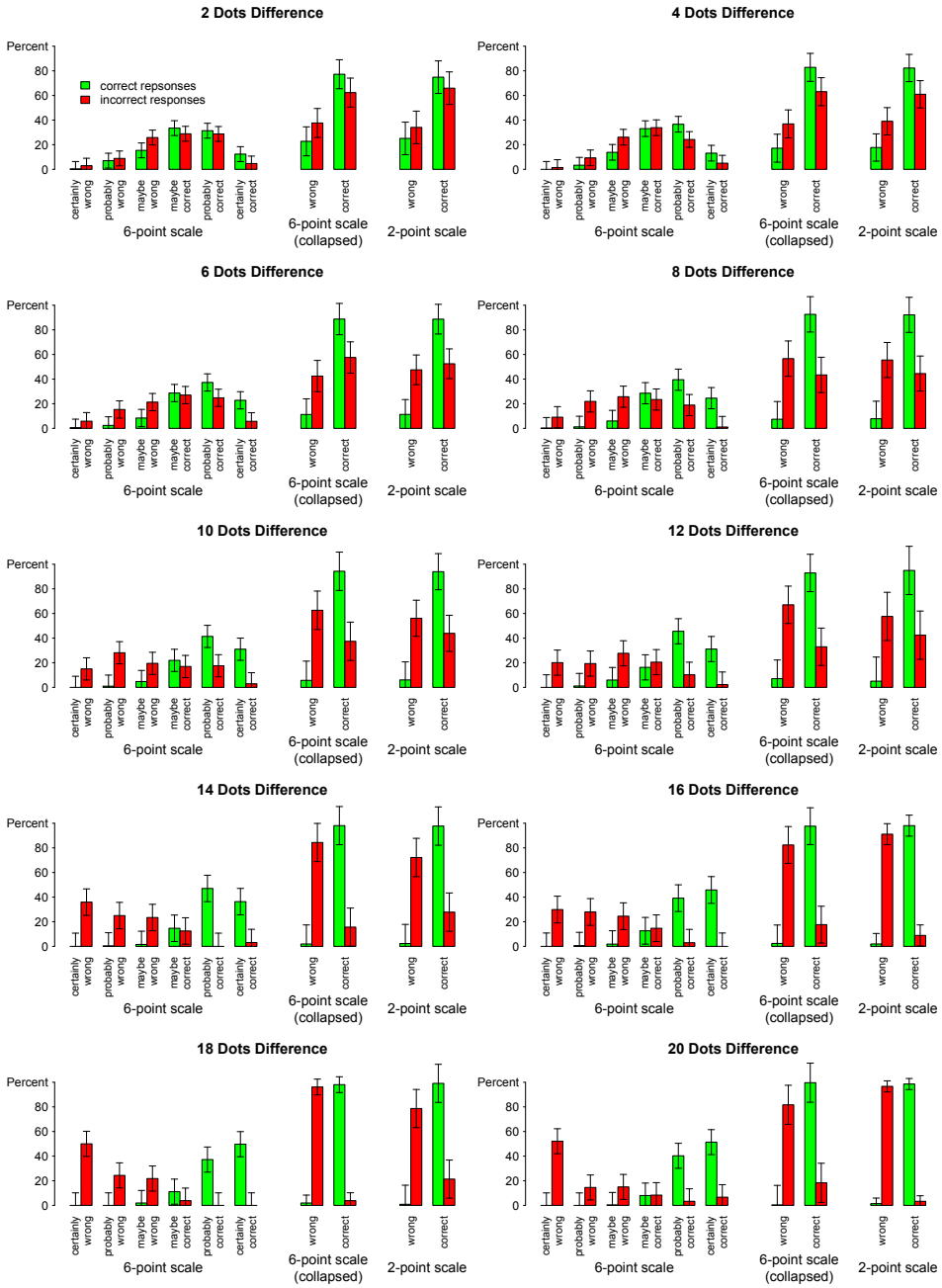

Figure 11: Distributions of confidence responses as a function of difficulty, confidence scale, and objective accuracy. Error bars represent within-subject confidence intervals for the confidence factor within each difficulty and scale condition. The different panels represent difficulty condition, starting at the most difficult condition (2 dots difference) in the top-most, left panel. Within each panel, data from the 6-point scale is shown on the left and data from the 2-point scale is shown on the right. In-between, collapsed data from the 6-point scale is shown, which was used to compare the scales directly.

that is, how participants mapped their responses to the confidence scale. The key question here is whether the 6-options scale led to reduced metacognitive sensitivity and overall low confidence, due to participants' confusion over the different confidence categories. A second question is whether there is an effect of difficulty on metacognitive insight. As described in the general introduction (Section 1.7), different approaches exist to estimate type-II SDT parameters, from which I chose to report metacognitive efficiency, *M-ratio* (Maniscalco & Lau, 2012), and metacognitive bias, B_{ROC} (Kornbrot, 2006) for reasons discussed above. I first tested whether participants' metacognitive efficiency differed for the two scales. The upper panel of Figure 12 shows *meta-d′*, that is metacognitive sensitivity, as a function of difficulty and scale. There was a significant effect of difficulty, $F(9,\ 207) = 41.0,\ p < 0.001,\ \eta_p^2 = 0.64$, but no effect of confidence scale, $F < 1$, and no interaction, $F < 1$. But does this really mean participants have better metacognitive insight in the easier conditions? Given a metacognitively optimal observer – that is, an observer who uses all available evidence from the first-order decision when making the second-order choice – *meta-d′* reflects the level of evidence that was available from the first order decision to arrive at the judgements of the second-order choice – *d′*. They therefore had to be compared to *d′*, which is shown in the middle panel of Figure 12. This measure also scales with difficulty, $F(9,\ 207) = 127.6,\ p < 0.001,\ \eta_p^2 = 0.85$. There was a difference between the two scales, $F(1,\ 23) = 5.6,\ p = 0.03,\ \eta_p^2 = 0.20$, however, this can only be assumed to be a type-I error, given that participants did not know which scale would be shown while completing the first-order decision as the trial order was pseudo-randomised. There was no interaction between the two factors, $F(9,\ 207) = 1.1,\ p = 0.40,\ \eta_p^2 = 0.04$.

Both measures can then be compared as a ratio, the logarithm of which

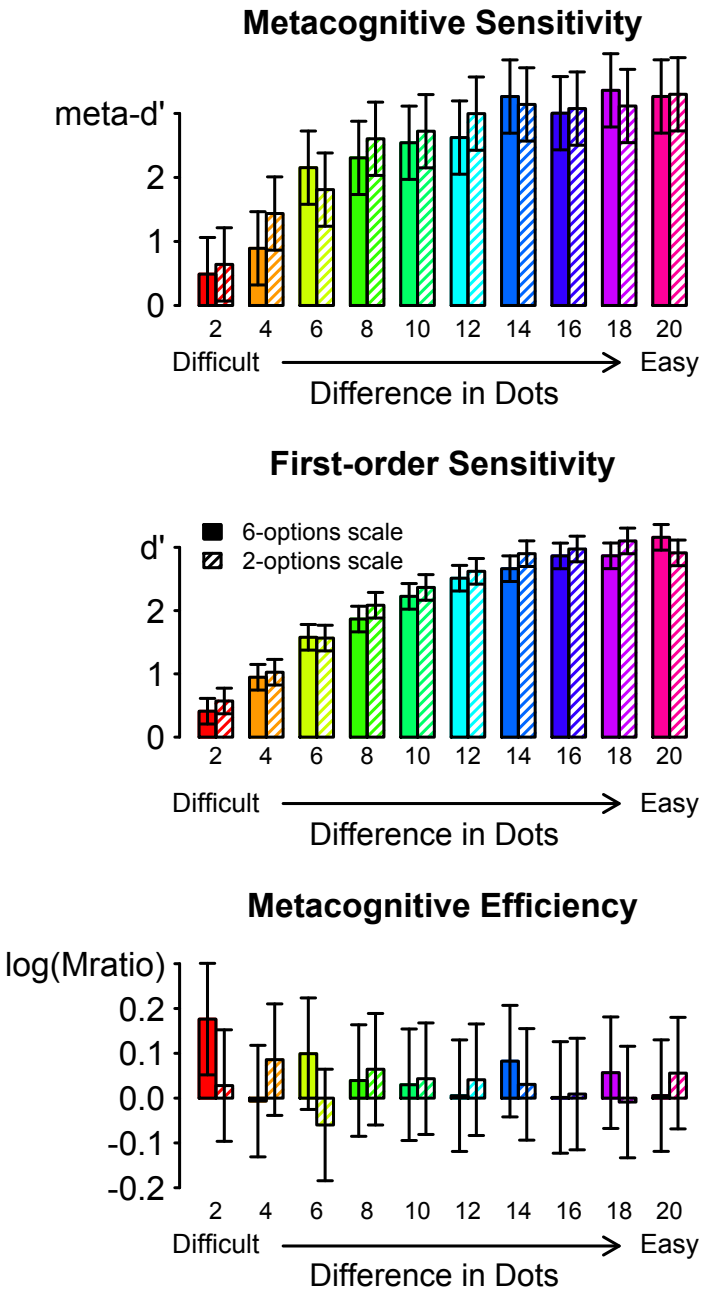

Figure 12: Metacognitive sensitivity (*meta-d′*; top panel), first-order sensitivity (*d′*; middle panel), and the common logarithm of metacognitive efficiency (*log*(*M-ratio*); bottom panel) as a function of difficulty and confidence scale. Both *meta-d′* and *d′* are coded in the same units. Values of *log*(*M-ratio*) larger than 0 can be interpreted as super-optimal metacognitive behaviour.

is shown in the bottom panel of Figure 12. Importantly, most averages are greater than 0, meaning that participants had more evidence available during their second-order judgement than during their first-order judgement, which can be interpreted as evidence for post-decision processing. For the metacognitive efficiency, there was no effect of difficulty, $F < 1$, as predicted. Also, no effect of scale was found, $F < 1$, and also no interaction between the two factors, $F(2.8,\ 59.7) = 1.0,\ p = 0.39,\ \eta_p^2 = 0.05$. We can therefore neither conclude that people have better metacognitive ability *per se* in easier conditions, nor that their metacognitive ability is significantly better when recorded on a binary scale. These findings furthermore suggest that the previously discussed effect of difficulty on confidence distributions was simply a reflection of differences in first-order performance and disappears as soon as this is taken into account.

One advantage associated with using the 6-options scale is that it is possible to construct ROC curves with it. Such curves are shown in Figure 13, again for five levels of difficulty. Difficulty had an effect on the area under the ROC curve (AUC), $F(2.2,\ 34.6) = 22.3,\ p < 0.001,\ \eta_p^2 = 0.58$. This parameter is often used as a non-parametric measure of metacognitive accuracy with 0.5, that is the diagonal, representing no metacognitive insight. Seven participants had to be excluded from this analysis due to an error rate of zero in the easy conditions.

These ROC curves can then be used to estimate SDT parameters, using the non-parametric approach by Kornbrot (2006) with A_{ROC} as a measure of metacognitive sensitivity and B_{ROC} as a measure of metacognitive bias. I have already discussed metacognitive sensitivity and efficiency as reflected in the measures $meta\text{-}d'$ and $log(M\text{-}ratio)$, respectively. I therefore do not present analyses of A_{ROC}, as this measure fails to take into account differences in

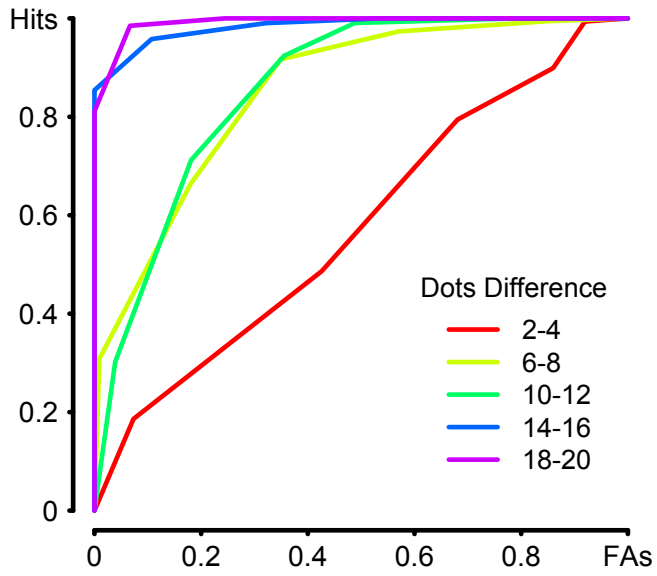

Figure 13: ROC (receiver operating characteristic) curves for the five aggregated difficulty conditions. Hits: hit rate. FAs: false alarm rate.

first-order performance, as previously discussed. Metacognitive bias according to B_{ROC}, however, is arguably the best approach to measure participants' tendency towards relative over- (high B_{ROC}) or underconfidence (low B_{ROC}). Given that this approach relies on ROC curves, we can again only use data from the 6-point scale. From visual inspection of the data presented in Figure 14, it seems like B_{ROC} is higher for the easier conditions, in other words, participants are more overconfident when the decision is easy compared to when it is difficult. However, there was no such main effect of difficulty for metacognitive bias, B_{ROC}, $F(2.1,\ 39.5) = 1.0,\ p = 0.38,\ \eta_p^2 = 0.05$. The linear trend was not reliable either, $F < 1$. For this analysis, two steps were taken to limit the number of excluded participants to four: First, to avoid division by zero, padding with arbitrarily small number (0.0001) was used. Second, difficulty was once more averaged over 5 conditions.

Taken together, these results suggest that both the 2- and the 6-point confidence scale are suitable tools for measuring metacognition. The higher

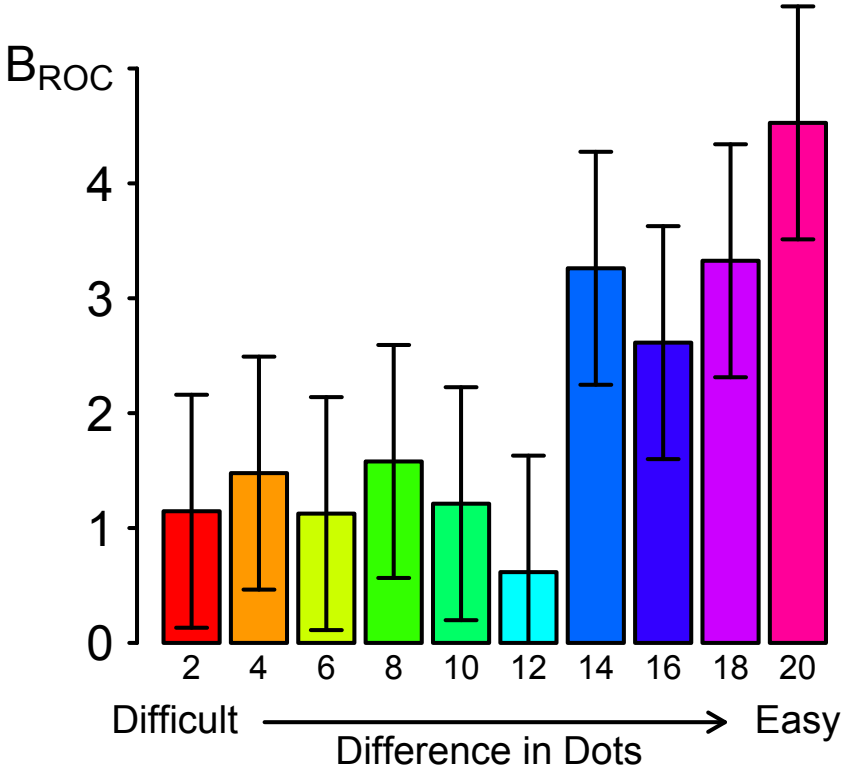

Figure 14: Non-parametric metacognitive bias, B_{ROC}, as a function of difficulty and confidence scale. Larger values relative to lower ones suggest (relative) overconfidence.

measurement resolution gained by using a 6-point scale as opposed to a binary scale therefore does not come at a cost of reduced metacognitive sensitivity. However, the 6-point scale has the advantage that ROC curves can be constructed with it, and that non-parametric SDT measures can be constructed from these ROC curves, including a bias measure that estimates whether participants were over- or underconfident in their use of the confidence scale.

One final comparison for the two scales focuses on how quickly people rated their confidence. Hick's law states that when facing a larger number of choice alternatives, participants take longer to decide (Hick, 1952). So even if the more fine-grained 6-point scale does not lead to decreased metacognitive efficiency, participants might still need more time to judge their confidence on such a scale. This could arguably be problematic for paradigms in which the overall duration of a trial is crucial. Confidence RTs for both scales are presented in Figure 15. They were first analysed separately before they were compared directly using a repeated-measures ANOVA.

Not every participant made use of the entire scale, which means there was missing data for some participants, especially for the *certainly wrong* category in case of correct trials as well as the *certainly correct* category in case of error trials. For the 6-point scale (shown in the left panel of Figure 15) the data were therefore aggregated over levels of confidence, so that two levels were formed: The first three levels form the category *wrong* and the last three levels form the category *correct*. This also allowed direct comparison with the 2-point confidence scale. Data from both confidence scales were then submitted to a repeated-measures ANOVA to test the effect of scale, confidence, and objective accuracy on confidence RTs. Critically, there was no difference between the two scales, $F < 1$. People were, if anything, a few milliseconds slower in responding to the 2-point scale, which goes directly against what we would

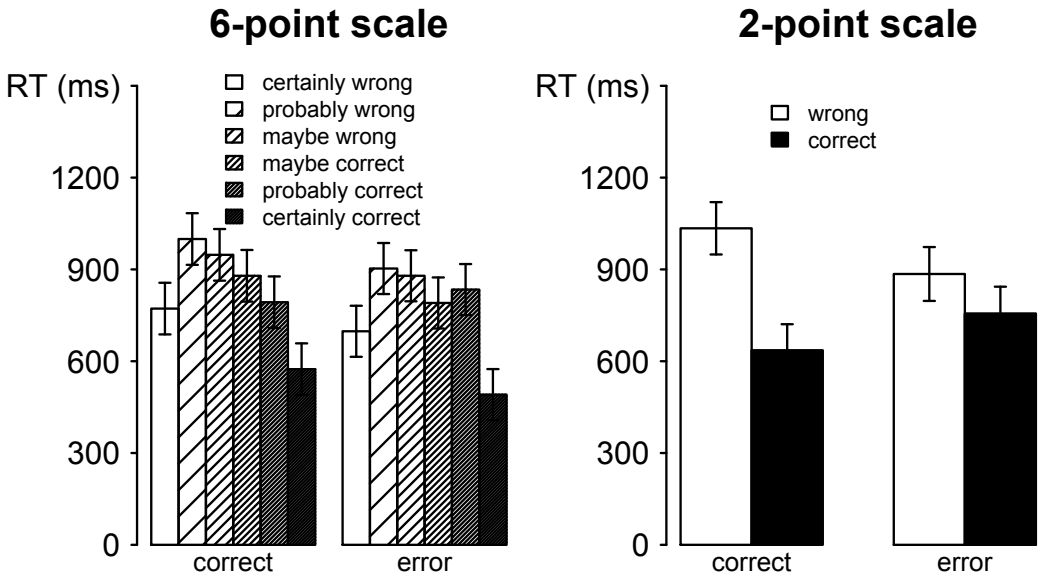

Figure 15: Confidence RTs as a function of objective accuracy and subjectively-rated level of confidence, presented for the 6-point scale (left panel) and the 2-point scale (right panel); ms: millisecond.

expect according to Hick's law, $M_6 = 825\ ms$; $M_2 = 828\ ms$. Confidence, $F(1,\ 23) = 29.4,\ p < 0.001,\ \eta_p^2 = 0.56$ had a significant effect on confidence RTs. This means that participants judged their trials classified as correct faster than trials classified as errors, $M_{sigcor} = 722\ ms$; $M_{sigerr} = 931\ ms$. There was no reliable effect of objective accuracy on confidence RTs, $F < 1$. This analysis furthermore revealed a reliable interaction between confidence and accuracy, $F(1,\ 23) = 18.3,\ p < 0.001,\ \eta_p^2 = 0.44$. This reflects that for correct trials, the just described effect of confidence on confidence RT was larger, $M_{cor} = 319\ ms$, compared to error trials, $M_{err} = 99\ ms$. Objective accuracy and scale did not interact, $F < 1$, and there was also no three-way interaction, $F(1,\ 23) = 2.2,\ p = 0.15,\ \eta_p^2 = 0.09$. There was, however, a significant interaction between confidence and confidence scale, $F(1,\ 23) = 12.4,\ p < 0.01,\ \eta_p^2 = 0.35$. This effect was caused by the fact that the difference in confidence RTs between trials classified as correct versus trials

classified as incorrect was larger in the 2-point scale, compared to the 6-point scale, $M_6 = 154\ ms$; $M_2 = 264\ ms$.

Taken together, these results suggest that participants do not take more time when judging their confidence on a 6- as opposed to a 2-point scale. One possible interpretation would be that this absence of scale effects is due to the fact that participants are able to prepare their confidence rating in advance, that is while judging the primary stimulus. However, the fact that confidence RT is sensitive to the influence of other factors, such as the effects of confidence on confidence RT, speaks against this interpretation. Once more, these findings suggest that using a more fine-grained scale such as the 6-point scale does not impose additional processing costs on the participant.

2.1.2.3 Practice effects

It is useful to know how much practice participants required before a confidence experiment. The present section therefore focuses on the question of how confidence changes over blocks, that is whether participants gain better insight into their responses as the experiment progresses and whether they also get increasingly more confident (i.e., overconfident). Figure 16 shows how average confidence developed over blocks, separately for correct and error trials. The blocks start at block 2, which was the first one on which confidence judgements were recorded. This analysis revealed a significant effect of objective accuracy, $F(1,\ 22) = 181.2,\ p < 0.001,\ \eta_p^2 = 0.89$, with higher judgements for correct compared to error trials, $M_{cor} = 5.7$ versus $M_{err} = 3.6$. Critically, however, there was no effect of block, $F < 1$, and also no interaction, $F < 1$. However, if tested separately for correct and error trials, there was indeed a reliable effect of block for correct trials, $F(2.6,\ 56.2) = 3.3,\ p = 0.03,\ \eta_p^2 = 0.13$, with early blocks (blocks 2 to 8) associated with lower confidence ratings than late

blocks (blocks 9 to 16), $M_{early} = 3.6$ versus $M_{late} = 5.8$. Such an effect was not present on error trials, $F < 1$. From all these analyses, one participant had to be excluded because of not committing any errors in one of the conditions.

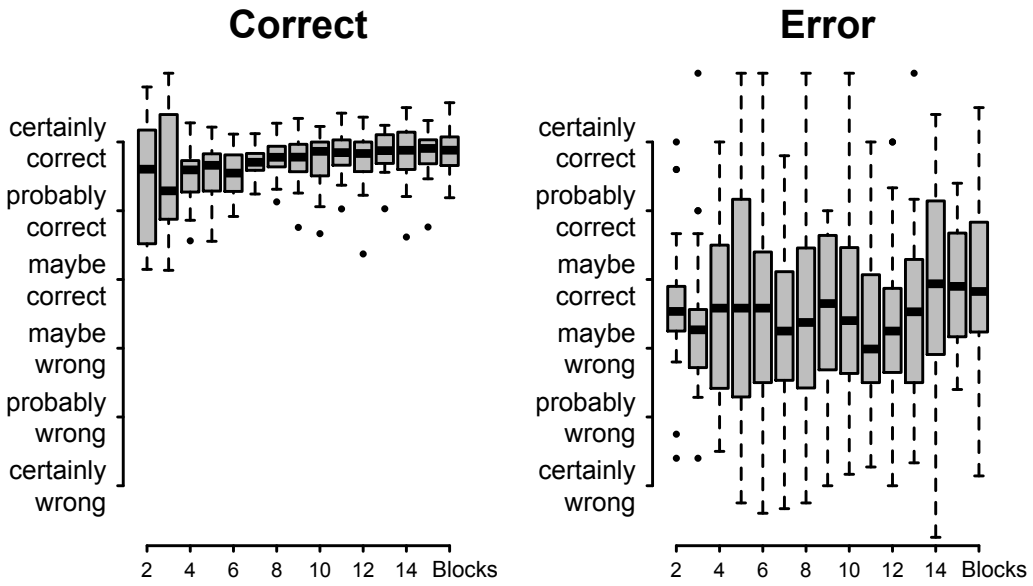

Figure 16: Development of average confidence over blocks; 50% of the data lie within the box, the thick black horizontal line within each box indicates the median, the whiskers extend to 1.5 times the interquartile range (IQR); outliers are displayed as dots. The first block in which confidence responses were collected was block 2; left panel: correct trials; right panel: error trials.

The fact that confidence did not vary over blocks in general but just for correct trials speaks for the fact that if any practice effects exist, they should be reflected by changes in metacognitive efficiency, that is metacognitive sensitivity in relation to first-order performance, and not bias. The data was therefore analysed using an SDT approach.

The first analysis focused on metacognitive efficiency. As shown in the upper panel of Figure 17, *meta-d′* indeed varied reliably over blocks, $F(14,\ 322) = 1.8,\ p = 0.04,\ \eta_p^2 = 0.07$, with lower metacognitive sensitivity for the earlier blocks (blocks 2 to 8), compared to the later ones (blocks 9 to 16), $M_{early} = 2.24$ versus $M_{late} = 2.57$. There was also an effect of d' (middle

panel of Figure 17), $F(14,\ 322) = 1.8,\ p = 0.04,\ \eta_p^2 = 0.07$, however, in the opposite direction, that is participants were getting less sensitive over blocks, $M_{early} = 2.07$ versus $M_{late} = 1.92$. With type-II performance getting better over time, and type-I performance getting worse, one can expect an increase in metacognitive efficiency over blocks. Metacognitive efficiency, that is the common logarithm of the *M-ratio* parameter and therefore *meta-d′* divided by d', is represented in the lower panel of Figure 17. Two participants had to be excluded from the analysis on $log(M\text{-}ratio)$ due to the fact that the negative values of *M-ratio* were estimated for these participants and the common logarithm of a negative number is undefined. There was indeed a reliable effect of block, $F(14,\ 294) = 2.6,\ p = 0.001,\ \eta_p^2 = 0.11$, with $log(M\text{-}ratio)$ lower for the earlier blocks (blocks 2 to 8), compared to the later ones (blocks 9 to 16), $M_{early} = 0.01$ versus $M_{late} = 0.12$. Moreover, there was a reliable linear trend, $F(1,\ 21) = 11.9,\ p < 0.01,\ \eta_p^2 = 0.36$. Taken together, these findings suggest that metacognitive efficiency increased over blocks.

One possible mechanism, which might have led to increasing metacognitive efficiency over blocks, would be the above-reported increases in correct-trial confidence as opposed to error trials. However, one alternative mechanism would be that participants increased their response speed over blocks. Speeding leads increases in error rate, that is decreases in d'. A large proportion of these errors can be assumed to be premature-response errors, which are easy to detect and therefore result in higher resolution (Pleskac & Busemeyer, 2010), which in turn is equivalent to increases in metacognitive efficiency. Figure 18 shows how RTs decreased over blocks. This effect was reliable, $F(3.6,\ 74.7) = 25.9,\ p < 0.001,\ \eta_p^2 = 0.55$, as was the linear trend, $F(1,\ 21) = 68.2,\ p < 0.001,\ \eta_p^2 = 0.76$. This matches the finding that d' decreased over blocks (see middle panel of Figure 17). These analyses excluded

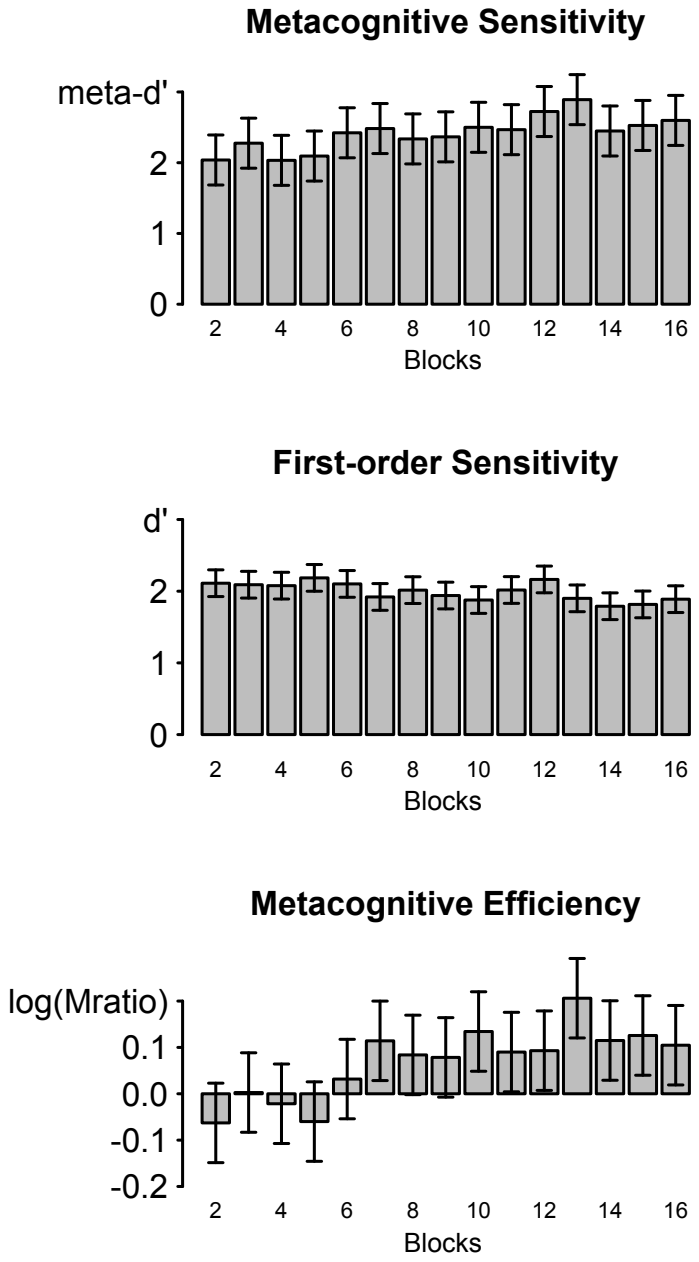

Figure 17: Metacognitive sensitivity (*meta-d′*; top panel), first-order sensitivity (*d′*; middle panel), and the common logarithm of metacognitive efficiency (*log*(*M-ratio*); bottom panel) as a function of block, beginning at the second block of the experiment as the first block did not contain confidence responses. Both *meta-d′* and *d′* are coded in the same units. Values of *log*(*M-ratio*) larger than 0 can be interpreted as super-optimal metacognitive behaviour.

the same two participants as above to match the analysis of the $log(M_{ratio})$ parameter. The first block was excluded for the same reasons. Taken together, the source of the effect on metacognitive efficiency cannot be clearly identified, but changes in response strategy constitute a likely explanation.

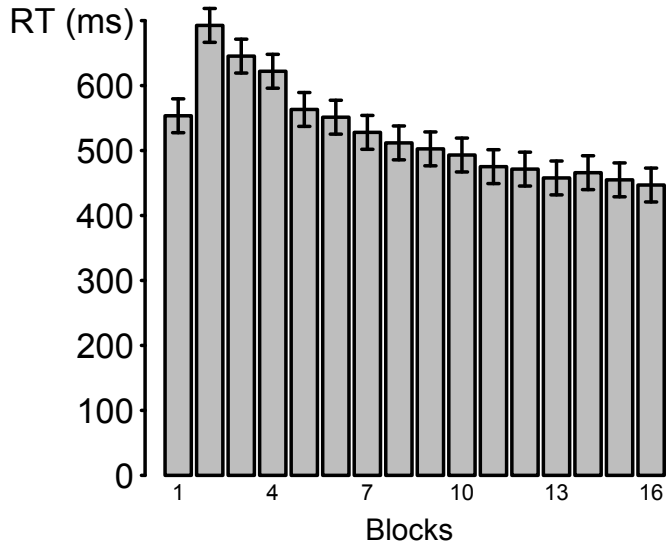

Figure 18: Block-wise RTs for correct and error trials combined.

To sum up, these findings indicate that participants' confidence ratings changed over blocks and that this change affected participants' metacognitive efficiency. The improvement in metacognitive performance over blocks was presumably at least partly caused a first-order speed-accuracy tradeoff, meaning that participants opted for a more speed-focused strategy the longer the experiment progressed.

2.1.3 Discussion

The present experiment tested the suitability of a perceptual decision-making task for the purpose of studying decision confidence, as well as directly comparing two confidence scales. First, the findings suggest that both the binary,

as well as the 6-point scale are sensitive tools for studying how confidence judgements covary with actual task performance. Second, the results suggest that the higher resolution gained by using a 6-point scale as opposed to a binary scale does not come at a cost of reduced metacognitive sensitivity, as reflected in analyses on actual error rates for different levels of confidence, confidence distributions, and metacognitive efficiency and bias. Moreover, rating one's confidence on a scale with a higher resolution was not accompanied by increased confidence RTs.

In the perceptual decision-making paradigm chosen in this experiment, difficulty was varied in ten steps by changing the number of dots in the two fields of which the one with more dots had to be chosen. This difficulty manipulation had a reliable effect on both on correct RTs, and on error rates. Moreover, the easier a trial, the larger the difference between correct-trial and error-trial confidence. This increased metacognitive insight for easier conditions was caused, however, by the fact that participants had a chance to accumulate more evidence during the first-order decision – evidence that could consequently be used in forming a metacognitive judgement. Moreover, difficulty also did not affect participants' tendency towards being more over- or underconfident, as reflected in the analysis on metacognitive bias. These results indicate that the decision paradigm used in this chapter is a suitable framework when measuring metacognitive judgements with precise experimental control over task difficulty, affecting only first- and not second-order performance. This is advantageous because it means that metacognitive performance should remain stable independent of how difficult the primary task is. Interestingly, these findings stand in contrast to the often-replicated *hard-easy effect*, which describes the finding that easier conditions are often accompanied by underconfidence while participants often exhibit overconfidence in the most difficult

conditions (Pleskac & Busemeyer, 2010; Gigerenzer, Hoffrage & Kleinbölting, 1991; Juslin, Winman & Olsson, 2000; Baranski & Petrusic, 1994). Such an effect does not seem to be present for the paradigm used in this experiment.

These findings regarding metacognitive bias highlight another advantage of using non-binary scales: The non-parametric SDT method (Kornbrot, 2006) chosen to estimate these bias parameters, B_{ROC}, is based upon ROC curves. Such curves cannot be constructed with a binary scale and metacognitive bias parameters can therefore not be estimated with such a scale. The present manipulation of difficulty did not lead to an effect on bias, as discussed above, but other experiments in this thesis will focus on these over- and underconfidence effects, for example Experiment 6.

One final goal of the present study was to assess whether participants get better at judging their confidence over trials. The results suggest that participants get indeed more confident in their correct decisions over blocks (but not their incorrect decisions), but this was not reflected in an increase in metacognitive efficiency, presumably caused by the fact that participants' RTs decreased over the course of the experiment, resulting in more premature responses and therefore more easy-to-detect errors. Even if this effect is explained by first-order performance, it needs to be taken into account because it might introduce unnecessary noise into the metacognition data. There is not much that can be done to avoid this overall strategy change, but we can conclude that changes in participants' response strategy have to be monitored over the course of the experiment and that re-instructing participants to keep sufficient balance between speed and accuracy might become necessary in individual cases. Moreover, these block effects suggest that it is necessary to include at least one practice block in which participants get adjusted to judging their confidence and thus settling into their speeded response strategy of the

task more readily.

All these results seem to point towards the finding that metacognitive judgements are relatively robust and that the higher resolution for the 6-point scale comes at no cost that might reflect on the key dependent variables, such as metacognitive sensitivity, metacognitive bias, or confidence RTs.

2.2 Experiment 2: Influence of timing of confidence judgements

With the second experiment in this chapter, I manipulated the time at which participants are asked to rate their confidence, that is the RSI between the primary choice and the onset of the confidence scale. Participants were not allowed to rate their confidence prior to the onset of the confidence scale and they were furthermore instructed to respond as quickly as possible upon onset of the confidence scale. Therefore, we can assume that manipulating RSI affects the time participants have to judge their confidence.

The results of the present experiment have practical significance in terms of whether confidence measures are sensitive to an RSI manipulation. This point is by no means trivial – if RSI did not affect confidence then experimenters could opt for an immediate onset of the confidence scale immediately after participants made their primary response to shorten the duration of a trial and therefore decrease the length of the entire experiment. On the other hand, we could instead find a time window for which metacognitive insight is maximal. For instance, in the metamemory literature a similar effect has been found judgements of learning, as briefly mentioned in the general introduction (Section 1.5): Judgements of learning require participants to judge the percentage of correctly remembered learning material – such as word pairs – in an

upcoming test after having previously studied them. These judgements can be made either immediately after having learned the pair association (immediate judgement of learning) or a few seconds or minutes later (delayed judgement of learning). Delayed judgements of learning have higher resolution compared to immediate judgements of learning (Nelson & Dunlosky, 1991).

The question of whether confidence is sensitive to an RSI manipulation also bears on an important theoretical distinction, that is the difference between decisional and post-decisional locus models (Yeung & Summerfield, 2012, 2014), as previously discussed in Section 1.5. Decisional locus models assume that information that was accumulated up to when the decision was formed then also forms the basis for the confidence judgement (e.g., Kiani & Shadlen, 2009; Vickers & Packer, 1982; Gherman & Philiastides, 2014). A consequence that follows from this model is that confidence judgements made immediately after the primary decision might be more accurate compared to slightly delayed confidence judgements due to memory decay: Metacognition can be regarded as an efference copy, or "corollary discharge" (Middlebrooks & Sommer, 2011, p. 11). This means that the memory trace of the just-performed action could presumably be forgotten over time. Such forgetting could be caused by signal decay, as assumed in many prominent theories of short-term memory (Barrouillet, Bernardin & Camos, 2004; Burgess & Hitch, 2006), or because it becomes more difficult to retrieve a memory trace due to increased interference with progressing time (Brown, Neath & Chater, 2007). Whether confidence is assumed to be based on a balance-of-evidence mechanism or another internal cue such as decision time – this detrimental effect on metacognitive accuracy is a natural prediction of decisional-locus models.

Post-decisional locus models, on the other hand, assume that even after the primary decision was formed, information accumulation continues with

the same mechanisms as the primary decision until a confidence judgement is made (e.g., Pleskac & Busemeyer, 2010; Van Zandt & Maldonado-Molina, 2004). Consequently, more time to accumulate post-decision evidence (that is, longer RSIs) should lead to higher metacognitive accuracy because participants have an ultimately larger number of samples on which they can base their confidence rating. More evidence samples means that the overall noise in this rating is reduced. In extreme cases, such additional evidence might even lead to reversals of the initial choice, or changes of mind. For instance, the model proposed by Van Zandt and Maldonado-Molina (2004) assumes two counters representing the choice options. The counter that reaches the decision threshold first determines the choice, however, evidence accumulation continues towards a second threshold representing the confidence rating. If the losing counter accumulates a sufficient amount of evidence during this post-decision processing stage, it can reach the second threshold first. In such a case, the participant would signal a change of mind, that is *low confidence*.

In summary, in the present experiment, I manipulated RSI to test whether such a manipulation had an effect on metacognitive processing. If such an effect was found, then it would have important theoretical implications regarding the question of whether confidence is formed at a decisional or a post-decisional locus: If participants were found to have highest metacognitive insight at shorter RSIs then this could be interpreted as preliminary evidence in favour of a decisional-locus model. If, on the other hand, they show higher metacognitive accuracy for longer RSIs then that is most congruent with a post-decision processing locus. In this latter case, increasing the time between the response and confidence judgement should increase the metacognitive accuracy because that the brain had more time to accumulate additional information, potentially overcoming its own errors.

I chose two difficulty conditions from the previous experiment. The reason for including two levels of difficulty stems from the idea that – according to the post-decision processing model – the same decision processes are involved in forming the confidence judgement. This would have the effect that for more difficult trials, trials with a lower drift rate, confidence judgements take more time to develop. This is also assumed in the evidence accumulation model by Audley (1960). If confidence RTs are held constant, as was the aim in the present paradigm, then metacognitive accuracy should be even more impaired at shorter RSIs for the difficult condition. This hypothesis therefore predicts an interaction between difficulty and RSI.

2.2.1 Methods

2.2.1.1 Participants

Twenty participants were tested, who were between 18 and 23 years old (M = 18.6). All participants were undergraduate students receiving course credit for the 40 minute duration of the experiment, including instruction and debriefing. All procedures were approved by the local ethics committee, and all participants gave informed consent prior to the start of the experiment. Fifteen participants were female, one participant was left-handed. All had normal or corrected-to-normal vision.

2.2.1.2 Task and procedure

The paradigm used in this experiment was very similar to the one described in Experiment 1. I will therefore only report where it deviated from the methods described in Section 2.1.1.2. In the present experiment, only two difficulty conditions were used, the 6- and 10-dots difference conditions. The

findings of Experiment 1 suggest that these conditions lie closest to 15 and 25% errors, therefore resulting in a more easy and a more difficulty condition. Furthermore, only the 6-options confidence scale was used.

The experiment began with one block of dot-task practice. This was the only block with feedback tones and longer ITIs following mistakes, as in Experiment 1. The second block introduced participants to the confidence scale, which appeared at a constant 600 ms post-response. Participants received written instruction that there would not be a time limit for this judgement but that they were encouraged to respond according to their first impression. The third, and last practice block introduced the five RSI durations of 100, 300, 600, 1000, 1500 ms, also letting participants know that from now on the confidence scale would appear on screen at varying times after they made a response. The sequence of such a trial is shown in Figure 19. Participants were instructed to give their answer "as soon as the scale is presented on screen". They were furthermore instructed that they would not be able to rate their confidence before the scale appeared on screen. The order of the RSIs was randomised over trials. All three practice-blocks were 40 trials long. Following these practice blocks, they then completed ten experimental blocks, similar to the third practice block but 60 trials long.

As in the previous experiment, participants were given performance feedback at the end of each block, that is their average correct RTs and error rates. If this error rate was lower than 10%, a sentence was presented underneath them: "Please keep in mind to react as quickly as possible!" If, on the other hand, their error rate was higher than 25%, this sentence read "Please try to be more accurate!" instead.

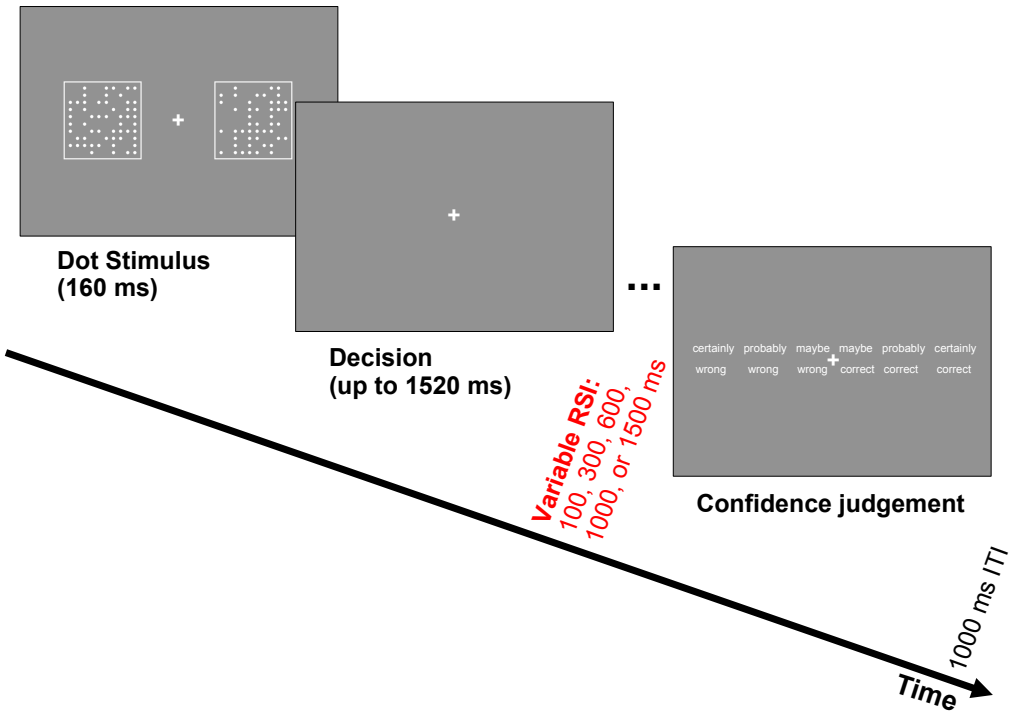

Figure 19: Methods of the dot task. Participants had to say which of two fields contained more dots by pressing the left or right key. A 6-point verbal confidence scale appeared on screen after a variable duration. RSI: response-stimulus interval; ITI: inter-trial interval; ms: millisecond.

2.2.2 Results

2.2.2.1 Replications

As in the previously reported experiment, there was a significant effect of difficulty on correct RTs, $F(1,\ 42) = 10.3,\ p < 0.01,\ \eta_p^2 = 0.20$, with the more difficult condition eliciting slower RTs compared to the easier one, $M_6 = 490\ ms$ versus $M_{10} = 472\ ms$. A between-experiment ANOVA, using experiment as a between-subject factor and difficulty as a within-subject factor revealed that – compared to the results of the previous experiment – correct RTs are similar but numerically faster, $M_6 = 508\ ms$ versus $M_{10} = 486\ ms$, $F < 1$. The factors experiment and difficulty did not interact, $F < 1$. Error rates also differed over levels of difficulty, $F(1,\ 42) = 164.3,\ p < 0.001,\ \eta_p^2 = 0.80$, with higher error rates for the more difficult condition, $M_6 = 24.6\%$ versus $M_{10} = 14.6\%$. These results were again very similar to the previously reported study, $M_6 = 23.3\%$ versus $M_{10} = 12.9\%$, $F < 1$. The factors experiment and difficulty again did not interact, $F < 1$.

Figure 20 shows how participants' confidence once again reflected objective performance: When participants reported that they were *certainly wrong*, they indeed had an average error rate of $M = 87.6\%$, whereas for trials classified as *certainly correct*, error rates were very low, $M = 2.7\%$. This effect was reliable, as revealed by significant Spearman's rank correlations for each participant individually, ranking from $r = -1.00$ to $r = -0.89$, $ps <= 0.02$, mirroring the findings from the previous experiment.

Figure 21 shows the distributions of responses as a function of confidence and objective accuracy, both for the difficult (left panel) and the easy (right panel) condition. As in the previous study, the proportions of trials classified as correct versus incorrect are complementary and the statistical

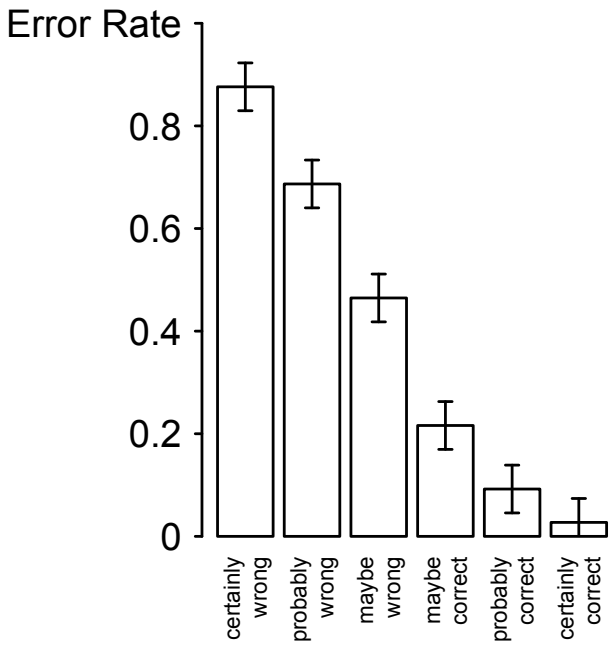

Figure 20: Error rates as a function of subjective confidence ratings.

tests are therefore calculated over proportion of trials classified as correct (aggregated over levels 4 to 6 on the confidence scale). There was a significant effect of objective accuracy on whether or not a trial was classified as correct, $F(1,\ 19) = 278.6,\ p < 0.001,\ \eta_p^2 = 0.94$, with more correct trials being classified as correct compared to error trials, $M_{cor} = 90.9\%$ versus $M_{err} = 48.6\%$. There was also a reliable effect of difficulty on the proportion of trials classified as correct, $F(1,\ 19) = 6.8,\ p = 0.02,\ \eta_p^2 = 0.26$, as well as a significant interaction between difficulty and objective accuracy, $F(1,\ 19) = 70.8,\ p < 0.001,\ \eta_p^2 = 0.79$, that is people discriminated between correct and error responses more clearly in the easier condition. Taken together, all these findings replicated the results from the previous study, leading to the conclusion that the perceptual decision-making paradigm with its dot-difference difficulty manipulation yields stable, replicable results.

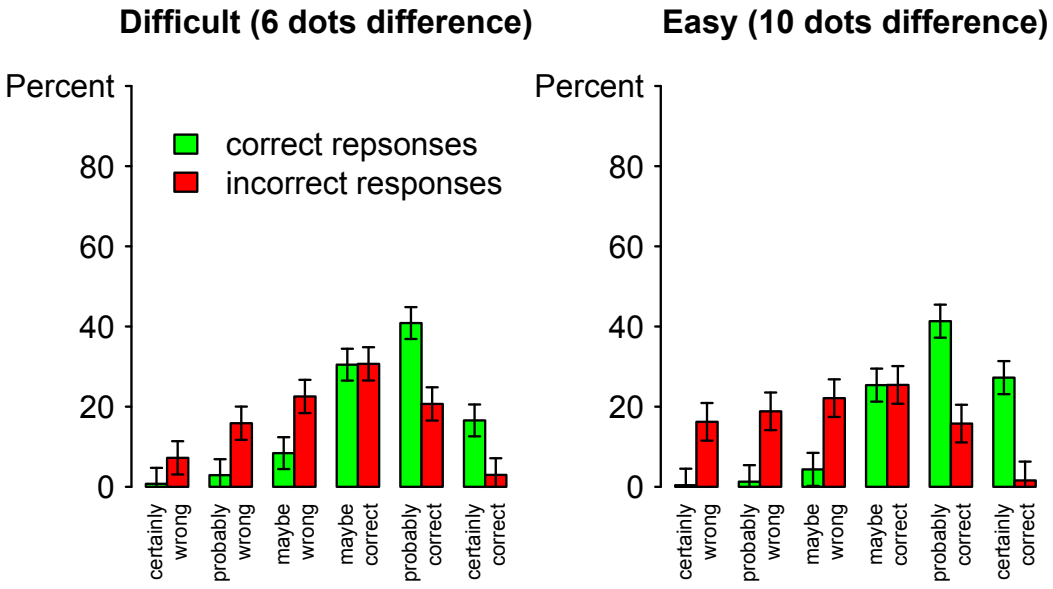

Figure 21: Distributions of confidence responses as a function of difficulty and objective accuracy. Error bars represent within-subject confidence intervals for the confidence factor within each difficulty condition. The left panel presents data from the difficult (6 dots difference) condition and the right panel presents data from the easy (10 dots difference) condition.

2.2.2.2 Effects of RSI

The goal of this next section is to analyse the effect different RSIs had on metacognitive processing. It was first analysed whether participants' responses followed the onset of the confidence scale independent of RSIs. This matters because participants had been instructed to respond immediately following presentation of the scale. Response slowing in the confidence response as a function of RSI might reduce possible effects of RSI.

Figure 22 shows confidence RT as a function of RSI and objective accuracy. The fact that the lines have a negative slope means that participants took more time to give a confidence response on trials with a short period of time between the primary response and the onset of the confidence scale. This effect was statistically reliable, $F(1.5,\ 27.7) = 53.0,\ p < 0.001,\ \eta_p^2 = 0.75,$

together with a reliable linear trend, $F(1,\ 18) = 67.3,\ p < 0.001,\ \eta_p^2 = 0.79$. Moreover, there was also a reliable effect of objective accuracy, $F(1,\ 18) = 16.7,\ p < 0.001,\ \eta_p^2 = 0.48$, reflecting that participants had longer confidence RTs on error trials, $M_{err} = 521\ ms$, compared to correct trials, $M_{cor} = 448\ ms$. The effect of difficulty on confidence RT was not reliable, $F < 1$, replicating Baranski and Petrusic (1998). There was a marginally significant interaction between objective accuracy and difficulty, $F(1,\ 18) = 4.2,\ p = 0.05,\ \eta_p^2 = 0.19$, indicating that the confidence RT difference between correct and error trials was smaller for difficult, $M_6 = 40\ ms$, compared to easy trials, $M_{10} = 108\ ms$. None of the other interactions was found to be reliable, $Fs <= 1.9,\ p >= 0.18$.

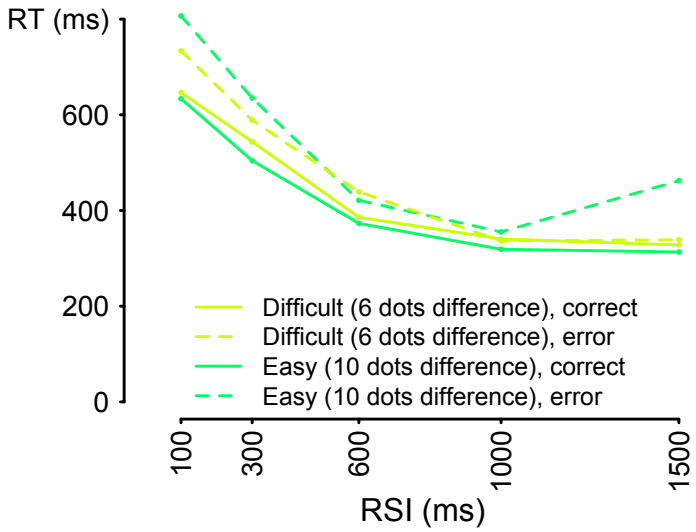

Figure 22: Confidence response times (RTs) as a function of objective accuracy and response-stimulus interval (RSI). No error bars are shown because they are somewhat overlapping and hinder interpretation of the figure; ms: millisecond.

The effect of RSI on confidence RTs could presumably be interpreted as support for post-decision processing – if participants had made up their mind about their confidence rating during the primary decision, there would have been no slowing for shorter RSIs. However, at the same time, this main effect

of RSI poses a problem for the main analyses of this study. In fact, this effect can be thought to decrease the RSI manipulation, given that participants to some extent evened out the RSI manipulation by taking more time when the RSI was short and vice versa. However, if we adjust the RSI levels by their respective confidence RTs, adding the average RT to each of them, this gives us conditions of 805, 867, 1005, 1338, and 1861 ms, which still provides a range of different timings at which confidence responses are given. I will therefore analyse the effects of RSI, keeping in mind that the power of the design might be somewhat reduced.

Figure 23 shows average levels of confidence as a function of RSI, difficulty and objective accuracy. The upper two panels show the difficult condition (6 dots difference) and the lower panels show the easy condition (10 dots difference). The left panels display data from correct trials and the right panels data from error trials. All data were submitted to a repeated-measures ANOVA with objective accuracy, difficulty, and RSI as factors and average confidence as a dependent variable.

This analysis replicated several effects already found in the previous experiment, for example higher confidence for correct trials compared to error trials, $F(1,\ 18) = 257.9,\ p < 0.001,\ \eta_p^2 = 0.93$; $M_{cor} = 4.7$ versus $M_{err} = 3.3$. There was also an effect of difficulty, $F(1,\ 18) = 4.6,\ p < 0.05,\ \eta_p^2 = 0.20$, despite this not showing in the overall averages, $M_6 = 4.0$ versus $M_{10} = 4.0$. Also, there was again a significant interaction between the two factors, $F(1,\ 18) = 243.6,\ p < 0.001,\ \eta_p^2 = 0.93$. This interaction was caused by the fact that the difference between error- and correct-trial confidence was larger for the easy, $M_{10} = 1.78$, compared to the difficult condition, $M_6 = 1.07$.

The key question here was whether the RSI manipulation had an effect on confidence, or more specifically on how confidence on error and correct trials

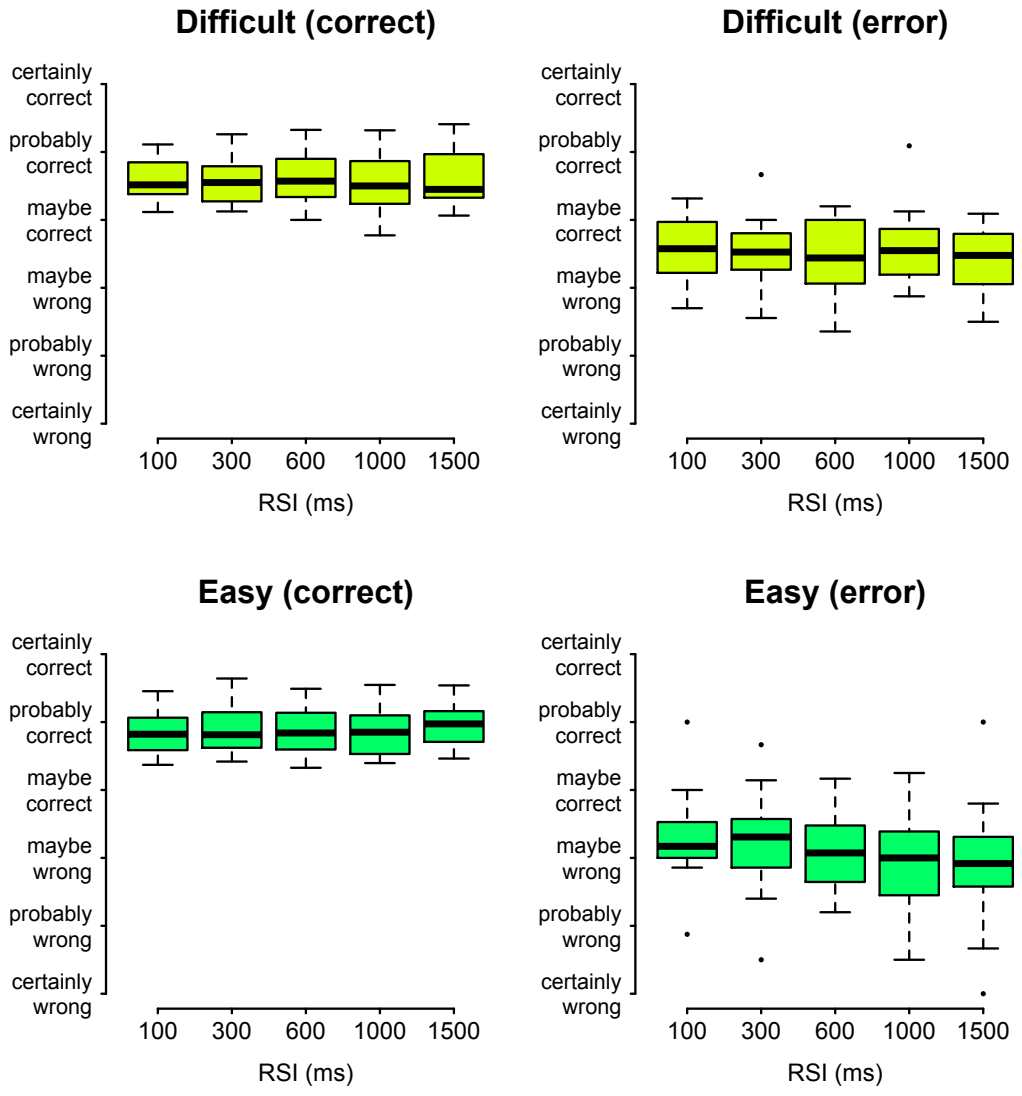

Figure 23: Confidence as a function of difficulty, RSI, and objective accuracy presented as box plots; 50% of the data lie within the box, the thick black horizontal line within each box indicates the median, the whiskers extend to 1.5 times the interquartile range (IQR); outliers are displayed as dots; upper panels: difficult condition (6 dots difference); lower panels: easy condition (10 dots difference); left panels: correct trials; right panels: error trials.

diverged. According to the post-decision processing hypothesis, additional time should increase the accuracy of metacognitive judgements. If so, we should find correct-trial confidence to increase with longer RSIs and error-trial confidence to decrease with longer RSIs. Statistically, this should be expressed in an interaction between the factors RSI and objective accuracy. Importantly, however, there was no significant effect of RSI, $F < 1$, and no reliable linear trend, $F(1,\ 18) = 2.5,\ p = 0.13,\ \eta_p^2 = 0.12$. There was also no reliable interaction of RSI with any of the other factors, $Fs <= 1.5,\ ps >= 0.20$.

There might not be an effect of RSI on raw confidence, but that does not mean that RSI does not affect metacognitive accuracy, which is what the next analysis will focus on. Figure 24 shows estimated SDT parameters as a function of difficulty and RSI, using the approach by Maniscalco and Lau (2012). Metacognitive sensitivity *meta-d′* is presented in the upper panel. Replicating the effects reported in Section 2.1.2.2, there was an effect of difficulty, $F(1,\ 19) = 28.0,\ p < 0.001,\ \eta_p^2 = 0.60$, with higher type-II sensitivity for the easy (10 dots difference) condition, $M_6 = 2.2$ versus $M_{10} = 3.2$. There was no effect of RSI, however, $F(4,\ 76) = 1.4,\ p = 0.25,\ \eta_p^2 = 0.07$, no reliable linear trend, $F(1,\ 19) = 1.9,\ p = 0.18,\ \eta_p^2 = 0.09$, and no interaction, $F < 1$.

As mentioned above, this measure should be compared to type-I sensitivity, d', which is presented in the middle panel of Figure 24. As in the previous study, this measure varied reliably with difficulty, $F(1,\ 19) = 153.0,\ p < 0.001,\ \eta_p^2 = 0.89$, with higher sensitivity for the easier (10 dots difference) condition, $M_6 = 1.6$ versus $M_{10} = 2.4$. There was also no effect of RSI and no interaction, $Fs < 1$. There was also no reliable linear trend for RSIs, $F < 1$.

The lower panel of Figure 24 depicts metacognitive efficiency, that is the common logarithm of the ratio between *meta-d′* and d'. Values larger than zero mean that more information was used for the second-order decision

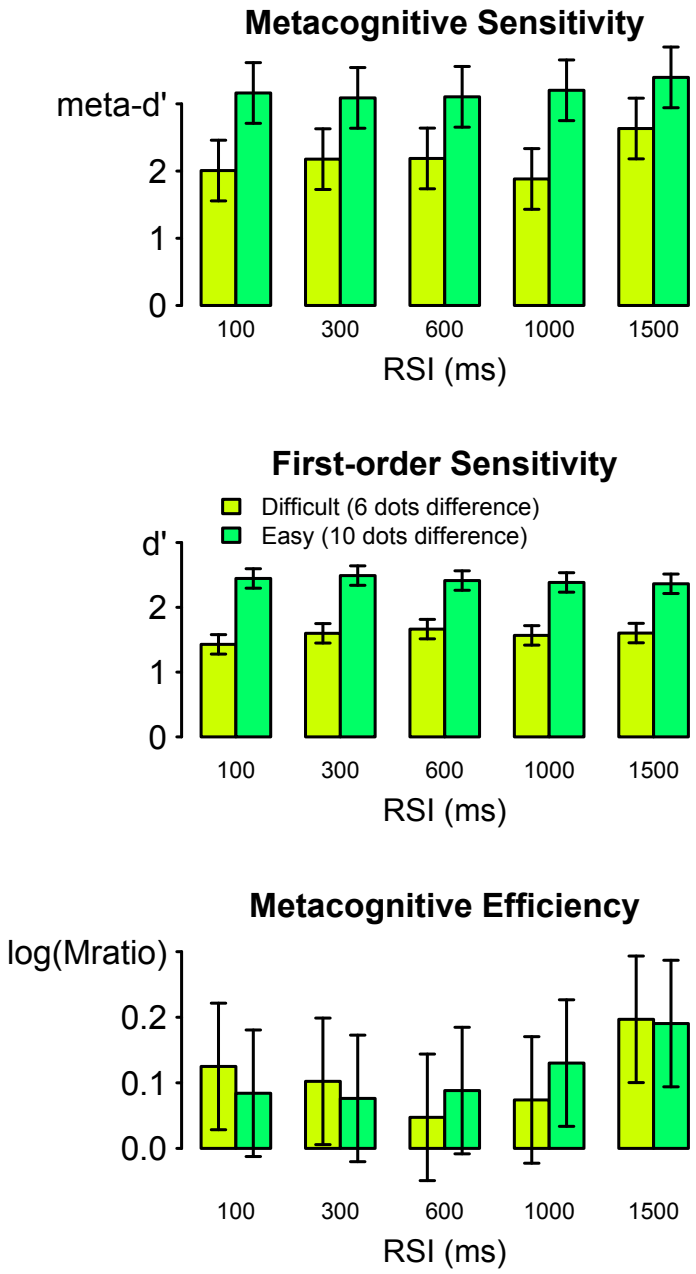

Figure 24: Metacognitive sensitivity (*meta-d′*; top panel), first-order sensitivity (*d′*; middle panel), and the common logarithm of metacognitive efficiency (*log*(*M-ratio*); bottom panel) as a function of difficulty and RSI. Both *meta-d′* and *d′* are coded in the same units. Values of *log*(*M-ratio*) larger than 0 can be interpreted as super-optimal metacognitive behaviour.

compared to the first-order decision. Two participants had to be excluded from this analysis. For these participants the ratio between second- and first-order sensitivity was zero or negative and the logarithm of this ratio could therefore not be calculated. Critically, there was no effect of RSI, $F(4,\ 68) = 1.6,\ p = 0.17,\ \eta_p^2 = 0.09$, and no linear trend, $F(1,\ 17) = 2.4,\ p = 0.14,\ \eta_p^2 = 0.12$. A post-decisional locus model would predict metacognitive efficiency to be lowest at the shortest RSIs, however, here I found metacognitive efficiency to be lowest for the 600 ms long RSI and higher for shorter and longer RSIs. The quadratic trend was only marginally reliable, though, $F(1,\ 17) = 3.2,\ p = 0.09,\ \eta_p^2 = 0.16$. There was also no effect of difficulty and no interaction effect, $Fs < 1$. The linear trend was also not reliable, $F < 1$. This suggests that the apparent difference in metacognitive sensitivity seems to stem from differences in sensitivity in the first-order decision. Taken together, these results indicate that how long participants had to wait before rating how confident they were does not have an influence on the accuracy of their rating, at least not within the time range studied in this experiment.

There exists yet another possibility by which RSIs could affect metacognitive processes – by influencing participants' metacognitive bias, that is how they map the confidence scale to their responses independent of how well these responses distinguish between correct and error trials. According to the post-decision processing hypothesis mentioned above, the longer participants wait, the more evidence they accumulate. One could assume that already integrated evidence influences how participants focus their attention, actively searching for more confirmatory evidence, that is evidence supporting the choice they have already made. According to this hypothesis, people should show more overconfident behaviour for longer RSIs, reflected in higher values of B_{ROC}. Figure 25 shows the values for this parameter as a function

of RSIs and difficulty. The effect of difficulty was not reliable, $F < 1$. This matched the findings from the previous study. Critically, there was no effect of RSI, $F(4,\ 60) = 1.1,\ p = 0.35,\ \eta_p^2 = 0.07$, and also no reliable linear trend, $F < 1$. There was also no interaction, $F(4,\ 60) = 1.1,\ p = 0.39,\ \eta_p^2 = 0.07$. Four participants had to be excluded for this analysis given that they did not have enough observations in one of the data cells. These findings suggest that the time participants had to wait before judging their confidence did not affect their over- or underconfidence.

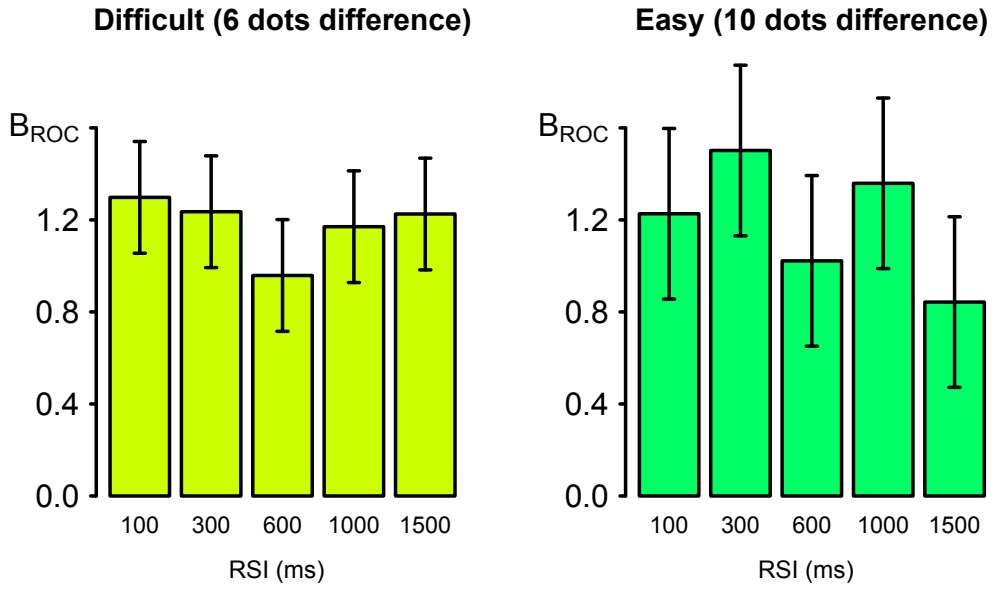

Figure 25: Non-parametric metacognitive bias, B_{ROC}, as a function of difficulty and response-stimulus interval (RSI). Larger values relative to lower ones suggest (relative) overconfidence; ms: millisecond.

2.2.3 Discussion

This study replicated a range of effects found with Experiment 1. There were only two levels of difficulty here, but they resulted in similar RTs and error rates as in the previous study, and difficulty itself had a similar effect on correct RTs as well as error rates. Once more, participants displayed good

metacognitive resolution; that is, they were more likely to rate a trial as highly confident when they were correct compared to when they had committed an error. Taken together, these findings indicate that the paradigm used in the previous and present experiments is a suitably robust tool to measure decision confidence.

The second part of the analyses focused on the effect RSIs might have on metacognitive processing, the main purpose of the present experiment. However, the analyses did not reveal any effects of RSIs on raw average confidence, metacognitive efficiency, or metacognitive bias. These findings can therefore be interpreted as support for the idea that when using retrospective confidence paradigms, differences as to when participants rated their confidence has little or no influence on these confidence judgements.

The question arises as to whether these findings should be understood as evidence against the general idea of post-decision processing. I would argue that such a conclusion cannot be drawn from these data – only that if post-decision processing contributes to confidence, as previous studies have shown (see Yeung & Summerfield, 2014, 2012, for a review), such processing must happen very fast after the decision is formed. The manipulation here started only at an RSI of 100 ms, which was presumably not fast enough to measure any such effect of post-decision processing. Moreover, the time participants took to rate their confidence varied over RSIs, which could be interpreted as direct evidence in favour of a post-decision processing account: In case of the shorter RSIs, participants took more time to rate their confidence, presumably finishing internal confidence judgement processes after the choice had been made.

There is one important confound that could presumably have reduced the power of the present experiment's RSI manipulation: Despite all attempts

to keep confidence RTs as constant as possible – by imposing an explicit speed pressure on participants for confidence RTs, while at the same time asking participants not to rate their confidence before the scale was presented – these confidence RTs were found to vary with RSI. However, if the extent to which this was the case is used to correct the RSIs, there is still a substantial difference in RSI and therefore the manipulation might still have worked despite its power being somewhat reduced.

There were, however, certain limitations with this study, which could be addressed in future studies. One of these limitations concerns the problem of motor preparation of the confidence response: Participants could possibly have formed their judgement as early as immediately after the response, then using any available time to prepare their motor response. This would lead to faster confidence RTs for the longer RSI conditions, which was indeed the case (see Figure 22). In future studies, the design should therefore be improved to be able to exclude such a possibility. This could be done by changing the confidence scale, similar to the wheel shown in Figure 26. The fraction highlighted as the confidence scale could change position and width on every trial, thereby not allowing participants to motorically prepare for their rating response. Another improvement would aim at imposing more speed pressure for the confidence responses – both by means of a temporal deadline and a reward scheme that lets participants earn extra credits for responding within a certain time window. The third improvement was already mentioned above – the range of RSIs used should be extended to include a fastest RSI level (scale is shown immediately after primary choice is made) as well as longer RSIs (2 to 3 seconds or more). The latter change would be aimed at testing the *memory-decay hypothesis* more directly in future studies.

Taken together, it can be concluded from the findings of Experiment 2

Figure 26: Circular confidence scale, which I suggested to address the issue of motor preparation as discussed in the main text. Participants judge their confidence on this wheel, of which a fraction is highlighted to represent the confidence scale. On each trial, this fraction would be located somewhere random on the wheel with a random width. Participants judge their confidence by clicking on a location on this wheel using a computer mouse. The starting point of the cursor is at the centre of the circle to ensure the precise measurement of confidence RTs.

that the precise timing as to when a metacognitive judgement is required does not have a reliable influence on the accuracy of such judgement. This suggests once more that confidence judgements are reliable and stable and can be recorded easily. Moreover, this means that studies on confidence can be replicated more easily than expected.

2.3 Experiment 3: Does judging confidence create task-switching costs?

The third methodological issue on which this chapter focuses concerns the effect introspecting might have on primary processing. Despite being one of the main methods of early experimental psychology (James, 1890), the age of behaviourism brought an intense skepticism towards this method which still holds today (see Wilson & Schooler, 1991; but also: Petitmengin, Remillieux, Cahour & Carter-Thomas, 2013; Jack, 2013). One reason for this skepticism is the assumption that judging one's confidence constitutes another task, and therefore an additional cognitive burden. In support of this hypothesis, several studies have reported findings of increased primary choice RTs when confidence responses were required (Petrusic & Baranski, 2003; Baranski & Petrusic, 2001; Grützmann, Endrass, Klawohn & Kathmann, 2014). These findings could suggest that confidence judgements impair first-order performance. Despite not finding a speed-accuracy trade-off, however, Grützmann et al. (2014), offered a different interpretation for this effect, assuming that increases in choice RTs were instead caused by participants assuming a more accuracy-focused strategy when facing confidence judgements. Such a strategy shift could presumably be caused by the fact that in a confidence rating condition, attention is more focused on error monitoring, using additional cognitive

resources. This interpretation was also supported by findings of enhanced error-related EEG activity for conditions in which participants were required to rate their confidence (Grützmann et al., 2014; Ullsperger & Von Cramon, 2006).

Yet another study suggests a co-variation of first- and second-order response strategy: Steinhauser and Yeung (2010) reported a trend of an effect of higher error rates in a condition in which a reward structure encouraged participants to adopt a low error signalling criterion, that is to avoid error-detection misses. Conversely, error rates were lower in a condition that encouraged them to adopt a high error signalling criterion, that is to be more cautious about their error signalling. In other words, if participants opted for a more cautious error signalling strategy, their first-order response strategy was also more cautious, whereas a more liberal error signalling strategy was associated with a more liberal first-order response strategy.

All of these findings seem to suggest that asking participants to rate their confidence leads to slower responses in the primary decision task, which could be due to a more accuracy-focused response strategy. A recent study by Hartwig and Dunlosky (2014, Experiment 3), on the other hand, led to the conclusion that no such difference exists between conditions in which people judge their confidence or just focus on the first-order decision task. In their study, half of the participants were required to judge their confidence after every response, whereas the second group did not perform such a judgement. The authors reported that there was no difference between the two groups. Assuming that judging one's confidence does not affect primary task performance is by no means implausible. For example, it could be assumed that confidence judgements are intrinsic in the decision process, given that they emerge from the same processes (see also Gherman & Philiastides, 2014). Reading out such

confidence measures would then therefore not pose an additional burden.

Taken together, these findings paint a somewhat mixed picture of the influence that confidence instructions might have on first-order performance. Experiment 3, therefore assessed to what extent making metacognitive judgements affects first-order performance. The key question was whether metacognitive processes are fundamentally different from basic task processes. If this was the case, then we should expect to find that making a judgement of confidence impairs task performance, similar to switch costs in the domain of task-switching studies (Monsell, 2003). Switch costs can be found when participants start a new task, reflected in higher RTs and error rates compared to task-repetition trials. It has been argued that switch costs arise from the necessity of reconfiguring the current task-set to match that of the new task, and thus reorganisation of mental resources (Allport et al., 1994; Jersild, 1927; Meiran, 1996). In the present experiment, I compared three experimental conditions, each of which was presented sub-block-wise; a baseline condition with just the primary choice task, a confidence condition similar to the previous two experiments, and an explicit task-switching condition in which participants had to alternate between the primary decision task and a second, unrelated task: After a dot decision, participants were shown a digit presented in one out of six possible locations on screen and they have to press one out of the six keys that are also used for the confidence judgement – without imposed time pressure to match the characteristics of the confidence judgement. If making confidence judgements relies on different cognitive processes, then we could expect to find switch costs if comparing the baseline to the confidence condition, similar to those observed when participants alternated between the primary decision-making task and an unrelated secondary task.

In the present experiment, different conditions were presented blocked,

as it is done in most confidence studies. This means that participants could always predict whether the current dot decision was going to be followed by a pause, a confidence judgement, or a digit-task stimulus. People were therefore able to prepare their switches, which presumably reduced switch-costs. However, one could argue that residual costs could still be expected, like the overall mixing costs that arise from letting participants alternate between two types of trials. These mixing costs are confounded with switch costs, in Jersild's method (Monsell, 2003). To my knowledge, the question of whether confidence judgements decrease task performance has not been addressed with such carefully selected baseline conditions.

2.3.1 Methods

2.3.1.1 Participants

I tested 24 participants, 18 of whom were female, and one was left-handed. The participants were 18 to 23 years old ($M = 19.6$). All participants received identical written instructions presented on screen. The first 4 participants also received verbal encouragement to focus on the speed of their responses rather than their accuracy, but this strategy was abandoned from participant 5 onwards. This change in instruction strategy was reflected in faster overall dot-task RTs for the first 4 participants compared to the others, $M_{first} = 388\ ms$ versus $M_{last} = 536\ ms$. This difference was reliable according to a Welch's t-test for unequal variances, $t(19.9) = 6.4,\ p < 0.001$. However, given that the present design focused on within-subject effects, participants are analysed together.

All participants had normal or corrected-to normal vision. All gave informed consent prior to the experiment, and testing was approved by the local

ethics committee. Each session lasted approximately 60 minutes, including the time to instruct prior to and debrief after the experiment. Three participants were paid for taking part in the study (£8); the remaining participants received course credit.

2.3.1.2 Task and procedure

The methods of the third experiment were again very similar to Experiments 1 and 2. This section will therefore only highlight in what way the methods differed from the previous studies.

There were three different conditions in this experiment. The first was a baseline condition with the dot task alone: Participants made a dot decision and were then shown a blank screen with a fixation cross for a variable length of time before the next stimulus was presented. This condition will be referred to as the Pauses condition. The durations of these pauses were sampled from a distribution at the beginning of each block. This distribution was formed by taking the mean and standard deviation of the confidence RTs and the RTs of the other task (see below) from the last block. These RTs were then averaged, so that one mean and one standard deviation for an exponentially modified Gaussian distribution were obtained. From this distribution, pauses were then sampled (bounded at 0 and 1500 ms). In the second condition, confidence judgements were measured on a 6-option scale. This condition will be referred to as the Confidence condition. The third condition required participants to alternate between the dot-task and an unrelated digit task with no time limit, the Digit condition. Participants had to press one out of six buttons according to a number shown on screen. This number was presented in one out of six horizontally-arranged rectangles, but the position of the number had to be ignored by the participant. After pressing a key, there was a 1000 ms blank

with the fixation cross before the next dot stimulus was shown. There were only incongruent digit-position pairings, that is the digit "1" was never shown in the first position, and so forth. Figure 27 shows all three conditions. The dot task in all of these conditions had a difference of 10 dots, equivalent to the easy condition in Experiment 2.

Only one out of the three conditions was presented within each sub-block, that is each block contained 60 trials and there were three sub-blocks with 20 trials each. Participants were told the order of the sub-blocks was random, but they would know which would follow next. Before the start of each sub-block, there was therefore a 5-seconds countdown, displaying for instance the words "Dot only part of this block will start in 5 sec."

The experiment started with a block of practice of the dot task. This was the only block with feedback tones and prolonged ITIs in case of a mistake, as in Experiments 1 and 2. The second block introduced confidence judgements. In the third block, participants familiarised themselves with the digit task. They practised alternating between this task and the dot task in the fourth practice block. These four blocks were 30 trials long. The fifth practice block was 60 trials long and exactly the same as the blocks in the experiment, therefore containing all three sub-blocks. Participants then completed 12 blocks of the main experiment. The order of the three parts within these 12 blocks was counterbalanced using a 6x6 balanced latin square. The sample size was thus set to a multiple of 6. The finger assignment for the confidence scale was also balanced over participants, as in Experiments 1 and 2. There was no such balancing for the digit task, with the numbers ranging from left to right for all participants.

As in Experiments 1 and 2, participants received feedback after every block regarding their average correct RTs and error rates of the dot task. After

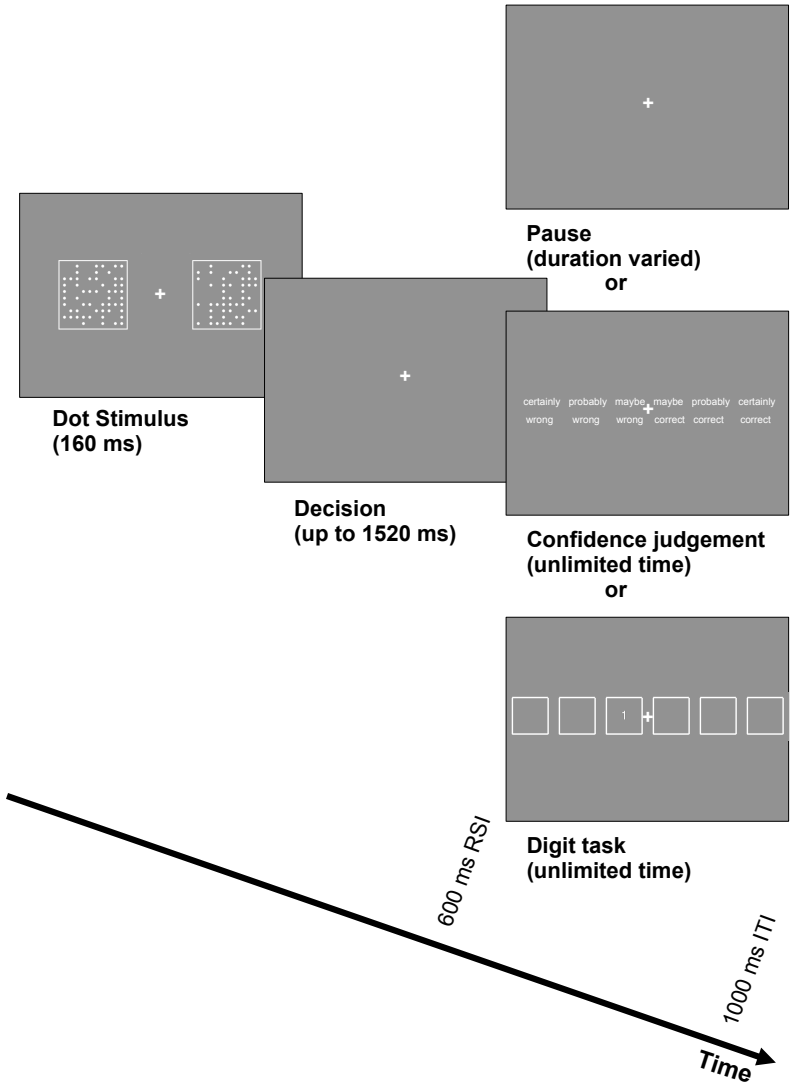

Figure 27: Methods of the dot task. Participants had to state which of two fields contained more dots by pressing the left or right key. After a 600 ms long response-stimulus interval (RSI), one out of three possible second halves of the trial followed, depending on the condition. In the Pauses condition, participants simply waited a variable duration before the next dot stimulus appeared. In the Confidence condition, participants judged their confidence on a 6-point verbal confidence scale. In the Digit condition, participants had to press one out of six keys depending on the identity (not the position) of a digit presented in one of the six horizontally presented squares. After a 1000 ms inter-trial interval (ITI), the next trial began; ms: millisecond.

blocks including digit-task trials, there was also feedback regarding the average error rate for this task. No mean RTs were shown for this task, as responding fast was not encouraged in these parts of the block.

2.3.2 Results

2.3.2.1 Replications

The first set of analyses focused on the question whether findings from Experiment 1 could be replicated. RTs on correct trials and error trials, as well as overall error rates, did not differ from the corresponding condition of Experiment 1 (all $ts < 1$).

Figure 28 plots error rates across confidence categories. Again, participants showed impressive resolution with error rates of 1.6% on trials classified as *certainly correct* and 80.2% on trials rated as *certainly wrong*. This pattern held over all participants, as expressed by an analysis that calculated individual participants' Spearman rank correlations. Such correlations were significant and highly negative for 23 out of 24 participants, $rs >= -0.83,\ ps <= 0.04$. For this one outlier, the correlation was $r = -0.14,\ p = 0.79$, which was caused by the fact that the participant only chose the *certainly wrong* option in three occasions, which happened to be correct trials. If the correlation was calculated over just five response categories, the correlation was perfect for this participant, $r = -1.00,\ p < 0.001$.

The final replication analyses focused on raw confidence, as well as SDT parameters. Replicating the findings from the two previous studies, confidence was higher on correct than error trials, $t(23) = 11.8,\ p < 0.001$; $M_{cor} = 4.6$ versus $M_{err} = 3.0$. Average metacognitive sensitivity according to the method developed by Maniscalco and Lau (2012), *meta-d′*, was $M = 1.84$. With an

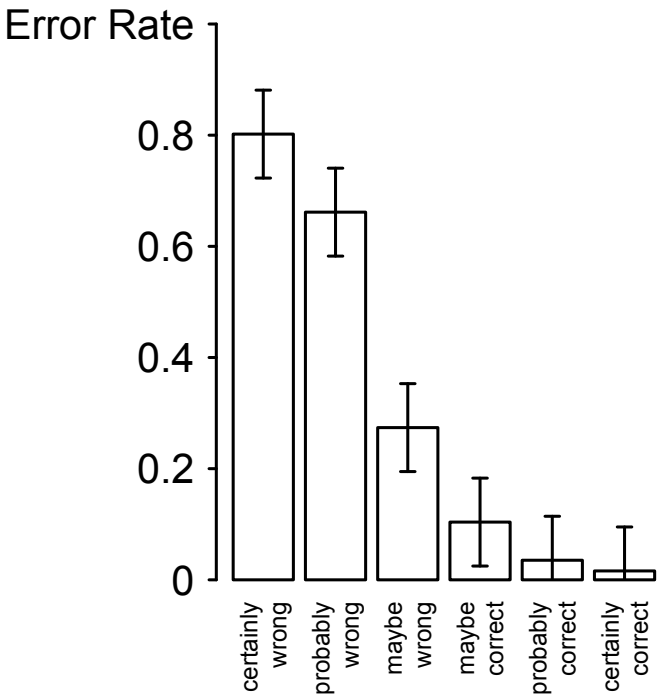

Figure 28: Error rates as a function of subjective confidence ratings.

average first-order sensitivity, d', of $M = 2.55$, this led to a metacognitive efficiency, $log(M\text{-}ratio)$, smaller than zero, $M = -0.17$, which means participants used less information for the second-order decision compared to the first-order decision. This finding differed from the previous two studies in which average metacognitive efficiency was larger than zero, reflecting additional evidence contributed to the confidence judgement compared to the primary decision, $M_{Exp1} = 0.04$ versus $M_{Exp2} = 0.11$. Both values were just taken from the same difficulty condition and – in case of EXPERIMENT 1 – only the 6-options scale. This difference was indeed reliable, $F(2,\ 65) = 8.5,\ p < 0.001,\ \eta_p^2 = 0.21$, and post-hoc comparisons revealed that metacognitive efficiency in present experiment was reliably lower than both in EXPERIMENT 1, $t(44.9) = 3.0,\ p < 0.01$, and EXPERIMENT 2, $t(41.9) = 3.9,\ p < 0.001$. This difference may reflect the reduced speed stress in the present experiment, therefore reducing the overall number of premature responses and thus the number of detected errors.

2.3.2.2 Task-switching effects

As a first manipulation check, I tested whether RTs of the digit task roughly matched the participants' confidence RTs. Ideally, RTs in the two tasks should be roughly similar, however, participants were reliably slower at judging the digit stimuli compared to their confidence RTs, $t(23) = 6.5,\ p < 0.001$; $M_{Digit} = 659\ ms$; $M_{Confidence} = 466\ ms$. Participants committed on average $M = 5.6\%$ errors in the digit task, but there was no goal to match these error rates to the performance on the confidence scale. To compensate for such possible differences, the RSIs in the Pauses condition had been constructed in a way so that they would be equally similar to both other conditions. The difference between confidence RTs and pauses was $M = 113\ ms$, which was reliable, $t(23) = 6.1,\ p < 0.001$. The difference between digit RTs and pauses

was $M = 79\ ms$, which was also reliable, $t(23) = 6.3,\ p < 0.001$. It can therefore be concluded that despite insufficient matching of response latencies in the Confidence and Digit condition, at least both of these conditions were roughly equidistant to the sampled RSIs of the Pauses condition. This is crucial given that key analyses will focus on condition contrasts.

The key question addressed with this experiment was whether including confidence judgements in a task leads to switch costs – reflected in increased RTs and error rates in the primary task – given that confidence relies on separate cognitive processes. Alternative to leading to switch costs, however, confidence ratings could also affect response strategies: Having to judge one's confidence presumably could have caused participants to adopt a more accuracy-focused response regime. According to this idea, we should expect increased RTs and decreased error rates for a condition in which confidence had to be rated after every trial.

Correct RTs are presented in the upper left panel of Figure 29. There was a reliable main effect of condition, $F(2,\ 46) = 6.9,\ p < 0.01,\ \eta_p^2 = 0.23$, with fastest RTs for the Pauses condition and slowest in the Confidence condition, $M_{Confidence} = 532\ ms$; $M_{Digits} = 507\ ms$; $M_{Pauses} = 493\ ms$. Bonferroni-corrected post-hoc t-tests revealed a significant difference between only the Confidence and the Pauses condition, $t(23) = 3.3,\ p < 0.0083$. There was only a marginally significant difference between the Confidence and the Digits condition, $t(23) = 2.3,\ p = 0.03$, and no reliable difference between the Digits and the Pauses condition, $t(23) = 1.6,\ p = 0.12$.

Error rates in the three conditions are shown in the upper right panel of Figure 29. Error rates were lowest in the Confidence condition and highest in the Pauses condition, $M_{Confidence} = 11.1\%$; $M_{Digits} = 13.0\%$; $M_{Pauses} = 13.3\%$. This difference was reliable, $F(2,\ 46) = 3.3,\ p = 0.04,\ \eta_p^2 = 0.13$. Post-hoc t-

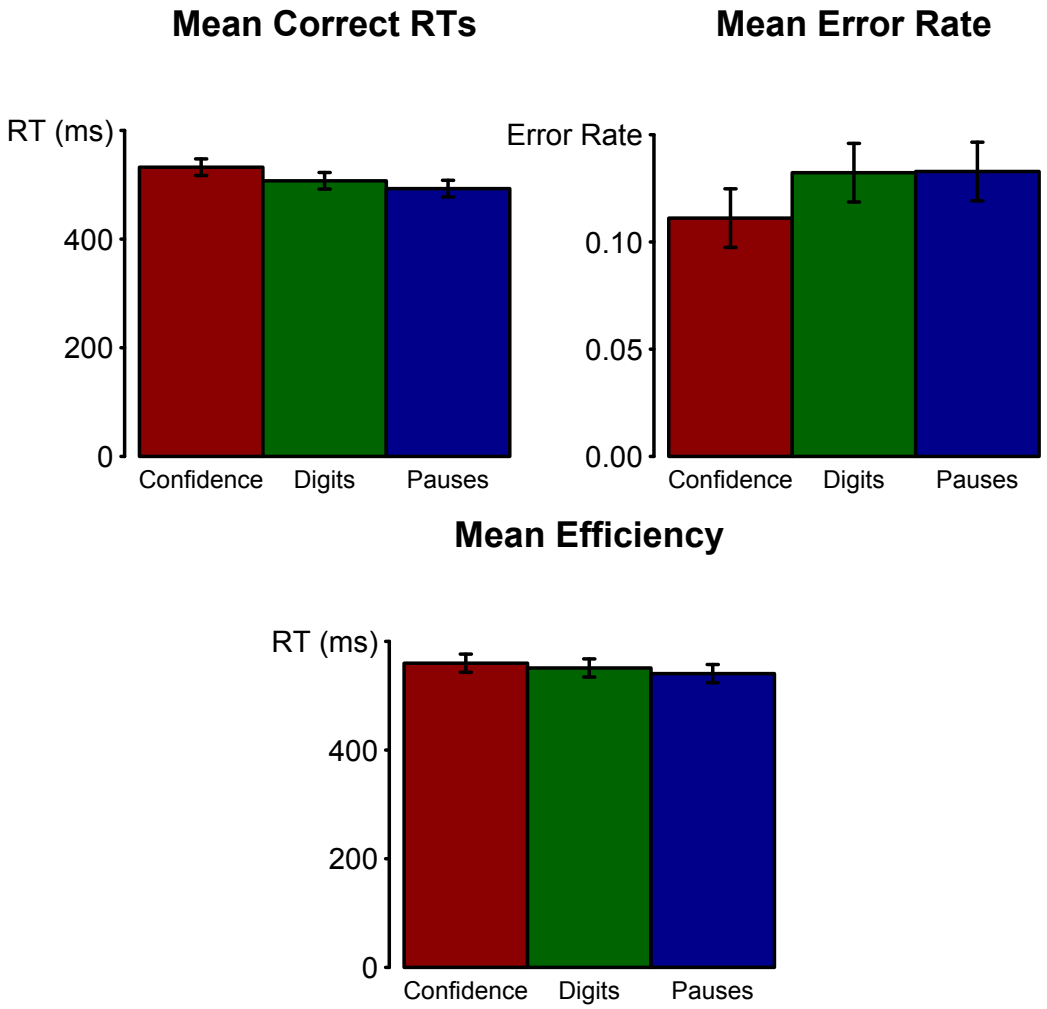

Figure 29: First-order task performance as a function of condition. Upper left panel: mean correct response times (RTs); upper right panel: mean error rates; lower panel: mean efficiency, that is RTs penalised for errors; ms: millisecond.

test with Bonferroni-corrected α-levels (0.017) revealed a significant difference only between the Confidence and the Pauses condition, $t(23) = 2.7,\ p = 0.01$, but not between the Confidence and the Digit condition, $t(23) = 2.2,\ p = 0.04$, or the Digit and the Pauses condition, $t < 1$.

Taken together, participants were slowest but most accurate in the Confidence condition. These results therefore speak against overall task-switching effects for the Confidence and Digits conditions compared to the Pauses condition. Instead, they hint at the possibility that the effect could lie in the response strategy chosen by the participant, that is people opt for a more cautious response strategy in the primary task whenever the confidence scale requires them to introspect. If this strategy hypothesis held true, overall task performance should be matched over conditions. Such task performance was arguably reflected in the efficiency averages shown in the lower right panel of Figure 29. This efficiency score is formed by dividing the median correct RT by accuracy (inverse efficiency score; *IES*; Bruyer & Brysbaert, 2011). These accuracy-corrected RTs were highest for the Confidence condition and lowest for the Pauses condition, $M_{Confidence} = 560\ ms$; $M_{Digits} = 551\ ms$; $M_{Pauses} = 541\ ms$. However, there was no reliable main effect of condition, $F(2,\ 46) = 1.3,\ p = 0.27,\ \eta_p^2 = 0.05$, nor were any post-hoc comparisons between conditions significant, $ts <= 1.6,\ ps >= 0.13$. Importantly, with a conventional frequentist ANOVA as used here, the null hypothesis can only ever be rejected. Bayesian models were therefore fitted to the data (Rouder, Morey, Speckman & Province, 2012) to compare all possible combinations of main and interaction effects. There was most evidence in favour of the null effect, $BF_{NULL} = 3.14$. These findings suggest that there is no overall impairment in task performance due to switch costs for both the Confidence as well as the Digits condition compared to the Pauses control condition. Taken to-

gether, the findings discussed here did not reveal any substantial task-switching effects. Participants were overall slower and more accurate when judging their confidence.

2.3.3 Discussion

The main goal of this study was to investigate whether confidence judgements relied on different processes compared to primary decision making, which should lead to task-switching effects. The analyses revealed no such impairments in task performance. Instead, the data seem to support a more cautious response strategy for the confidence trials, with slower RTs and lower error rates for the Confidence compared to the Pauses condition.

Perhaps the null effect of task switching was due to the similarity of the primary response and the confidence judgement. As argued above, a higher similarity between two tasks could presumably mean than the task-set has to be reconfigured less. However, there was no reliable difference between the Confidence and the Digit condition or the Pauses and the Digit condition in RTs or error rates, which speaks against this interpretation.

The present experiment replicated several findings from Experiments 1 and 2. First, the RTs and error rates for the difficulty conditions chosen here roughly matched those of Experiments 1 and 2. Second, confidence ratings covaried with objective task performance, which is also reflected in higher average confidence levels for correct as opposed to error trials, and the distributions of confidence answers were similar to the previous experiments. However, metacognitive efficiency was significantly smaller than in Experiment 1 or 2, as well as lower than zero. Values less than zero reflect that the participant had less information available when making the second-order judgement compared to the first-order decision. Values larger than one, on the other hand,

can arguably reflect post-decision processing. The present experiment was the first in which the speed stress was slightly lifted, which might in fact have led to confidence being judged more during the time leading up to the decision (Baranski & Petrusic, 1998), therefore supporting a more decisional locus of confidence.

Taken together, the findings from this study point towards the conclusion that if confidence instructions cause task-switching effects on the primary performance, such inference must indeed be very small and therefore need not to be further taken into account. However, substantial task-switching effects would have been expected in the Digit control condition. Not finding at least some effect of task-switching when alternating between the dot task and such an unrelated task suggest that the present paradigm lacked power to even detect task-switching effects. This might have been caused by several limitations of the design. First, in the present study, conditions were presented blocked. This was done to encourage participants to prepare their confidence judgement while responding to the primary stimulus when within the Confidence part of the block, thereby maximising possible interference effects. Critically, most studies that include metacognitive judgements record these on every trial, therefore the present experiment simulated these conditions very closely. On the other hand, intermixing confidence trials with other conditions might reduce the strategy-effect we found in this study, which is a prediction that remains to be tested. Therefore, the present study could be extended to include a condition in which different conditions are intermixed within a block.

Yet another change that could be considered, as frequently done in the task-switching literature, is the use of bivalent stimuli: Using bivalent stimuli means that the stimuli for the tasks between which participants al-

ternate are identical, but the different tasks focus on different properties of these stimuli, for example judging a number as larger or smaller than 5, or judging odd versus even. However, this is not entirely possible, given that the confidence judgement is estimated internally. One could therefore imagine a re-designed version of the experiment in which the digit task is replaced by another judgement that has to be made internally, from short-term memory. I would, for instance, use the same dot task as before but now present the dots in six different colours. Instead of the digit task, participants have to press a key according to the colour the just-judged dot stimulus was presented in. Such changes to the original design could increase the power to detect possible task-switching effects.

2.4 General discussion

In the present chapter, a paradigm was tested for its suitability for measuring metacognitive judgements. This paradigm constitutes the basis for the experiments in the following two chapters. The findings of Experiment 1 suggest that the dot task is a suitable task with fine-grained levels of difficulty. These levels range from very easy (20 dots difference; 4.4% errors) to very difficult (2 dots difference; 39.9% errors). Crucially, the difficulty manipulation was also reflected in different ratings of confidence, with participants being more confident on easier correct trials, due to increased metacognitive insight facilitated by more available information at the time of the primary decision. At the same time, errors are rated as less confident on easy trials, which can also be explained by more insight (in this case, insight that a response was incorrect). Critically, difficulty did not affect metacognitive efficiency, meaning that all differences in insight were fully explained by differences in first-order

performance. Moreover, no hard-easy-effect (higher level of overconfidence for difficult conditions) was found in the present paradigm. The difficulty effects were replicated in the three different experiments. Taken together, this task provides the experimenter with precise experimental control and is therefore a suitable paradigm for the purpose of the questions addressed in this thesis.

Moreover, the results demonstrate that a verbal, discrete confidence scale is a suitable tool to assess decision confidence with. Using a discrete rather than a continuous scale has the advantage of precise measurement of confidence RTs. The main goal of Experiment 1 was to test whether a more fine-grained version of this scale with six as opposed to only two levels could be used. The findings supported that this was indeed the case with no processing costs for the more fine-grained scale, and comparable metacognitive accuracy and speed with which confidence was rated. This was a key result in the investigation of confidence in decision making. Yet another advantage of using a more fine-grained confidence scale is that the confidence ratings can then be used to calculate ROC curves, as well as SDT measures based on these curves. I will therefore use this version of the scale, a scale which allows me to assess on which trials participants changed their minds, in all subsequent experiments of this thesis. While participants arguably treat the binary and the 6-point scale very similarly, the question remains as to whether confidence and error detection are distinct or two sides of the same coin (Yeung & Summerfield, 2012). I will address this question further in Chapter 3.

Moreover, my findings suggest that measuring confidence judgements is straightforward and can be implemented without the need to train participants given the verbal scale used here. Instead, participants repeatedly showed intuitive use of the confidence scale right from the very first block onwards. However, there was an increase in metacognitive efficiency over trials,

presumably caused by participants getting faster and therefore committing more speed errors, which are easier to detect. I therefore concluded that participants should be monitored for changes in response speed and accuracy and reminded verbally and through performance feedback to trade off speed and accuracy within their first-order responses throughout the experiment.

Findings from Experiment 2 suggest that the precise timing of the confidence judgement does not affect metacognitive processing. More precisely, if people are given more time to judge their confidence neither does their metacognitive accuracy increase – as could be expected from the post-decision processing locus hypothesis – nor does their metacognitive accuracy decrease – as could be expected from a memory decay hypothesis. There were certain limitations with regard to this study: The speed with which participants rated their confidence varied with the delay before the onset of the scale, suggesting that participants formed their decision much earlier during the delay and then used the remaining time to prepare for the button press motorically. This could be reduced by using a variable scale and enforcing speed pressure for the confidence judgement even more rigorously, potentially even by introducing a reward scheme for both first- and second-order responses. However, this limitation does not affect the conclusion that decision confidence is a stable phenomenon, which goes unperturbed by timing specifics of the measurement of metacognitive judgements.

The results of Experiment 3 clarify the nature of task-switching effects that might occur when confidence judgements are required. It was somewhat surprising that the results presented here suggest that asking participants to rate their confidence in a perceptual decision-making paradigm does not have a detrimental effect on first-order task performance. Instead, my data support the notion that participants opt for a more accurate response strategy

when they have to judge their confidence, that is they tend to be slower and more accurate. One interpretation of the findings from EXPERIMENT 3 is that the absence of task-switching effects was caused by a high degree of shared resources between the confidence and the dot task itself. This finding can be interpreted as further support for the idea that requiring participants to judge their performance does not negatively affect such performance due to not having to reconfigure the task-sets in question. However, the absence of task-switching effects in the unrelated task, another control condition, render the results somewhat ambiguous in the sense that the overall design might have lacked power necessary to detect any such task-switching effects.

To conclude, the design developed in the current chapter provides a stable paradigm to carry out research on confidence in decision making. The three experiments described in this chapter highlight the fact that confidence judgements are both reliable and stable, increasing reproducibility of studies on metacognition. Confidence judgements moreover do not appear to have a detrimental effect on first-order performance, but rather a small effect of response strategy leading to a more accurate response mode, as well as a slightly enhanced effect of post-error slowing, which can be taken into account when designing experiments.

Chapter 3

Investigating the relations between confidence and error detection

This chapter focuses on the relationship between two key metacognitive evaluations: error detection and confidence judgements. These types of judgements have separately been studied in detail but have rarely been directly compared (except for Yeung & Summerfield, 2014, 2012; Scheffers & Coles, 2000; Fernandez-Duque et al., 2000). Similar methodological approaches have been used in prior work on those two lines of research: The participant makes a first-order perceptual decision and is then asked to evaluate his or her choice by being asked either "how confident are you that you were correct?" or "did you make an error?" While in research on confidence, a graded scale is usually given to participants to rate the correctness of their response (e.g., Fleming, Huijgen & Dolan, 2012; Zylberberg et al., 2012; Bahrami et al., 2012; Fleming et al., 2010; Baranski & Petrusic, 1994; Pleskac & Busemeyer, 2010; Baranski & Petrusic, 1998; De Martino et al., 2013; Petrusic & Baranski, 2003), participants are asked to judge the opposite in research on error detection, that is they have to evaluate incorrectness of the response, usually in a binary man-

ner (e.g., Rabbitt, 1968; Charles et al., 2013; Endrass et al., 2005; Wessel et al., 2011). These methodological differences have been discussed in detail in Section 2.1.

Despite the many similarities in approach, there is little compatibility between current theories of confidence and error detection (Yeung & Summerfield, 2012, 2014). For example, popular models of confidence, such as the balance-of-evidence hypothesis (Vickers & Packer, 1982), can explain graded confidence judgements but not why participants sometimes state with certainty that an earlier response was incorrect (one exception is the post-decisional balance-of-evidence model proposed by Van Zandt & Maldonado-Molina, 2004). Conversely, many theories propose error detection to be all-or-nothing (e.g., Falkenstein et al., 1991; Gehring et al., 1993) and therefore struggle to explain graded judgements of confidence. Given the previously reported findings from Experiment 1, in which participants showed equally good metacognitive insight on both graded and binary confidence scales, however, there is evidence that such graded judgements exist and that participants are intuitively capable of reporting their confidence using graded confidence scales. Taken together, over the last decades, little systematic effort has been made to link the two lines of research theoretically (Yeung & Summerfield, 2012, 2014; Fernandez-Duque et al., 2000; Davelaar, 2009).

Empirical findings are similarly discrepant. For example, Charles et al. (2013) recently observed dissociations in the use of binary error judgements in conscious and non-conscious conditions: Whereas above-chance confidence judgements were observed even on trials in which stimuli remained subliminal due to visual masking, error-related EEG activity was only evident on conscious trials. Charles et al. (2013) therefore concluded that these judgements reflect the existence of two separate metacognitive systems: An all-or-nothing

error signal, which is conscious, and a graded non-conscious system for confidence judgements. Scheffers and Coles (2000), on the other hand, had found that error-related EEG activity varies in a graded way with subjectively-rated confidence, implying that confidence and error detection might be two sides of the same coin.

Here, I argue that linking error detection and decision confidence would have substantive implications for current theories in the respective fields. For instance, the neural basis of metacognitive monitoring remains a highly debated topic (see for instance Fleming & Frith, 2014, for a review). If decision confidence could be linked to well-characterised EEG correlates of error processing, then this would provide useful constraints on theories of the neural bases of metacognition in decision making. Furthermore, another implication would be to hypothesise the existence of graded error signals and include them in current theories of error monitoring. Finally, the existence of *sure errors* remains a challenge for many current theories of decision confidence, which assume a decisional-locus model, and are therefore unable to explain why participants sometimes change their mind. These theories should be expanded to include post-decision processing (Yeung & Summerfield, 2012, 2014).

The previously reported results from Experiment 1 already shed some light on the question how confidence and error detection judgements can be linked, observing no difference in how participants used the binary and graded scale: They were equally accurate in their metacognitive judgements and showed no difference in their tendency to rate their responses as highly confident. Moreover, graded error detection was observed (*certainly wrong*, *probably wrong*, *maybe wrong*), which is rarely measured in metacognition studies, which either use a binary scale or a confidence scale with *guessing* as the lowest confidence category (Baranski & Petrusic, 1994; Wallsten et al., 1993;

Buratti & Allwood, 2012; Van Zandt & Maldonado-Molina, 2004; Scheffers & Coles, 2000).

Findings from Experiment 1 were interpreted mainly from a methodological point of view, that is they were taken to mean that a graded scale, given its advantage of a higher measurement resolution, does not impose additional processing costs. The focus in the present chapter, however, was whether those different forms of judgements constitute different forms of the same underlying metacognitive processes. Experiment 4 aimed to establish the relationship between error detection and confidence. More specifically, I asked whether well-characterized EEG correlates of error detection are sensitive to, and are predictive of, fine-grained differences in correct-trial confidence. Two such EEG correlates have been extensively studied in prior research: The ERN and the Pe. In contrast to prior work focusing on the ERN, a fronto-central component observed immediately following errors, here I focus on the subsequent parietal-focused Pe because of its established link to subjective error awareness (Overbeek, Nieuwenhuis & Ridderinkhof, 2005; Steinhauser & Yeung, 2010; Endrass, Klawohn, Preuss & Kathmann, 2012). The core rationale is that if error detection and confidence judgements share common underlying mechanisms, well-characterised neural correlates of error awareness should then be predictive of participants' decision confidence on a trial-by-trial basis.

In the second part of this chapter, I then summarise how I believe metacognitive judgements are formed given the findings reported so far in this thesis. The assumed mechanisms will be formalised in a computational model and simulation results will be reported. These results suggest that the model captures the most basic first- and second-order results reported for the dot-count task. The model assumes a common internal metacognitive signal (*post-*

decisional balance of evidence) from which both confidence and error detection judgements are derived from evidence accumulated in a single decision process. As an even stronger test of the model, data from a visibility manipulation task are then simulated. In particular, this model aimed to capture the fact that metacognitive accuracy strongly increases for trials reported as *seen* as opposed to trials rated as *unseen*. The model successfully replicated these data patterns observed by Charles et al. (2013), which contradicts the hypothesis proposed by Charles and colleagues that these data can only be explained by a model that assumes two separate routes that accumulate information: one conscious and slow, but accurate, the other route noisy, fast and automatic (see also Del Cul, Dehaene, Reyes, Bravo & Slachevsky, 2009).

3.1 Experiment 4: Do confidence and error detection rely on similar processes?

The present experiment focused on whether there is a shared neural basis of the two judgements, that is whether EEG correlates of error detection are also sensitive to fine-grained changes in correct-trial confidence. Two such correlates were considered, given that errors in speeded decision tasks are associated with a characteristic sequence of ERP components, the ERN and the Pe (Falkenstein et al., 1991).

First, the ERN is a negative deflection in the EEG, peaking within 100 ms of the incorrect motor response. The ERN is larger for errors compared to correct trials and this difference is largest at fronto-central electrode sites. The negative activation observed on correct trials is often referred to as the correct related negativity (CRN). It has been argued by some authors, such as Pailing and Segalowitz (2004), that the CRN reflects response uncertainty, that

is correct responses that have been incorrectly perceived as errors. Similarly, the amplitude of the ERN has been interpreted as a reflection of error awareness (see Wessel, 2012, for a review). This means that the difference between correct and incorrect trials should be larger for reported compared to unreported errors. Indeed, Scheffers and Coles (2000) found that the ERN amplitude correlates with error detection, meaning a larger (i.e., more negative) ERN is observed for low-confidence trials. Wessel (2012), however, proposed that the ERN should be interpreted not as a consequence of error awareness but instead as a precursor of it. This matches what has been proposed in a recent study by Steinhauser and Yeung (2010), who suggest that the ERN is related to probabilistic information of whether or not an error has occurred, such as detection of response conflict (Yeung, Botvinick & Cohen, 2004) rather than error awareness *per se*. Therefore, findings as to whether the ERN reflects error awareness remain contradictory, instead the second error-related EEG component, the Pe, has previously been shown to vary reliably with the degree to which participants report awareness of their errors (Nieuwenhuis et al., 2001; Hester, Foxe, Molholm, Shpaner & Garavan, 2005; Steinhauser & Yeung, 2010; Murphy, Robertson, Allen, Hester & O'Connell, 2012). The Pe is an extended centro-parietal positivity, which usually peaks around 200 to 300 ms after a response and is larger for error compared to correct trials. More specifically, Steinhauser and Yeung (2010) found that Pe amplitude reflects the internal evidence accumulated in favour of an error having occurred. Similarly, Nieuwenhuis et al. (2001) have previously found this component to be more enhanced when the error had been detected than when it was undetected. Moreover, Murphy et al. (2012) reported that Pe peak latency correlated with how fast participants had reported their errors, suggesting that the Pe reflects the temporal dynamics of emerging error awareness. Taken together, while

findings regarding the relationship between the ERN and error awareness is mixed, compelling evidence exists in support of the hypothesis that the Pe reflects conscious error detection. This component therefore ought to be taken as the best index of error awareness.

The question addressed with this experiment is whether such well-characterised error-related EEG components also vary with confidence. To test this prediction, multivariate pattern classifier techniques were used to assess the degree of overlap between neural signatures of error detection and decision confidence. Specifically, I assessed the degree to which cross-classification is possible between error detection and decision confidence. In other words, will a classifier algorithm that is trained to distinguish correct versus error trials be predictive of more subtle variation in correct-trial confidence? Previous research has shown that a multivariate classification approach can robustly index single-trial Pe amplitude to distinguish objectively correct and incorrect responses (Steinhauser & Yeung, 2010). The novel question addressed here was whether a classifier trained in this way would similarly predict variations in confidence on a single-trial level. Specifically, the question was whether a classifier trained to discriminate errors versus a matched subset of correct trials would be predictive of varying levels of confidence on the remaining set of (untrained) correct response trials. To the extent that the multivariate classifier generalises successfully in this way – in particular, to predict subtle variation in correct-trial confidence – this would provide evidence for shared neural correlates of error detection (as studied extensively in past research focusing on the Pe) and subjective confidence (as assessed here).

3.1.1 Methods

3.1.1.1 Participants

Seventeen right-handed participants (8 female) between 21 and 30 years of age ($M = 23.9$), all with normal or corrected-to-normal vision, gave informed consent and were paid for their time. One participant had to be excluded due to a technical recording error, leaving 16 participants in the final sample. Each session lasted between 80 and 120 minutes, including EEG setup, instruction, and debriefing. All procedures were approved by the local ethics committee.

3.1.1.2 Task and procedure

The paradigm used for this experiment was similar to the one previously described, so here I describe only the parts that differed from the previous studies. Only one level of difficulty was used in this version of the protocol, a difference of 10 dots. Again, only the 6-options confidence scale was used. Speed was again stressed so that participants made sufficient numbers of errors to permit planned contrasts of neural activity on correct versus error trials. In the main experiment, participants completed 18 experimental blocks of 48 trials. Prior to the main experimental blocks, participants first completed two practice blocks to become familiarised with the perceptual decision task (including auditory feedback) and the confidence rating scale, respectively.

3.1.1.3 EEG recording

Participants were seated in a dimly lit, electrically shielded room. EEG data were recorded using Ag-AgCl electrodes embedded in a fabric cap (QuikCap, Neuroscan, El Paso, TX) from 32 channels: FP1, FPZ, FP2, F7, F3, FZ, F4,

F8, FT7, FC3, FCZ, FC4, FT8, T7, C3, CZ, C4, T8, TP7, CP3, CPZ, CP4, TP8, P7, P3, PZ, P4, P8, POZ, O1, OZ, O2, as well as the left mastoid. All electrodes were referenced to the right mastoid online and re-referenced to linked mastoids offline. The vertical and horizontal electrooculogram (EOG) was measured from above and below the left eye and the outer canthi of the two eyes. Electrode impedances were kept below 50 kΩ. EEG and EOG data were continuously recorded using SynAmps2 amplifiers (Neuroscan, El Paso, TX) at a sampling rate of 1000 Hz, with a band-pass filter of 0.1 – 200 Hz, a gain of 2816, and a resolution of 29.8 nV.

3.1.1.4 Data analysis

EEG preprocessing. EEG data were preprocessed using the approach described in detail in Section 1.8. Response-locked epochs were then extracted from the continuous data. Those epochs were baseline-corrected to -100 to 0 ms pre-response for the main analyses, and -100 to 0 ms pre-stimulus for the additional analyses of the stimulus-locked P3. Trials were rejected as containing artefacts if the signal exceeded -100 to 100 μV in the electrodes FZ, FCZ, CZ, CPZ, and PZ (due to noisy channels these channels were used without FZ for one participant and without FZ and PZ for another; later interpolating these channels). The data were then low-pass filtered offline at 12 Hz.

EEG analysis focused on the 600 ms interval between the participant's key-press response in the perceptual decision task and the subsequent appearance of the confidence scale. In the averaged ERP data, the Pe was quantified as the difference between error- and correct-trial waveforms in an interval from 250 to 350 ms after the response.

Single-trial EEG analysis. Of critical interest here was the relationship between the Pe and participants' confidence judgements. In particular, it was predicted that variation in the error-related Pe, as measured in the period after the participant's response but before their confidence rating, would be predictive of fine-grained changes in subjectively-rated confidence.

To quantify Pe amplitude robustly on single trials, a classifier based on spatial linear integration (Parra et al., 2002) was trained to distinguish between objectively correct and incorrect responses, using a subset of the data from each participant. This classification method has been used successfully in the past to distinguish between experimental conditions (Philiastides, Ratcliff & Sajda, 2006; Ratcliff, Philiastides & Sajda, 2009; Steinhauser & Yeung, 2010; Macdonald et al., 2011). The subsets were matched-size samples of correct and incorrect responses, all baseline-corrected and also taking into account the hand with which the response was made so that the classifier would not reflect differences in motor activity. Linear integration aims to find a spatial ERP component by identifying a classifying vector that maximally discriminates between the two conditions in question – in this case errors and correct responses. If $x(t)$ is taken to be the vector of activity across electrodes at time t, then logistic regression can be used to compute a spatial filter v, which results in the discriminating component

$$y(t) = v^T x(t). \tag{1}$$

This component is maximally discriminating between the two conditions. Through this approach, signal-to-noise ratio (SNR) is improved on the single-trial level by combining data across electrodes, much as SNR is improved in conventional ERP analyses by averaging across trials. Once the optimal

weighting coefficient v has been determined, single-trial short-time averaged discrimination activities $\overline{y}_k$ can be estimated for each trial k by summing over all T samples – in my case 100 samples – from each trial. These single-trial short-time averaged discrimination activities vary between 0 and 1, with larger values reflecting a higher probability that the trial in question was an error.

To identify the classifying vector that maximised activity specific to one of the two conditions while minimising non-specific activity, that is processes contributing to both conditions, however, it is crucial to train the classifier on a suitable subset of data from each trial. Here, I used 100 ms long time windows from each trial. The width of this time window was set to 100 ms *a priori* given that it had previously been shown to be a suitable window width analyses focusing on the Pe (Steinhauser & Yeung, 2010). The starting point of the time window on which the classifier would be trained was then found using sweeping windows by 10 ms from the response onwards up to 600 ms post-response. This procedure resulted in a total of 60 time windows, for each of which classification performance was estimated, so the time window with the highest classification performance could be found. Within each time window, all samples are treated as independent.

Once the discriminating component $y(t)$ has been estimated, this vector can be applied back to the data across time points, thereby collapsing the – in the present case – 32-dimensional data of the EEG to one-dimensional data in discriminating component space. Such a reduction of dimensionality inherently provides a solution to the multi-comparison problem often occurring in EEG studies, where several different electrodes are analysed, sometimes even with results being then selectively reported. Back-projected data can be plotted and analysed in ways similar to conventional ERP analyses. For instance, I would expect to find errors and correct trials to differ around the time of the

ERN and the Pe.

The spatial distribution of the electrode weights can be visualised by calculating a *sensor projection* (Parra, Spence, Gerson & Sajda, 2005; Parra et al., 2002). This sensor projection a shows the amount of activity at each electrode that correlates with the discriminating component:

$$a = \frac{Xy}{y^T y}, \tag{2}$$

with X being the re-arranged sensor data and y the discriminating component. For visualisation purposes, a can be plotted like an ERP topography. I expect to find more activity correlated with the discriminating component at centro-parietal electrodes, given that these are the locations on which the Pe is usually found to be most pronounced.

3.1.2 Results

3.1.2.1 Behavioural data

Participants made primary perceptual decisions with a mean RT of 427 ms. This was reliably faster than data from the same difficulty condition in Experiment 1, $t(33.5) = 2.6,\ p = 0.01$, which was on average 490 ms. Moreover, participants committed on average 17.5% errors, reliably more than the equivalent mean error rate of 12.9% found in Experiment 1, $t(29.2) = 2.2,\ p = 0.03$. These findings hint at a slightly more speed-oriented response strategy in the present experiment compared to Experiment 1.

As in previous experiments, accuracy varied monotonically as a function of subjectively-rated confidence, showing high resolution, as well as high calibration: Participants made errors on 97.1% of trials judged *certainly wrong*, compared to an error rate of 1.4% on trials judged *certainly correct*. This pat-

tern held for individual participants, as revealed by individual Spearman's rank correlations of confidence and accuracy, which were reliable for all participants, and close to perfect, $rs <= -0.94,\ ps < 0.01$.

Figure 30 presents the distributions of confidence judgements as a function of accuracy. Once again, some overlap can be observed, with *probably wrong* as the mode for error responses and *certainly correct* as the mode for correct responses. These findings suggest that participants had good resolution in their metacognitive judgements in the present experiment. This interpretation was furthermore supported by the finding that participants were more confident on correct trials compared to error trials, $M_{cor} = 5.0$ versus $M_{err} = 2.7$, $t(15) = 14.4,\ p < 0.01$.

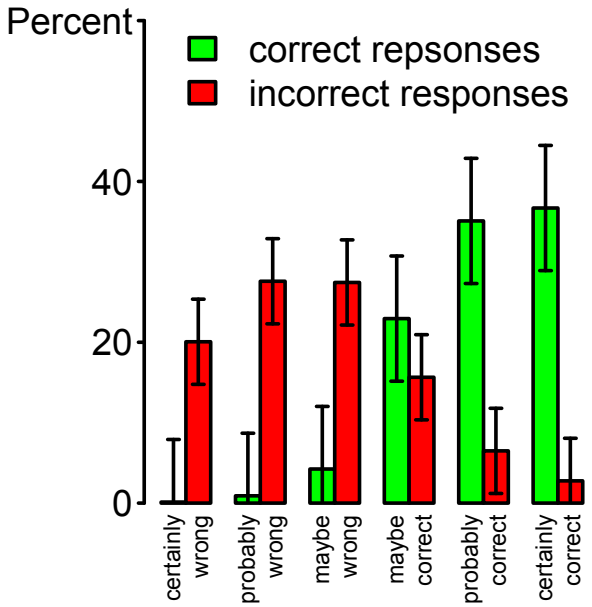

Figure 30: Distributions of confidence responses as a function of objective accuracy.

On average, participants rated their metacognition with a sensitivity of *meta-d'* $= 3.88$. With an average first-order sensitivity of $d' = 2.07$, this resulted in an average metacognitive efficiency of $log(M\text{-}ratio) = 0.28$.

These metacognitive efficiency parameters were significantly different from 0, $t(15) = 5.8,\ p < 0.001$. This means participants used more evidence in their metacognitive judgement compared to the dot decision.

3.1.2.2 ERP data

The key analyses of this study focused on error-related ERPs, that is both the ERN and the Pe. It therefore first had to be determined whether those ERPs were present in the EEG data. Figure 31 shows response-locked EEG activity at electrode CZ as a function of objective accuracy. The left dashed time window highlighted in the figure contained the ERN (-40 to 60 ms). The ERN was analysed with a two-way repeated-measures ANOVA with accuracy and anteroposterior scalp location as factors. Data recorded at five midline electrodes (FZ, FCZ, CZ, CPZ, and PZ) were submitted to the ANOVA. The ERN is usually found to be stronger (i.e., more negative) for error trials compared to correct trials. This was also the case here, $M_{err} = 0.5\ \mu V$; $M_{cor} = 2.1\ \mu V$. This difference was reliable, $F(1,\ 15) = 13.0,\ p < 0.01,\ \eta_p^2 = 0.46$. The ERN usually has a fronto-central, symmetrical topography, which can also be seen in the upper panel of Figure 31. I would therefore have expected to find an effect of location, which was indeed the case, as expressed in a reliable interaction between location and accuracy, $F(1.5,\ 22.4) = 4.2,\ p = 0.04,\ \eta_p^2 = 0.22$. There was furthermore also a main effect of location, $F(1.4,\ 20.8) = 22.7,\ p < 0.001,\ \eta_p^2 = 0.60$.

A similar analysis was conducted for the Pe, which was measured in a time window ranging from 250 to 350 ms, also highlighted in Figure 31. As expected, the amplitude of the Pe was larger for errors compared to correct trials, $M_{err} = 3.9\ \mu V$, $M_{cor} = 0.1\ \mu V$. This difference was also reliable, $F(1,\ 15) = 26.1,\ p < 0.001,\ \eta_p^2 = 0.64$. Moreover, there was a reliable main

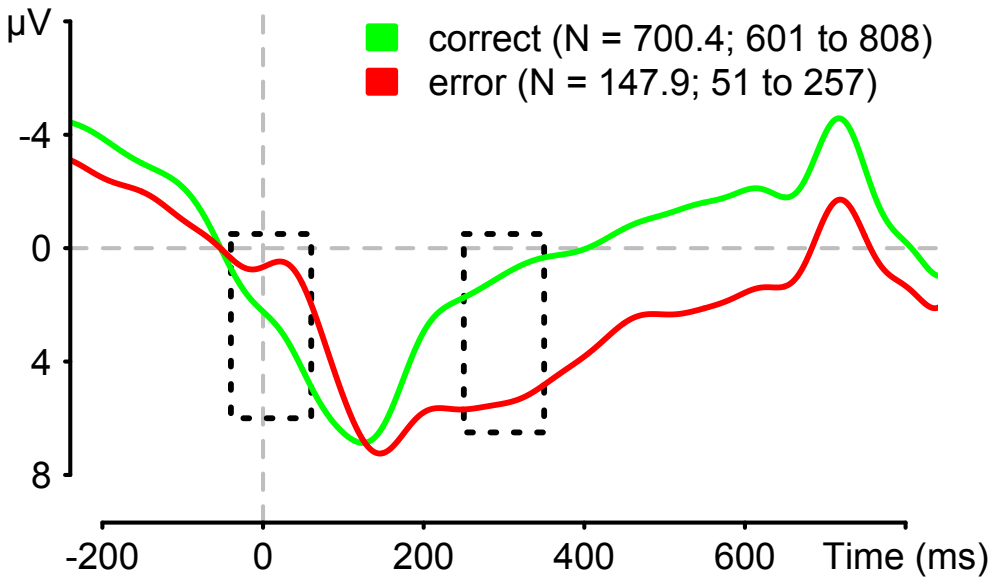

Figure 31: Error-related negativity (ERN) and error positivity (Pe) at electrode CZ, conditioned on objective accuracy; response-locked event-related potential (ERP). The two windows highlight the ERN (-40 to 60 ms) and the Pe (250 to 350 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

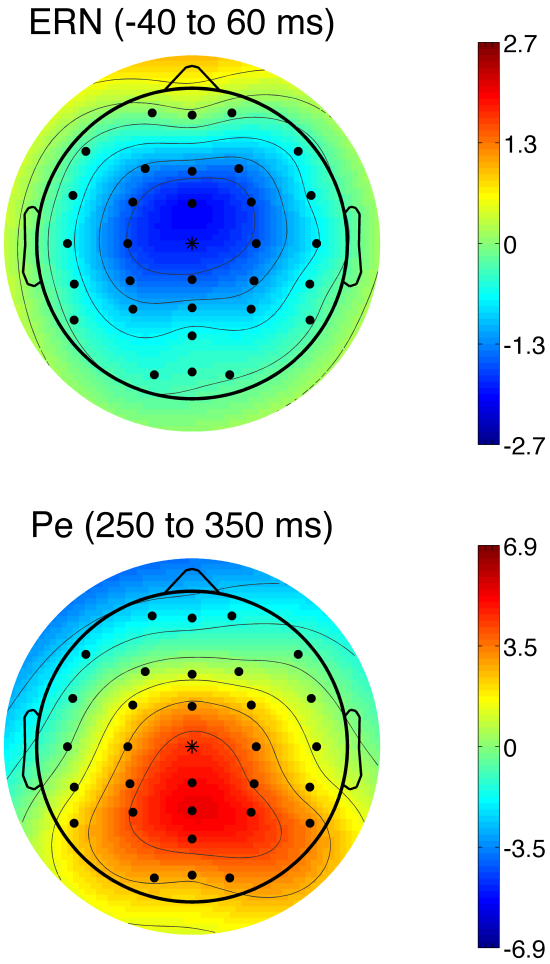

Figure 32: Topographies for the difference between errors and correct trials for both the ERN (top panel) and the Pe (bottom panel). The colours in the topographic plots indicate different values in micro-volt; ms: millisecond.

effect of anteroposterior scalp location, $F(1.3,\ 19.6) = 6.2,\ p = 0.02,\ \eta_p^2 = 0.29$, which was qualified by an interaction with accuracy, $F(1.4,\ 20.3) = 23.6,\ p < 0.001,\ \eta_p^2 = 0.61$, reflecting increased Pe amplitude at posterior sites.

I next investigated how ERN and Pe amplitude varied with confidence, rather than objective accuracy. In the averaged ERP data (Figure 33, collapsed across all trials), the amplitudes of both ERN and Pe components were strongly modulated by decision confidence: For the ERN, there was both a main effect of confidence, $F(1.8,\ 27.7) = 9.2,\ p = 0.001,\ \eta_p^2 = 0.38$, as well as a reliable linear trend, $F(1,\ 15) = 30.6,\ p < 0.001,\ \eta_p^2 = 0.67$. The ERN was strongest (i.e., most negative) for the lowest confidence category. There was once more a reliable effect of anteroposterior scalp location, $F(1.5,\ 22.5) = 19.8,\ p < 0.001,\ \eta_p^2 = 0.57$. The interaction between confidence and location was also reliable, $F(20,\ 300) = 3.0,\ p < 0.001,\ \eta_p^2 = 0.17$. This reflected once more the fact that the ERN effect was strongest at electrode FCZ, $\eta^2_{p(FCZ)} = 0.42$, compared to the other electrode locations, $\eta^2_{p(FZ)} = 0.39$, $\eta^2_{p(CZ)} = 0.36$, $\eta^2_{p(CPZ)} = 0.37$, and $\eta^2_{p(PZ)} = 0.26$. This can be seen in the upper panel of Figure 34, which shows the topography of the ERN for the difference between the two most extreme confidence categories.

Pe amplitude varied systematically with subjectively-rated confidence, $F(2.6,\ 39.0) = 8.2,\ p < 0.001,\ \eta_p^2 = 0.35$, with a reliable linear trend, $F(1,\ 15) = 4.9,\ p = 0.04,\ \eta_p^2 = 0.24$. The amplitude of this component was largest on trials rated *certainly wrong*, and gradually reduced in amplitude as subjective confidence increased. This analysis replicated the reliable effect of location, $F(1.3,\ 20.0) = 6.2,\ p = 0.02,\ \eta_p^2 = 0.29$, and the interaction between location and confidence, $F(20,\ 300) = 11.3,\ p < 0.001,\ \eta_p^2 = 0.43$, which reflected that the Pe effect was strongest at PZ, $\eta^2_{p(PZ)} = 0.53$, and de-

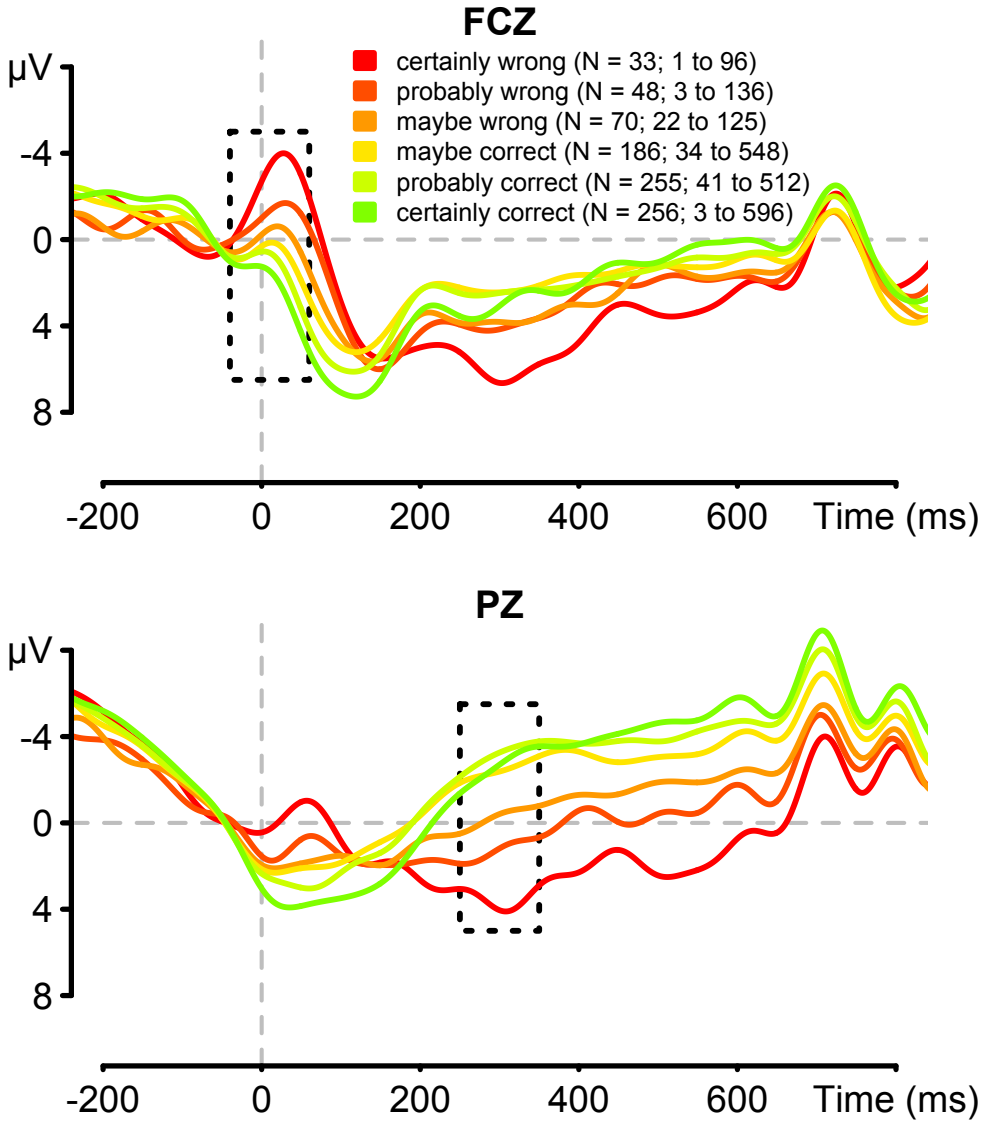

Figure 33: Top panel: Error-related negativity (ERN) at electrode FCZ; bottom panel: error positivity (Pe) at electrode PZ. Both response-locked event-related potentials (ERPs) were conditioned on subjectively-rated confidence. Plots show data combined for objectively correct and incorrect trials. The two windows highlight the ERN (-40 to 60 ms) and the Pe (250 to 350 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

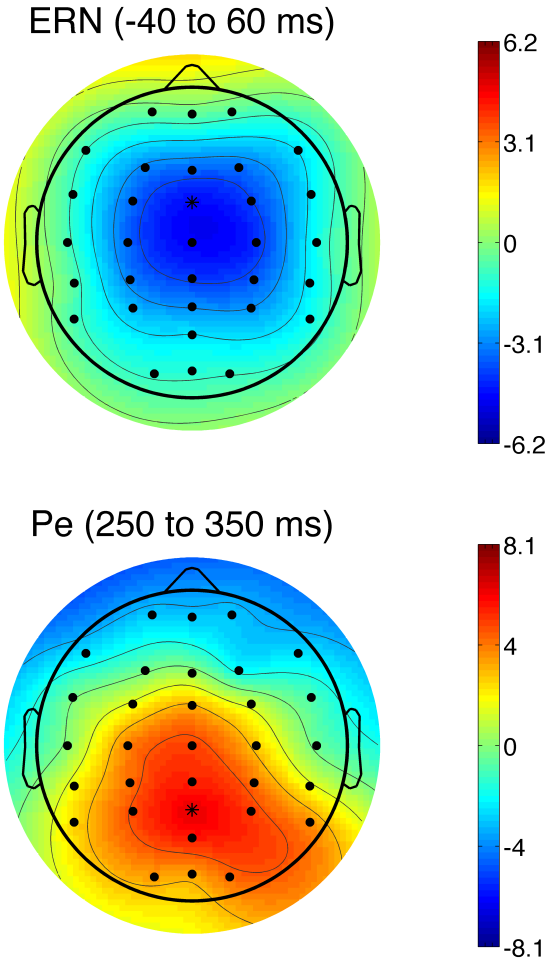

Figure 34: Topographies for the difference between *certainly wrong* and *certainly correct* condition for both the ERN (top panel) and the Pe (bottom panel). The colours in the topographic plots indicate different values in microvolt; ms: millisecond.

creased gradually towards the frontal electrodes, $\eta^2_{p(CPZ)} = 0.51$, $\eta^2_{p(CZ)} = 0.40$, $\eta^2_{p(FCZ)} = 0.18$, and $\eta^2_{p(FZ)} = 0.02$. This can be seen in the topography for the Pe in the lower panel of Figure 34, which again shows the difference between the two outer-most categories.

These results extend previous analyses of the ERN (Scheffers & Coles, 2000) to demonstrate a clear association between Pe amplitude and confidence. However, averaged ERP results are inherently ambiguous about the precise relationship between error-related neural activity and confidence: It could be that amplitude reflects graded variation in confidence across trials, as hypothesised here, but it could also be that error-related neural activity has an all-or-none quality (cf. Charles et al., 2013), with changes in amplitude across confidence bins simply reflecting variation in the proportion of trials on which this activity is triggered (i.e., from very rarely when participants feel *certainly correct*, to almost always when they feel *certainly wrong*). Figure 35 illustrates this alternative explanation: Each stacked bar represents a hypothetical distribution of all choices made at each confidence level, while colour represents the proportion of trials on which error related activity – such as the Pe – was present or absent. The *certainly wrong* category, for example, is hypothesised to have a large proportion of trials on which activity was present, leading to a large overall amplitude if trials are averaged. On the other hand, the *certainly correct* category is hypothesised to contain a small proportion of trials on which error-related activity was present, leading to a small overall amplitude if trials are averaged. This example illustrates how averaging EEG data, as is conventionally done in ERP analyses, can lead to incorrect conclusions regarding the true, underlying data patterns. The next analysis will address this question by analysing data on the single-trial level.

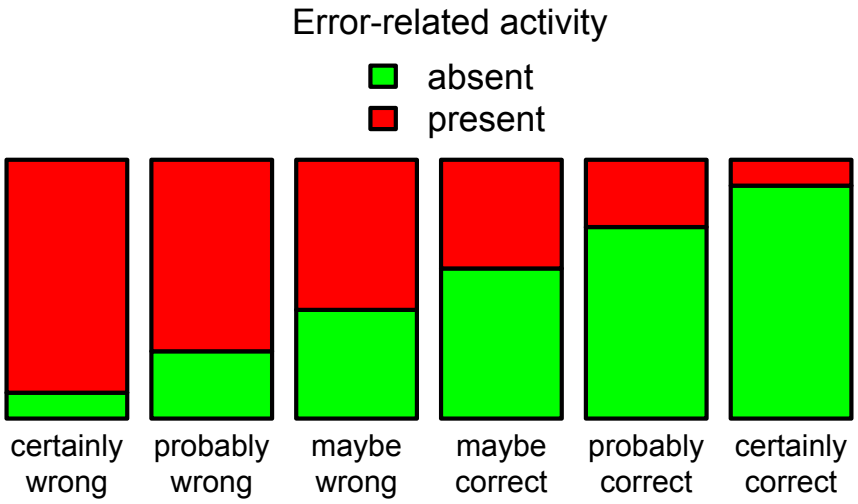

Figure 35: Schematic example of an alternative explanation regarding the systematic variation of confidence and Pe amplitude: The stacked bars represent data from different confidence categories; the colours within these bars represent the proportions of trials in which error-related activity was present or absent.

3.1.2.3 Single-trial EEG data

Gradations in confidence on correct trials. To distinguish between these alternative interpretations – which point to fundamentally different models of the relationship between error detection and decision confidence – multivariate classifier techniques were used to robustly index Pe amplitude on individual trials. Linear integration allowed me to derive a spatial filter (a discriminating component) that maximally distinguishes correct- and error-trial waveforms. The classifier was trained on the set of all error trials and a matched-sized set of correct response trials ($M = 296$ of error and correct responses combined, range 102 – 514).

First of all, the time window of classification had to be determined. I therefore fitted the classifier to 60 100-ms-wide time windows, shifting the current window 10 ms forwards in time to form the next window, as described in

detail in the methods section (Section 3.1.1.4). The classification performance over time windows is presented in Figure 36, that is the AUC of the classifier is applied back to the training set. Replicating previous findings (Steinhauser & Yeung, 2010), optimal classification performance was found using a training window of 250 to 350 ms post-response. For this window, mean single-trial discrimination performance for the 16 participants was very robust: $AUC = 0.83$ ($min = 0.74$; $max = 0.91$), where AUC refers to the area under the curve in an ROC plot with a value of 0.50 expressing behaviour at chance. Overall, AUC values were found to be reliably larger than 0.50, $t(15) = 24.0$, $p < 0.001$. The individual ROC curves are presented in the left panel of Figure 37. Interestingly, this time window coincides with the Pe rather than the ERN, which makes sense considering the former's demonstrated association with subjective error awareness (Overbeek et al., 2005). Nevertheless, below I repeated the same procedure but for the time window of the ERN with the result that classification based on the ERN failed to demonstrate consistent association with single-trial decision confidence.

These ROC curves only reflect how well the classifier fitted the data it was trained on if applied back to this data set. The question remains how consistent these estimates were, that is how well a classifier fitted on a proportion of trials could be generalised to the entire data set for this participant. Given that no errors were left that were not part of the training set, a k-fold cross-classification method had to be used. This means that all test sets were further divided into four non-overlapping folds. The classifier could then be trained on three of these folds and be tested on the remaining fold for all 4 possible combinations. The average AUC for this analysis was 0.71 ($min = 0.53$; $max = 0.85$), which was reliably larger than 0.50, $t(15) = 9.8$, $p < 0.001$. This value is numerically lower than for the training set, but does not affect

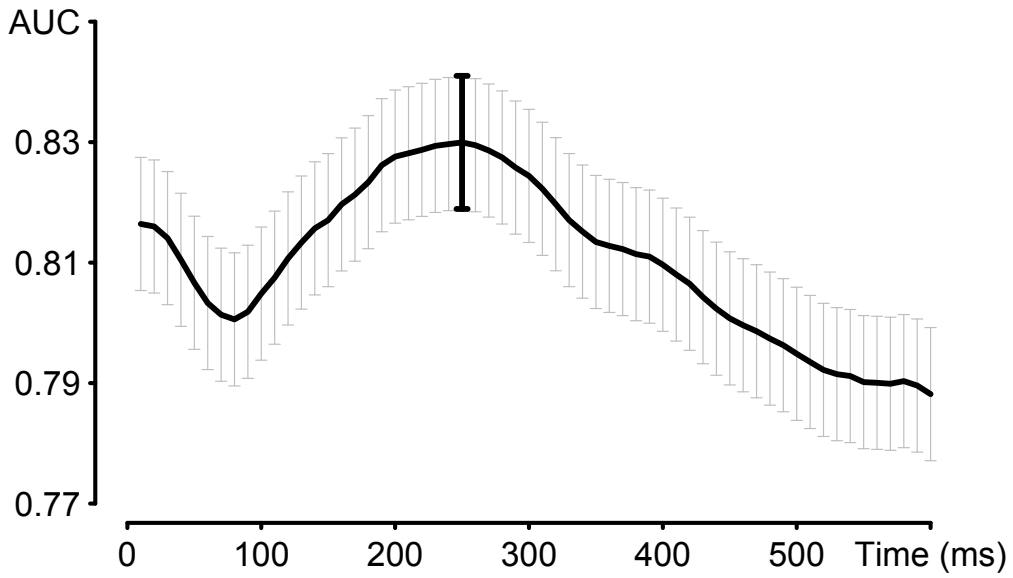

Figure 36: Classification performance (area under curve; AUC) when applied back to the training set for a range different time windows form which the training data was taken. The x-axis specifies the starting point from which onwards a window of 100 milliseconds (ms) width was extracted.

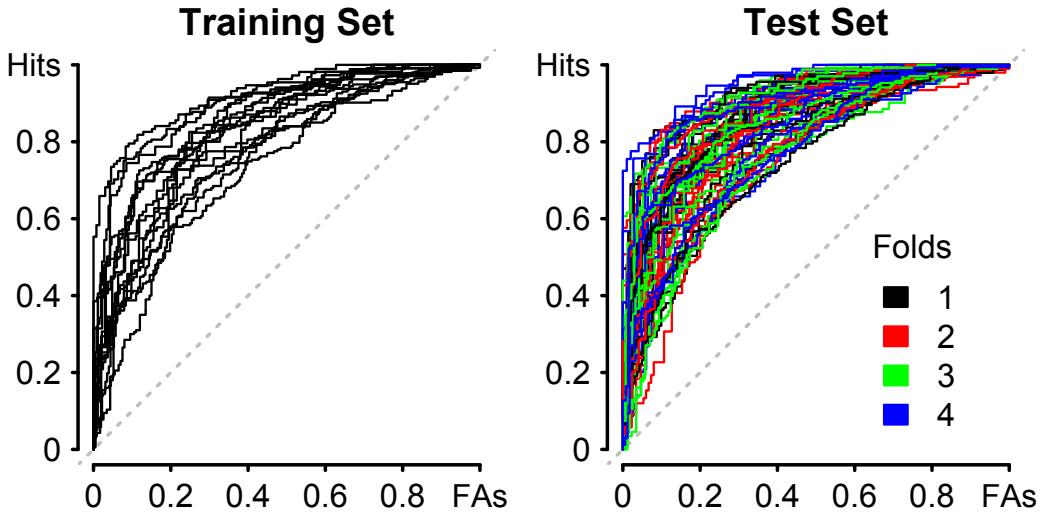

Figure 37: Classification performance expressed in receiver operating characteristic (ROC) curves for classifier trained on Pe time window. Left panel: classifier applied back to the training set; right panel: classifier applied to a new set of data, the test set, using a 4-fold algorithm; FAs: false alarms.

the validity of the analyses given that the cross-classification performance of the classifier was only of secondary interest here. The individual ROC curves for the k-fold cross classification are furthermore presented in the right panel of Figure 37. All the following analyses used the classifiers obtained by training on all the available data, not the averaged classifiers obtained using the k-fold approach.

Figure 38 furthermore presents classification performance over time using the same shifting-window approach described above, but for average 4-fold AUC. Classification performance in this analysis was very similar to the non-cross-validated analysis: The curves in Figures 36 and 38 share an almost identical overall morphology. However, there is more variability across time points in cross-validated AUC, reflecting the lower number of trials on which classification was based, such that peak AUC was observed for a classification window from 220 to 320 ms rather than 250 to 350 ms. However, for subsequent analyses I preferred to use the latter window because it is based on a less noisy estimate and is consistent with prior research (Steinhauser & Yeung, 2010).

The time course of the discriminating component is presented in Figure 39. This component was different for error and correct trials during the time window that was used for training the classifier, $t(15) = 8.1,\ p < 0.001$, as predicted above. Moreover, a second time window was chosen by shifting the window 100 ms forward, given that the Pe is a slow wave that exceeds the training window in time (see Figure 31). For this second time window, the discriminating component also varied between errors and correct responses, $t(15) = 7.2,\ p < 0.001$. Moreover, the sensor projection (Figure 40) of this discriminating component indicated that the extracted component corresponded closely to the Pe (cf. Figures 32 and 34).

The primary question here was whether this classifier, which was trained

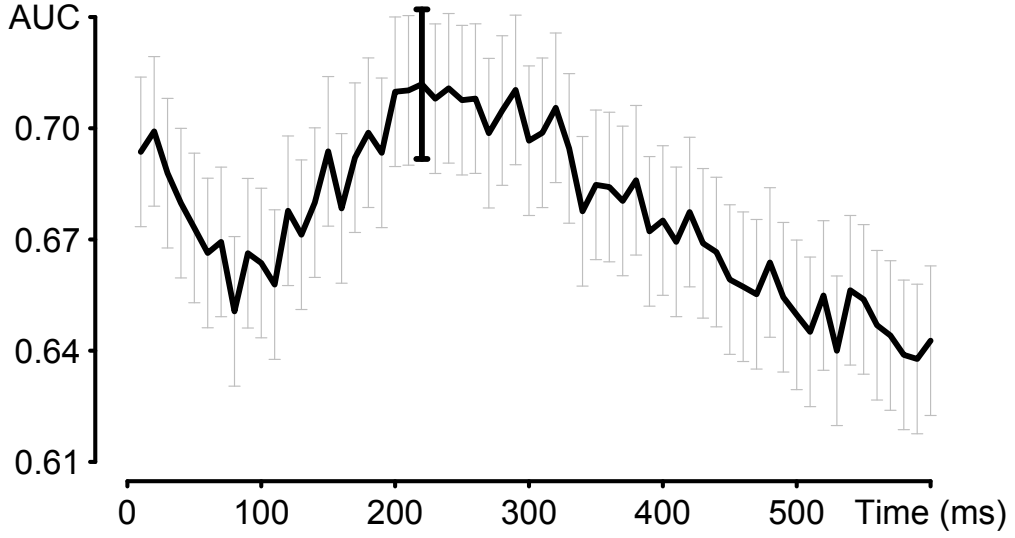

Figure 38: Mean classification performance (area under curve; AUC) using a 4-fold algorithm for a range different time windows form which the training data was taken. The x-axis specifies the starting point from which onwards a window of 100 milliseconds (ms) width was extracted.

to predict objective accuracy, would also predict variation in confidence on correct trials. I therefore applied the classifying component to the response-locked EEG data from the subset of correct trials not used in classifier training ($M = 553$ trials across participants, range 344 – 757), to yield an estimate of Pe amplitude for each time point on each of these trials. The resulting values were averaged across a moving window of 51 ms, and for each time point were then split into quintiles (smallest to largest Pe amplitude). Mean confidence within each quintile was then calculated. The results are presented in Figure 41. They indicate that correct-trial confidence indeed covaried with the amplitude of the discriminating component. This relationship was reliable for the training window (left window highlighted in Figure 41), $F(1.7,\ 25.9) = 5.9,\ p = 0.01,\ \eta_p^2 = 0.28$, with a reliable linear trend, $F(1,\ 15) = 6.7,\ p = 0.02,\ \eta_p^2 = 0.31$. The same held for a time window shifted

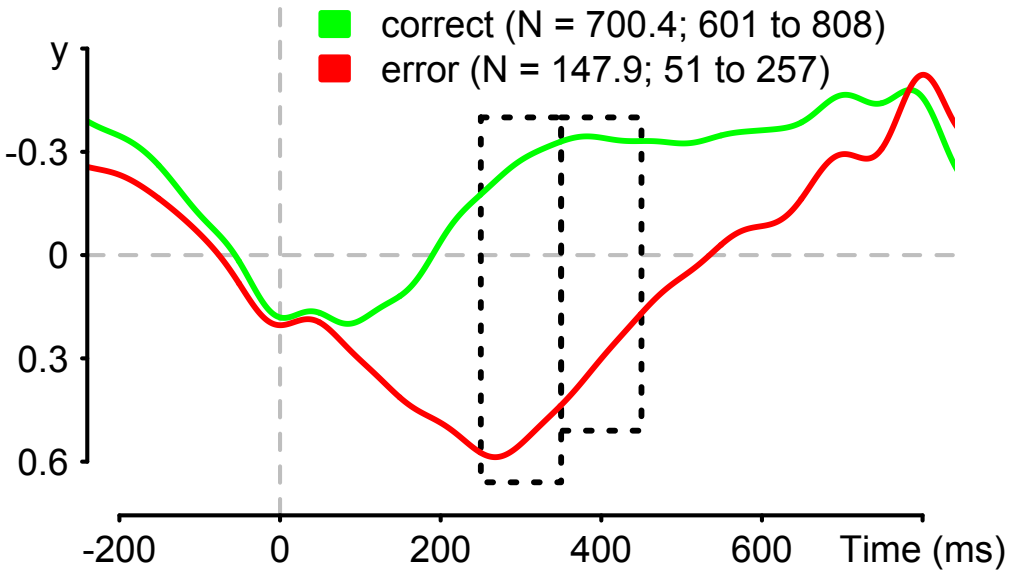

Figure 39: Time course of the discriminating component (y) for the Pe time window, identified by the logistic regression classification analysis of errors versus correct responses, coded in arbitrary units. The two windows highlight the training window for the classifier (250 to 350 ms) and a second window, assumed to capture later parts of the Pe (350 to 450 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond.

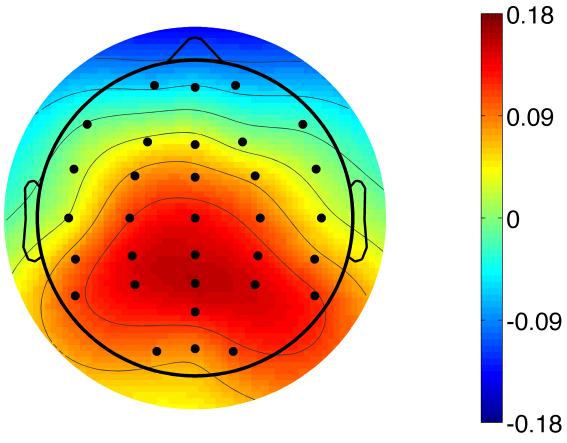

Figure 40: Sensor projection of the discriminating component identified by the logistic regression classification analysis of errors versus correct responses, trained on the Pe time window, coded in arbitrary units.

100 ms so that it began after the training window was over (350 to 450 ms; right window highlighted in the figure), $F(1.6,\ 23.6) = 5.5,\ p = 0.02,\ \eta_p^2 = 0.27$, again with a reliable linear trend, $F(1,\ 15) = 6.8,\ p = 0.02,\ \eta_p^2 = 0.31$.

It can therefore be concluded that confidence varied inversely, and monotonically, as a function of Pe amplitude. Thus, even on trials matched for objective accuracy, Pe amplitude varied in a manner predictive of confidence. Moreover, the resulting gradations in confidence were observed around a high mean value (5 = *probably correct*), suggesting that the information reflected in the Pe not only reflects graded certainty about having made an error (cf. Steinhauser & Yeung, 2010) but also reflects graded certainty of having made a correct response.

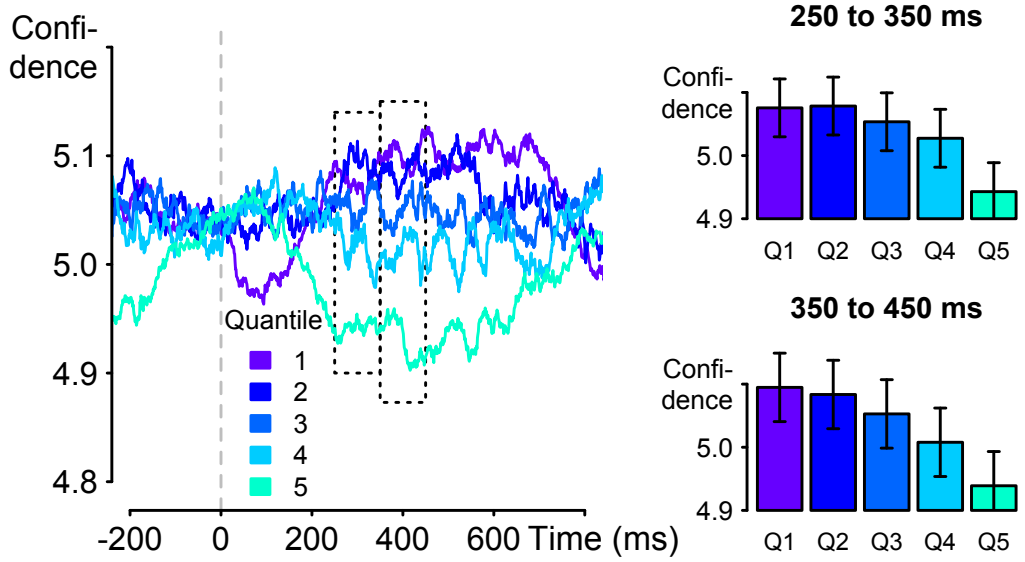

Figure 41: Response-locked, moving average confidence (window width: 51 ms) for quintiles of discriminating component; correct trials only; classifier trained on objective accuracy; ms: millisecond. The two windows highlight the training window for the classifier (250 to 350 ms) and a second window, assumed to capture later parts of the Pe (350 to 450 ms). The two right panels present averaged confidence over the two time windows and quintiles.

Gradations in confidence on correct trials, classified as correct. It remains possible, however, that the described association of confidence and Pe amplitude was driven by changes in the proportion of false error detections – rather than true gradations in correct-trial confidence – across Pe-classifier amplitude quintiles. The next analysis therefore aimed to rule out this possibility. The analysis paralleled the previous one, but now mean confidence was calculated only for hit trials, in which correct responses were rated by the participant as being correct in their secondary confidence judgement. Thus, for this analysis, all trials were objectively correct and subjectively-rated as such. Yet the prediction was that variation in the level of confidence on these trials would follow Pe amplitude. For the training window (left window highlighted in Figure 42), the effect was only marginally significant, $F(1.8,\ 27.7) = 3.0,\ p = 0.07,\ \eta_p^2 = 0.17$, with no reliable linear trend, $F(1,\ 15) = 2.9,\ p = 0.11,\ \eta_p^2 = 0.16$. However, consistent with this prediction, mean confidence varied significantly over Pe-classifier quintiles for these hit trials, $F(1.7,\ 25.6) = 4.1,\ p = 0.03,\ \eta_p^2 = 0.22$, with a significant linear trend, $F(1,\ 15) = 5.2,\ p = 0.04,\ \eta_p^2 = 0.26$ for a time window ranging from 350 to 450 ms, therefore non-overlapping with the training window. This is presented in Figure 42. Taken together, these findings indicate that changes in Pe amplitude are associated with subtle shifts in confidence, with increased amplitude associated with a change in the balance from *certainly correct* judgements to evaluations that responses are only *probably correct* or *maybe correct*.

I furthermore conducted an analysis to test whether the same results would also be found if the data were analysed with a classifier trained on data from the ERN time window rather than for a time window from when the Pe was observed in the data, given that some researchers have linked the ERN to error awareness (Wessel, 2012). As previously shown, the AUC was lower for

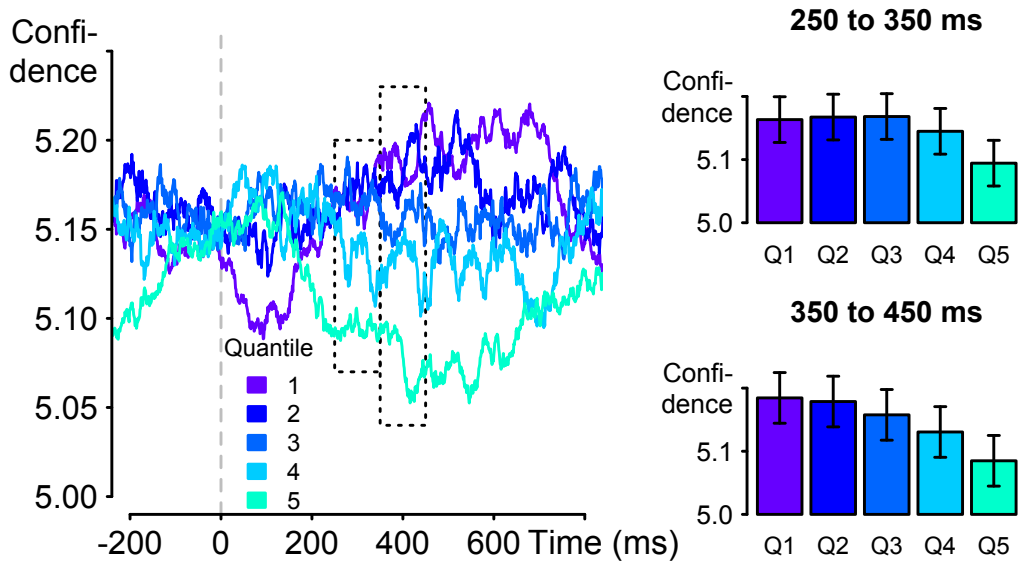

Figure 42: Response-locked, moving average confidence (window width: 51 ms) for quintiles of discriminating component; correct trials classified as *correct* only (i.e., hit trials); classifier trained on objective accuracy; ms: millisecond. The two windows highlight the training window for the classifier (250 to 350 ms) and a second window, assumed to capture later parts of the Pe (350 to 450 ms). The two right panels present averaged confidence over the two time windows and quintiles.

such an ERN time window. However, that does not necessarily mean that the cross-classification analysis would not work, given that this AUC value merely reflects how well the classifier discriminates between correct and error trials if applied back to the test set. The above-described analysis was therefore repeated for a classifier trained on data from a time window ranging from -40 to 60 ms.

Figure 43 shows the ROC curves for this analysis, both for the overall classification performance if applied back to the training set (left panel), as well as for a k-fold classifier applied to the remaining of four folds. For the training set, the AUC values were on average $AUC = 0.80$ ($min = 0.69$; $max = 0.93$). This was reliably lower if compared to the previous classifier trained on data from the Pe time window, $t(15) = 3.1,\ p < 0.01$. For the classifiers derived from the k-fold approach, the average AUC was again slightly lower, $AUC = 0.66$ ($min = 0.54$; $max = 0.88$). No difference in folds should have been expected, given that the data for the folds is selected randomly. Indeed, AUC did not differ between folds, $F(3, 45) = 1.3,\ p = 0.29,\ \eta_p^2 = 0.08$. However, AUC was again reliably lower for this analysis compared to the classifiers trained on data from the Pe time window, $F(1, 15) = 4.6,\ p < 0.05,\ \eta_p^2 = 0.23$. The two factors did not interact, $F(3, 45) = 1.3,\ p = 0.30,\ \eta_p^2 = 0.08$. As for the previous analysis, all the following analyses used the classifiers obtained by training on all the available data, not the averaged classifiers obtained using the k-fold approach.

The time course of the discriminating component is presented in Figure 44 as a function of objective accuracy. Given that the classifier was again trained to distinguish between correct and error trials, it should be expected that the time courses for correct and error trials should differ significantly. This was indeed the case, both for the training window ranging from -40 to

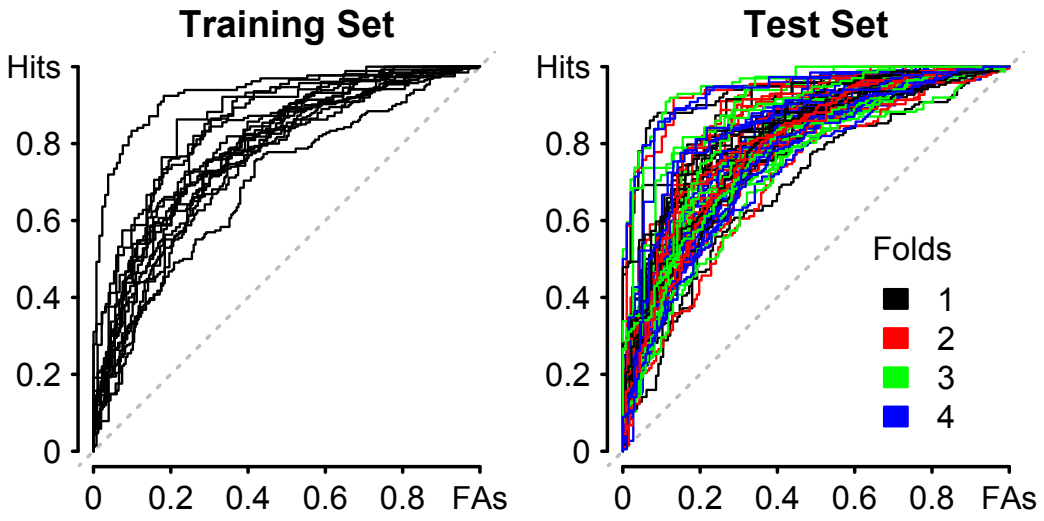

Figure 43: Classification performance expressed in receiver operating characteristic (ROC) curves for classifier trained on ERN time window. Left panel: classifier applied back to the training set; right panel: classifier applied to a new set of data, the test set, using a 4-fold algorithm; FAs: false alarms.

60 ms, $t(15) = 7.0$, $p < 0.001$, as well as for a time window shifted 100 ms forwards (60 to 160 ms), $t(15) = 5.6$, $p < 0.001$. The sensor projection, presented in Figure 45, indicated that the extracted component corresponded to the topography of the ERN (cf. Figures 32 and 34).

As for the previous set of analysis, the obtained classifiers were then applied back to the remaining correct trials, trials which had not been used for training the classifier ($M = 553$ trials across participants, range 334 - 757). The question here was again whether this classifier, which was trained on objective accuracy, would also predict variations in correct-trial confidence. Applying the classifier back to the test sets yielded estimates of the ERN amplitude for each time point. These estimates were again averaged over a sliding window of 51 ms, and then split into quintiles. For each of these quintiles and time points, average confidence was then again calculated. The

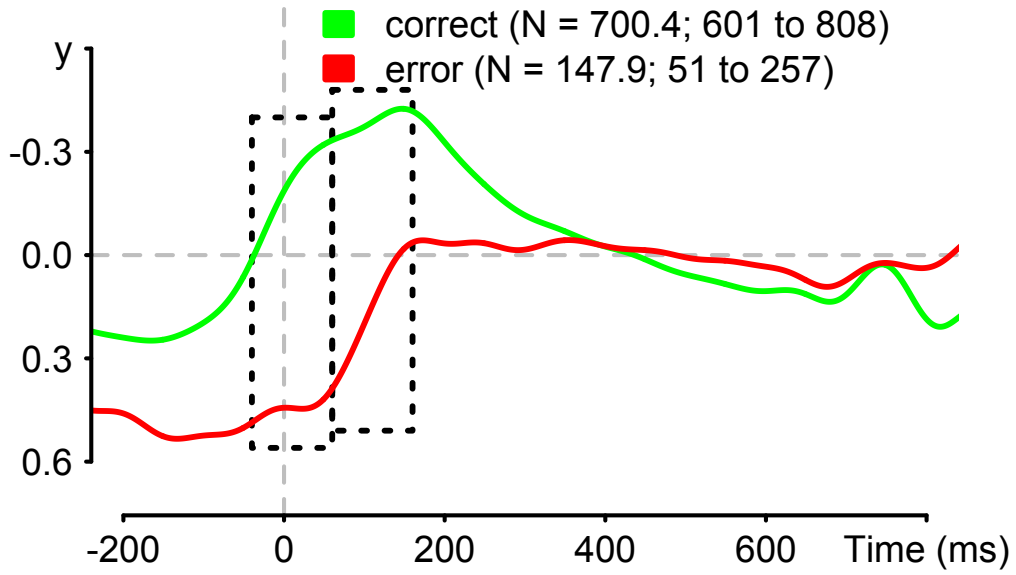

Figure 44: Time course of the discriminating component (y) for the ERN time window, identified by the logistic regression classification analysis of errors versus correct responses, coded in arbitrary units. The two windows highlight the training window for the classifier (-40 to 60 ms) and a second window (60 to 160 ms), to mirror the previous analysis of the Pe-trained classifier. The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond.

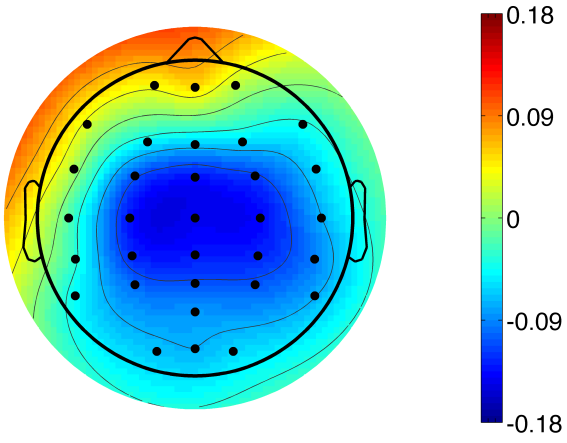

Figure 45: Sensor projection of the discriminating component identified by the logistic regression classification analysis of errors versus correct responses, trained on the ERN time window, coded in arbitrary units.

results from this analysis are present in Figure 46. In contrast to the previous classifier analyses, correct-trial confidence did not covary reliably with the amplitude of the discriminating component, $F(1.3,\ 19.9) = 1.1,\ p = 0.33,\ \eta_p^2 = 0.07$, and there was no reliable linear trend, $F < 1$. For a time window that was shifted 100 ms forwards (60 to 160 ms), the effect of quantile was marginally significant, $F(1.2,\ 17.9) = 3.4,\ p = 0.07,\ \eta_p^2 = 0.19$. There was no reliable linear trend, though, $F(1, 15) = 3.0,\ p = 0.10,\ \eta_p^2 = 0.17$. It can be concluded from these results that a classifier trained on data from the ERN window could not be used reliably to detect fine-graded variations in correct-trial confidence.

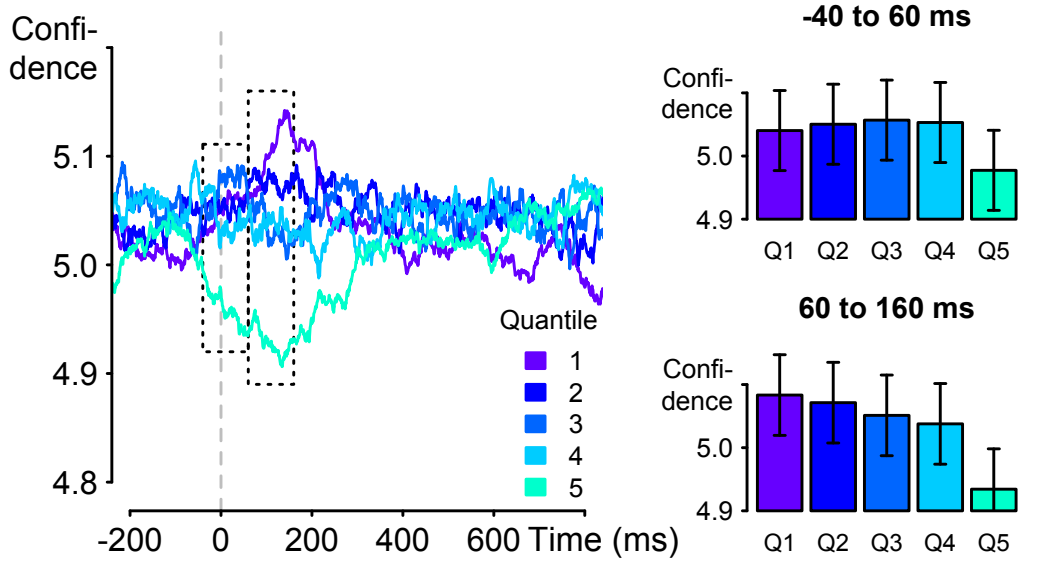

Figure 46: Response-locked, moving average confidence (window width: 51 ms) for quintiles of discriminating component; correct trials only; classifier trained on objective accuracy; ms: millisecond. The two windows highlight the training window for the classifier (-40 to 60 ms) and a second window (60 to 160 ms), to mirror the previous analysis of the Pe-trained classifier. The two right panels present averaged confidence over the two time windows and quintiles.

Analysis of the stimulus-locked P3. The P3, a positive deflection occurring approximately 300 ms after the onset of a task-relevant stimulus, has pre-

viously been linked to confidence (Squires, Squires & Hillyard, 1975a; Kerkhof, 1982; Squires, Hillyard & Lindsay, 1973; Squires, Squires & Hillyard, 1975b; Wilkinson & Seales, 1978; Sutton, Ruchkin, Munson, Kietzman & Hammer, 1982; Hillyard, Squires, Bauer & Lindsay, 1971; Selimbeyoglu, Keskin-Ergen & Demiralp, 2012). The question arises as to whether the effects reported here are not merely a reflection of the sensitivity of the stimulus-locked P3 to confidence. For instance, studies about the P3 have identified two subcomponents (Polich, 2007), and the argument has been raised that the Pe might be a late P3 that has continued through the ERN (Davies, Segalowitz, Dywan & Pailing, 2001; Wessel, 2012). More specifically, the idea has been that errors often happen because people respond before they have fully categorised the stimulus (premature responses). On such error trials, we might expect the P3 to happen after the response, whereas on correct trials the P3 will tend to occur earlier relative to the response.

However, the effects that were observed for the response-locked Pe are the precise opposite of those previously reported for the stimulus-locked P3, which usually has a larger amplitude for more confident responses. To assess how stimulus-locked P3 varied with confidence in this dataset, I quantified the P3 as average voltage in a time window ranging from 350 to 500 ms post-stimulus: Whereas Pe amplitude varied inversely with decision confidence, P3 amplitude correlated positively, $F(2.6,\ 39.5) = 5.1,\ p < 0.01,\ \eta_p^2 = 0.25$, with a reliable linear trend, $F(1,\ 15) = 11.1,\ p < 0.01,\ \eta_p^2 = 0.43$. This dissociation is illustrated in Figure 47, which plots stimulus-locked P3 amplitude as a function of decision confidence (left panel) together with a summary of our Pe results (right panel), $F(3.3,\ 49.3) = 16.1,\ p < 0.001,\ \eta_p^2 = 0.52$, with a reliable linear trend, $F(1,\ 15) = 36.7,\ p < 0.001,\ \eta_p^2 = 0.71$. The Pe results were similar but not identical to the ones shown previously (Figure 33), only that for this new

analysis, I used a pre-stimulus baseline, furthermore ruling out any potential baseline contamination artefact from the stimulus-locked P3. The key results reported here were also apparent in an analysis using such a baseline. All these analyses comparing the P3 and Pe were carried out using data that was filtered with a 20 Hz cutoff instead of 12 Hz, as in previous analyses, to test whether the filter used for all previous analyses did not result in shifting of the here-studied ERP components.

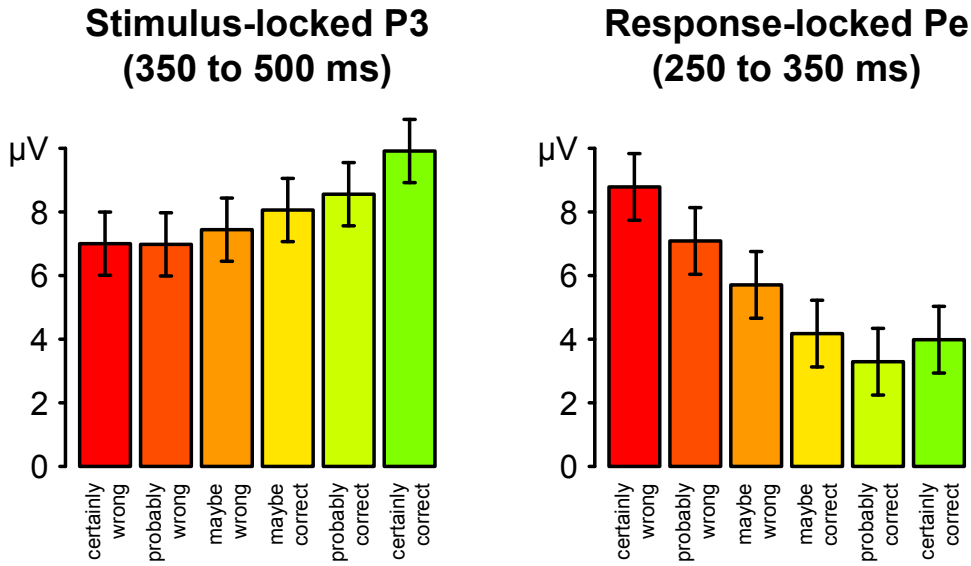

Figure 47: Averaged activity as a function of subjectively-rated confidence. Left panel: stimulus-locked P3 over a time window ranging from 350 to 500 millisecond (ms). Right panel: response-locked Pe over a time window ranging from 250 to 350 ms; μV: micro-volt.

3.1.3 Discussion

The present experiment provides new insight into the neural mechanisms of metacognition and the relationship between error monitoring and decision confidence. The findings indicate that the Pe, an EEG index of error processing, also varies with decision confidence on correct trials. Thus, a Pe classifier

trained to discriminate between objectively correct and incorrect trials was predictive of fine-grained differences in correct-trial confidence on single trials. Crucially, this association was not driven by changes in the proportion of trials classified as errors across confidence ratings, but rather reflected truly graded change in correct-trial confidence: Pe amplitude was predictive of subtle shifts in confidence (e.g., from *certainly* to *maybe*) on trials that were objectively correct and accurately judged so by participants.

Furthermore, I tested whether the reason why Pe amplitude was found to vary with confidence was merely a reflection of the P3 being sensitive to decision confidence (Hillyard et al., 1971) with the result that the stimulus-locked P3 exhibited precisely the opposite relationship to the one shown by the Pe: increasing in amplitude as confidence increased. These results demonstrate a clear functional dissociation between the P3 and Pe. In additional analyses, it was confirmed that the inverse relationship between Pe amplitude and confidence was replicated using pre-stimulus rather than pre-response baseline, thus ruling out any potential baseline contamination artefact from the stimulus-locked P3 effect.

The observed association of Pe amplitude with both error detection and decision confidence is consistent with the proposal that these two metacognitive evaluations reflect similar underlying mechanisms (Yeung & Summerfield, 2012, 2014). In prior work on error detection, binary error judgements have been studied, often with concurrent EEG recording and instructing participants to press a button when they think they have made an error. In the memory and perceptual decision-making literature, on the other hand, the focus has typically been on graded judgements of confidence or certainty, which ask participants to rate how correct they think they were in the decision that they just made. These two lines of research have addressed similar questions

with similar methods, but have rarely been linked. The findings from the present experiment suggest a strong link between neural correlates of error awareness and decision confidence, with substantive implications for current theories in the respective fields. First, linking decision confidence to well-characterised EEG correlates of error processing should place usefully strong constraints on emerging theories of the neural basis of metacognitive monitoring (Fleming & Frith, 2014). Second, an association between confidence and errors presents a significant challenge to many current models of decision confidence, which explain graded evaluations of correct-trial confidence but which struggle to account for error judgements: Many theories propose that confidence judgements are formed at the time of the primary decision, yet error judgements are known to depend on continued processing of stimulus and response information after the initial decision (Yeung & Summerfield, 2012, 2014).

Meanwhile, in research on error monitoring there has been debate over whether error detection is better characterised as a discrete, all-or-none process (e.g., Falkenstein et al., 1991; Wessel, 2012) or a graded continuous judgement (e.g., Scheffers & Coles, 2000; Steinhauser & Yeung, 2010). The findings reported here strongly support the latter hypothesis, and extend it to show that the continuum of error certainty is continuous with fine-grained judgements of certainty that a response is correct: On the 6-point confidence scale used in the present experiment, correct-trial confidence varied around a high mean value (5.0) with relatively small standard deviation (0.8). Correspondingly, I observed very fine-grained variation in correct-trial confidence as a function of Pe amplitude, spanning a relatively narrow range of values clustered around high confidence judgements (quintile range of 4.9 – 5.1).

As such, the present findings bear on the question of whether neural

correlates of error processing vary discretely or continuously, and help to resolve ambiguities in prior research on this question. Wessel (2012) suggested that the late Pe reflects error awareness as an all-or-none process, and Charles et al. (2013) have made a corresponding argument for the ERN based on evidence that this component is observed only on trials in which the stimulus is consciously perceived. Conversely, Scheffers and Coles (2000) reported systematic variation in ERN amplitude with confidence. In this context, it is noteworthy that I found the ERN and Pe both to vary in amplitude with subjective confidence, but only the Pe was predictive of graded changes in confidence across trials in our multivariate analysis. This difference between ERN and Pe might simply reflect greater amplitude and SNR for the latter component (even though the ERN is robustly measurable on individual trials; Parra et al., 2002). However, a more intriguing possibility is that ERN amplitude fails to predict variation in confidence on single trials because it is an all-or-none signal (cf. Charles et al., 2013), and that the observed association with confidence seen in averaged ERPs (Figure 3A; Scheffers & Coles, 2000) reflects variation in the probability that this all-or-none signal is triggered across trials with differing levels of confidence. If correct, this interpretation suggests a reconciliation of previously contradictory findings regarding the ERN. Regardless, the findings indicate that the Pe is a stronger correlate of error awareness (Nieuwenhuis et al., 2001; Hester et al., 2005; Steinhauser & Yeung, 2010) and can simultaneously index associated variation in decision confidence. The findings reported here also have practical implications in suggesting that Pe amplitude can provide a robust 'non-invasive' index of participants' confidence. In Experiment 5, reported in Chapter 4, I therefore use single-trial Pe amplitude as an unobtrusive measure of confidence.

3.2 A single-route model of decision confidence and error detection

The work shown in this thesis so far has led to the conclusion that confidence and error detection may be expressions of the same underlying metacognitive processes. However, the question remains as to precisely how both types of judgements might be generated by the same internal mechanism. This will be the key question addressed in the second half of the present chapter. For instance, one could assume that confidence is a graded signal derived by comparing evidence in favour of the chosen response option and the unchosen response option – a balance-of-evidence mechanism similar to the ones discussed in Chapter 1. This graded signal could then be mapped onto a continuous or categorical confidence scale (Moreno-Bote, 2010), and can even be dichotomised whenever binary error judgements are needed. In its simplest possible form, such a mechanism assumes a single processing route in which evidence is accumulated which then leads to both the first-order decision and the confidence judgement. In the present section, I report results from simulations, showing that such a simple account of decision making can serve as an adequate starting point for modelling the formation of metacognitive judgements.

3.2.1 Basic confidence mechanisms in a sequential sampling model

Several computational models of confidence have already been discussed in this thesis, many of them assume evidence accumulation towards a decision threshold as a core mechanism. These models can be roughly divided into two separate classes: (drift) diffusion models and race models, which differ mainly

in the assumed correlation between the neuronal integrators (Bogacz, Brown, Moehlis, Holmes & Cohen, 2006). Diffusion models assume that balance of evidence is tracked over time, that is the difference in evidence between the two response options. Assuming such difference coding has the advantage of being more parsimonious, in the sense that only one evidence counter is required as opposed to two. However, the disadvantage of such a model is that the overall balance of evidence at the time the decision threshold was reached – often assumed to reflect confidence (Vickers & Packer, 1982) – is irretrievably lost. Alternative solutions have been suggested as to how confidence can result from a diffusion model. Pleskac and Busemeyer (2010), for instance, assume post-decision processing in their 2DSD model, with an interpretation of post-decision evidence relative to several confidence boundaries. Here, however, I focus on the more intuitive balance-of-evidence assumption as the basis of confidence, and therefore assume a race model, that is a model with two evidence counters, which have to be at least partially uncorrelated. Examples of such models can be found in Vickers and Packer (1982) and Merkle and Van Zandt (2006), as discussed in Section 1.4.

The decision model that will be proposed in the present chapter also relies on such a balance-of-evidence mechanisms in the context of a race model, more precisely post-decisional balance of evidence: Two consecutive decision thresholds are assumed, similar to the model by Van Zandt and Maldonado-Molina (2004). The first threshold constitutes the response threshold, while the second elicits the metacognitive judgement. Different to the model by Van Zandt and Maldonado-Molina (2004), however, is that reaching this second threshold also means that a trial is classified as conscious. Confidence is assumed to be a function of the balance of evidence at the time of the second threshold or at a fixed deadline, whichever happens first. This mechanism

gives rise to a graded metacognitive signal, which can then be transformed into graded confidence judgements, as well as binary error detection judgements. My model therefore assumes that both types of judgements rely on the same mechanism. Results from simulations will be reported that qualitatively mimic the findings reported for the dot-count decision task. To foreshadow, simulated correct RTs and error rates vary as a function of difficulty, with changes in difficulty manipulated through changes in drift rate. Moreover, the model produces lower confidence signals for error trials compared to correct trials, but with a substantial overlap between the confidence distributions, as previously observed. Also consistent with previous findings, this overlap is smaller for easy compared to difficult conditions. These findings suggest that the model was able to simulate the confidence resolution observed in human participants.

However, arguing that stimulus discriminability affects both response speed and accuracy is by no means new and has been shown and discussed in the context of many different sequential-sampling models (for an overview, see Ratcliff & Smith, 2004; P. L. Smith & Ratcliff, 2004). The novel contribution of the confidence model described here will instead be to show how different levels of stimulus discriminability affect metacognitive insight, that is the difference between confidence on correct and error trials. I would assume that both the original model by Vickers and Packer (1982) as well as the changes-of-mind model by Van Zandt and Maldonado-Molina (2004) would already have produced similar data patterns, however, neither of these previous studies focused on this measure. Here, I argue that when studying confidence, it is crucial to not only focus on average levels of confidence but instead to extend the analysis to confidence as a function of objective accuracy. I therefore also present simulated distributions of confidence responses and I compare these to

empirical distributions reported in the context of EXPERIMENT 1 (as shown in Figure 11).

3.2.2 Extension of the model to also include metacognitive insight for unconscious decisions

A recent study by Charles et al. (2013) presented evidence in favour of the hypothesis that people have metacognitive insight even if they report that they were not consciously aware of the evidence that led to their primary decision. More specifically, the authors presented participants briefly with a number stimulus which was subsequently masked, as illustrated in Figure 48. The stimulus-onset asynchrony (SOA) between the stimulus and the mask was varied so that the amount of evidence that entered the visual system could be manipulated. After responding to the stimulus, participants had to also rate whether they saw the stimulus or were just guessing, as well as whether or not their response was correct or incorrect. One of the key findings, which have already partly been discussed above, was that even if participants reported a stimulus as *unseen*, they still showed weak metacognitive insight. They moreover found that subjective visibility ratings as well as objective accuracy were affected by the SOA difficulty manipulation, as were metacognitive ratings.

Varying visibility by means of such an SOA manipulation constitutes a qualitatively different manipulation of difficulty compared to the above described manipulation of discriminability: Rather than assuming changes in overall accumulation rate, a manipulation on SOA results in dynamic changes in drift rate over the course of a single decision. In other words, it is assumed that stimulus information only affects evidence accumulation during

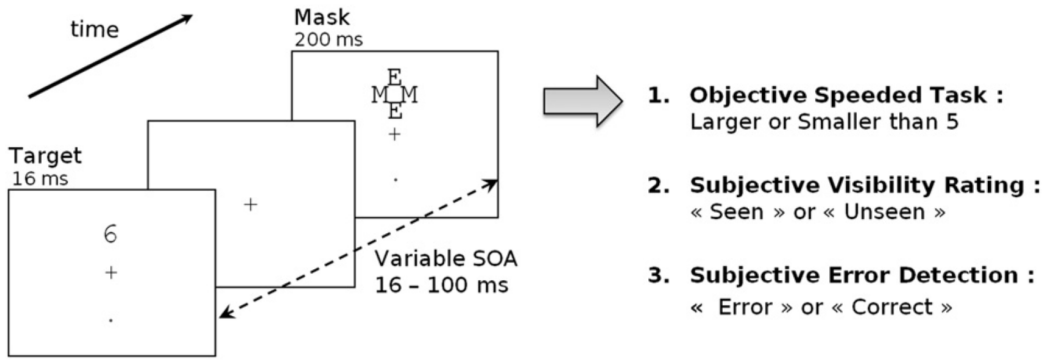

Figure 48: Figure reproduced from Charles et al. (2013): details of the decision paradigm, which included a manipulation on visibility by varying the duration (stimulus-onset asynchrony; SOA) between the presentation of the stimulus and a mask. After each choice, participants first classified the stimulus as *seen* or *unseen* and then the decision as an *error* or a *correct* trial; ms: millisecond.

early stages of the decision, whereas after onset of the mask, the decision process is governed only by noise. I therefore test whether my confidence model can also simulate the effects caused by such a manipulation, focusing on changes in accuracy, visibility and metacognitive insight.

The confidence model will therefore have to be able to distinguish between conscious and unconscious trials. This can be incorporated by the same two thresholds from which responses and metacognitive judgements arise: If the model reaches the second threshold before a pre-specified deadline, then this trial is consequently classified as conscious. If this second threshold is not reached within the pre-specified time window, the trial is reported as *unseen*. The main difference between my model and the model proposed by Van Zandt and Maldonado-Molina (2004), is therefore the implementation of such a consciousness threshold.

My model will address yet another challenging finding reported by Charles et al. (2013): On trials classified as *unseen*, metacognitive insight was lower compared to trials classified as *seen* and importantly, this effect was in-

dependent of SOA. To foreshadow, the model captured such a 'jump' in metacognitive performance dependent on subjectively-rated visibility. This finding is particularly interesting, given that Charles et al. (2013) suggested that such a data pattern could only be simulated with a *dual-route* model, similar to the one proposed by Del Cul et al. (2009, the full model description can be found in the supplementary material). In contrast, the model presented in this chapter could be described as a *single-route* model, meaning that there is only one set of counters – one for each choice alternative – which accumulate evidence over time. Within this same system, decisions and awareness depend on different criteria, whereas a dual-route model assumes that evidence is accumulated in two distinct routes. One of these routes is fast but noisy – the unconscious response route – while the other is slow but more precise – the higher-level conscious decision route. A dual-route model therefore consists of two separate sets of racing counters. A decision is elicited whenever a counter in one of these routes reaches a decision boundary. Errors are detected whenever there is a mismatch between the two routes. I focus more on the differences between a single- and a dual-route account of these metacognition and visibility data in the discussion, after presenting results from my simulations.

Importantly, I simulate only the behavioural results reported by Charles et al. (2013), whose main focus was on the absence of the ERN on unconscious trials. My aim is not to simulate these ERN findings here, but I will get back to these in Section 3.2.5. A key point made by Charles et al. (2013), however, is that their data suggest a dual-route model and therefore a challenge to any single-route theory like the one proposed here. So the question is whether my single-route model would be able to explain their findings, specifically by assuming a two-threshold model.

3.2.3 Model architecture

Figure 49 shows eight example decisions that illustrate the architecture of the single-route model of decision confidence. Those decisions are at the same time examples of the eight different trial types, being a combination of whether or not a trial was correct or incorrect, classified as correct or incorrect by the model, and whether the trial was reported as seen or unseen by the model. The latter factor, visibility, will be discussed in the second set of simulations (Section 3.2.4.2).

The model is first used to simulate data from the dot-count task used in all of my previous experiments. Two accumulators, which are completely independent meaning that they do not inhibit each other, integrate evidence in favour of the two response alternatives, in this case that the larger number of dots is presented on the left or on the right of the screen. A choice is elicited if evidence in one of the counters exceeds the response, or first-order threshold τ_1. Importantly, the race continues after the counter reached this threshold towards the second-oder threshold τ_2, at which a metacognitive judgement is given according to the current balance of evidence between the two counters. If the second threshold has not been reached within a prespecified time, however, confidence is assumed to be the balance of evidence at the time at which the race process was aborted. In the present section, I describe in detail the key features of the model.

In the two separate evidence counters, evidence is accumulated at every time step. One time step corresponds to a millisecond in this model. At any point in time, a noisy sample at the visual input $s_i(t)$, is created by adding the current visual input for the respective counter, $\upsilon_c(t)$, to Gaussian noise with a standard deviation of σ_i,

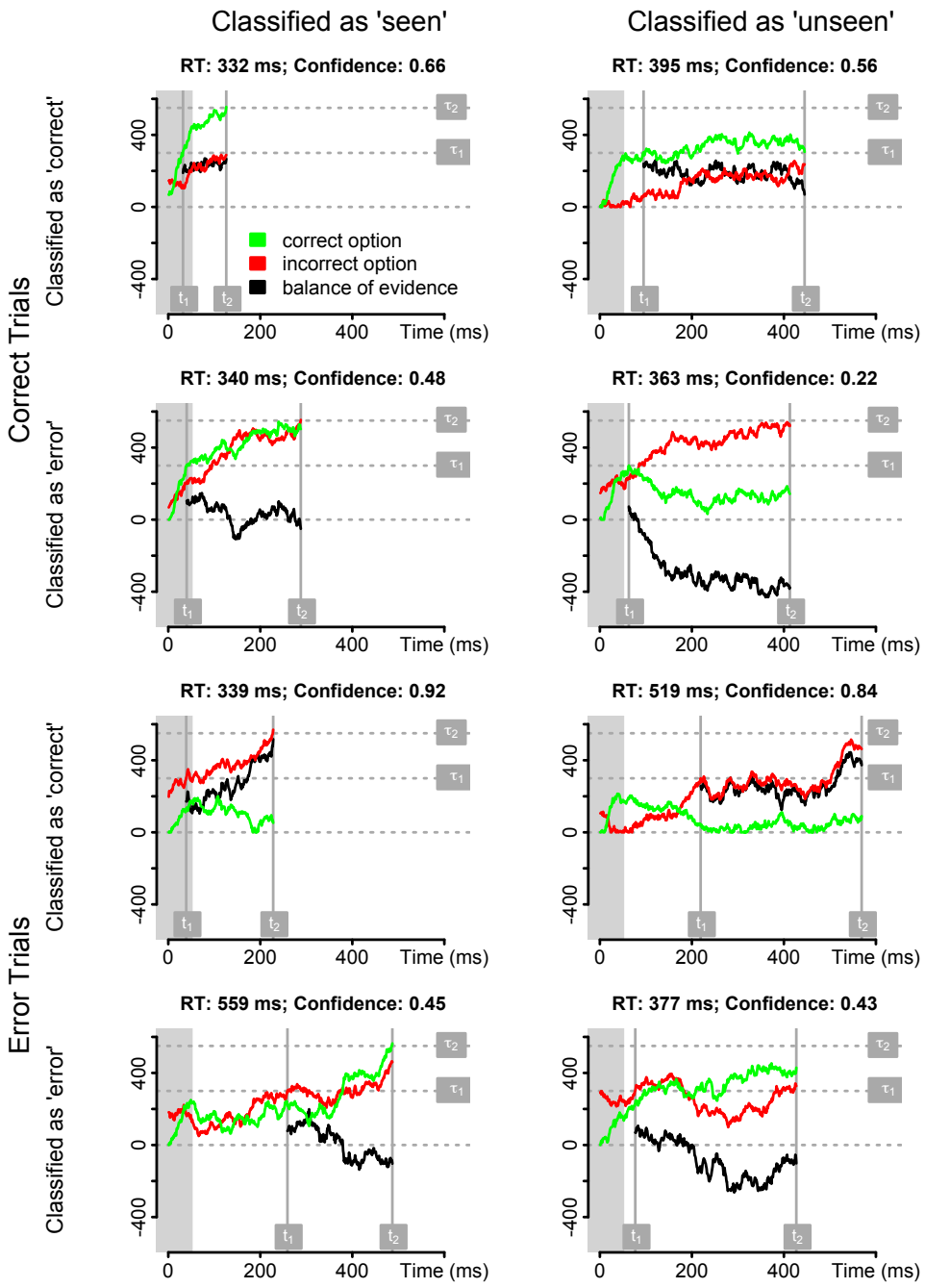

Figure 49: Eight possible trial types follow from the confidence model, depending on whether the trial was classified as *seen* (left panels) or *unseen* (right panels), whether the trial was objectively correct (upper panels) or incorrect (lower panels), and also depending on whether the trial was subjectively classified as *correct* (first and third row) or *incorrect* (second and bottom row). The traces of accumulated evidence (coded in arbitrary units) are displayed in green and red; balance of evidence in black. Grey shaded areas: time during which input from the visual stimulus was present; ms: millisecond; τ_1: first-order threshold; τ_2: second-order threshold; t_1: time at which τ_1 was reached; t_2: time at which τ_2 was reached or trial was aborted; RT: response time.

$$s_i(t) = v_c(t) + N(0, \sigma_i). \tag{3}$$

For the examples shown in Figure 49, the stimulus is shown for 50 ms, highlighted by the grey area. This noisy stimulus sample is then used to create an evidence sample for each counter separately, where it is contaminated again by normally distributed noise with a standard deviation of σ_r,

$$s_c(t) = s_i(t) + N(0, \sigma_r). \tag{4}$$

These evidence samples are then integrated within each counter over time,

$$x_c(t) = x_c(t-1) + s_c(t). \tag{5}$$

Most participants are right-handed, which often results in faster and more frequent responses with the right hand. This bias was also incorporated into the model. For the counter corresponding to the left hand, the starting point is normally distributed around zero with a standard deviation of one third of the response threshold,

$$z_l = N(0, \frac{\tau_1}{3}), \tag{6}$$

whereas for the right hand counter it is slightly biased with a Normal distribution around half of the response threshold and again a standard deviation of one third of the response threshold,

$$z_r = N(\frac{\tau_1}{2}, \frac{\tau_1}{3}). \tag{7}$$

Evidence accumulation continues like this in both counters. Moreover,

the evidence time series are clipped at 0. In other words if the evidence value in a counter is negative, it is set to zero for this time step (behaving like a reflecting lower boundary). A response is emitted once one of the counters reached the response threshold, or first-order threshold τ_1. In the examples in Figure 49, τ_1 is represented by the horizontal dotted line in the middle. The RT is then registered as $RT = t_1 + T_{er}$, where T_{er} denotes the non decision time and t_1 the time at which the first threshold was reached. In Figure 49, t_1 is the left vertical line in each subplot. The upper four panels in Figure 49 show examples in which the counter representing the correct response option reached the response threshold first, whereas the lower four panels present error trials in which the other, incorrect counter won the race.

After reaching the first threshold, evidence accumulation continues towards a second threshold, the *second-order threshold*, τ_2. In the examples in Figure 49, τ_2 is represented by the top-most horizontal dotted line. The trial would terminate if one of the counters reached this threshold (left panels in Figure 49), or if a deadline was reached at 350 time steps after the response (right panels in Figure 49). In Figure 49, this time point is highlighted as t_2, the right vertical line in each subplot. Only if the second threshold τ_2 has been reached the trial was classified as *seen*. Confidence is then defined as a function of the balance of evidence at the time when the second threshold or deadline was reached (t_2). This balance of evidence is calculated relative to the chosen option, which explains why in Figure 49, the black line representing the balance-of-evidence time course only starts after the vertical line denoting the response. The relative balance of evidence is calculated using an adapted version of the formula proposed by Merkle and Van Zandt (2006),

$$Confidence = \frac{x_{win}(t_2)}{x_{win}(t_2) + x_{lose}(t_2)}, \tag{8}$$

where $x_{win}(t_2)$ denotes the amount of evidence accumulated in the winning counter (corresponding to the chosen response option) at time t_2, that is when the second threshold or the deadline was reached. This results in confidence being scaled to a range from zero to one with 0.5 representing a cutoff between trials classified as errors (see negative balance of evidence in Figure 49) and corrects. In other words, the balance of evidence is negative when the order of the counters has changed. This leads to a confidence value smaller than 0.5 and represents a detected error or a change of mind. This model therefore assumes a stable error signalling criterion, despite recent findings that participants are able to flexibly adjust their criterion (Steinhauser & Yeung, 2010). However, for now I would like to argue that 0.5 ought to be fixed as a neutral, but arbitrary cutoff. Future versions of the model can explore changes in the error-signalling criterion.

3.2.4 Simulations

A good model of confidence should exhibit several crucial features of empirically observed data. First, I would expect to find that the model simulates basic effects of difficulty on first-order performance: higher error rates and increased RTs for more difficult compared to easier trials. Second, with regard to second-order performance, I would expect the model to simulate larger differences in correct- to error-trial confidence for easier trials compared to more difficult ones. Both first- and second-order effects are addressed in Section 3.2.4.1. Third, I would expect the model to also show sensitivity to an SOA manipulation, resulting in changing proportions of trials classified as either *seen* or *unseen*. More specifically, I would expect to find that longer masking times lead to decreased error rates and higher proportions of trials classified as *seen*, similar to the findings by Charles et al. (2013). Fourth, I expect

differences between correct- and error-trial confidence to be larger for longer masking times. As a fifth and final prediction, I expect the model to mimic weaker metacognitive insight for trials classified as *unseen*, as also reported by Charles et al. (2013). The latter effects will in turn be addressed in Section 3.2.4.2.

3.2.4.1 Basic first-order and confidence effects

Two simulations were run to test whether the model would be able to mimic the effects on RTs, error rates, and confidence given two different difficulty manipulations – the difference in dots and the above-mentioned SOA manipulation. First, the dot-count manipulation will be discussed. To simulate first-order behaviour from the findings of EXPERIMENT 1, ten different difficulty levels were chosen. All the parameter settings reported here resulted from hand-fitting the model, given that only qualitative patterns were analysed and compared. Those ten levels of difficulty were simulated by varying the visual input $\upsilon_c(t)$, which is the accumulation rate in this model. Rates for ten different levels of difficulty were chosen, increasing quadratically from 0.80^2 to 3.05^2 with the base increased in steps of 0.25. As in the previous experiments, the stimulus was shown for 160 ms. During this period, the stimulus input was set to the accumulation rate for the objectively correct response alternative. The other, objectively incorrect alternative was set to a small drift rate of 0.10. After this time window, the stimulus input was set to zero for both counters. A maximum of 4,000 time steps was simulated after which time the trial was aborted. The standard deviations for the normally distributed noise at input and response level were set to $\sigma_i = 0.1$ and $\sigma_r = 10$, respectively. The thresholds were set to $\tau_1 = 300$ and $\tau_2 = 550$ and a constant non-decision time of $T_{er} = 300\ ms$ was added to all resulting decision times to form the RTs. For

each of the ten levels of difficulty and the objectively correct response option (out of 2 response alternatives), 2,000 trials were simulated. This resulted in a total of 40,000 simulated trials.

Figure 50 presents how error rates and correct RTs varied with simulated difficulty. Faster RTs in easier conditions follow from the fact that higher drift rates allow the counters to reach the response threshold τ_1 faster compared to lower drift rates. Higher drift rates also increase likelihood that the incorrect counter reaches the response threshold τ_1 first, thereby resulting in lower error rates in easier conditions. This Figure closely resembles the data pattern found for human participants (Figure 8). The model was therefore successful in simulating the general first-order data patterns.

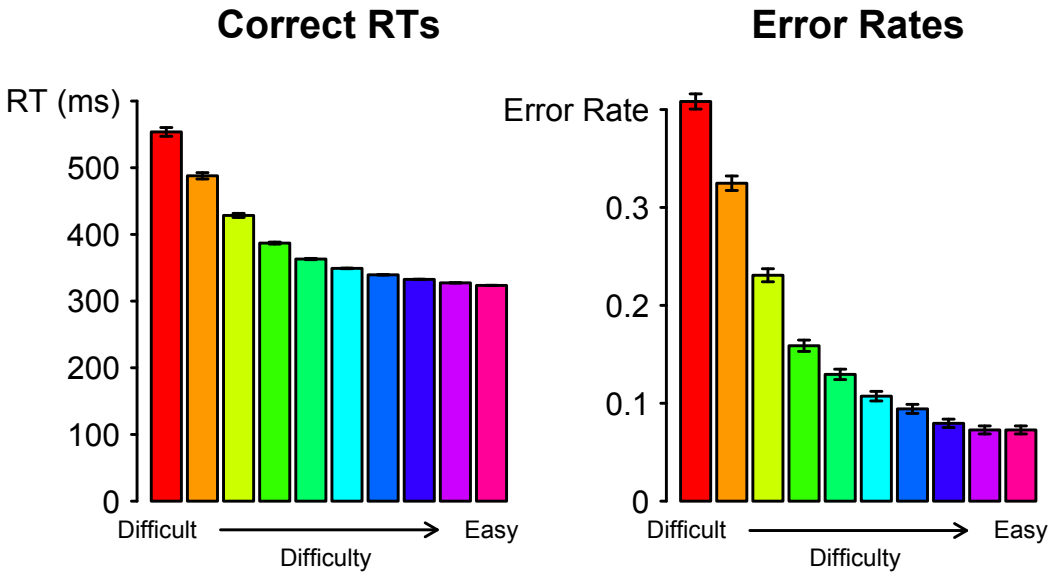

Figure 50: Simulated first-order performance as a function of difficulty, that is discriminability modelled as changes in drift rate. Correct response times (RTs, left panel) and error rates (right panel) ranging from the most difficult (red bar) to the easiest (magenta bar) condition. Error bars are standard errors of the mean; ms: millisecond.

Figure 51 shows average confidence values computed as a function of the balance of evidence at the end of the post-decision processing phase

according to Equation 8 for all ten simulated difficulty conditions. This data pattern was again very similar if compared to the empirical one presented in Figure 9, that is higher confidence on correct compared to error trials and a larger difference for easier dot decisions. This is caused by the fact that for easier conditions and consequently higher drift rates, most trials are correct trials with a large difference between the final totals of the two counters, leading to large correct-trial confidence. On the few trials on which the model chose the incorrect response, the counter accumulating evidence in favour of the correct response alternative, however, often 'caught up' with the (incorrectly) winning counter, reaching the second-order threshold τ_2 first and therefore signalling a change of mind (see second and third row of Figure 49). A lower drift rate in more difficult trials, in contrast, means that there is usually a smaller difference between the counters on correct trials, leading to lower correct-trial confidence. This also means that for error trials, the correct counter is lacking a large enough drift rate to 'catch up' with the winning counter, therefore resulting in higher error-trial confidence relative to errors on easier trials.

Indeed, the confidence distributions were distinct but overlapping for most difficulty conditions, as presented in Figure 52. This figure presents, for each of the simulated difficulty conditions, both the probability density functions for the confidence values as a function of objective accuracy, as well as dichotomised error detection distributions. If compared to the previously analysed experiments in this thesis (Figure 11) these distributions are very similar with a larger overlap for more difficult conditions, reflected both in the continuous confidence data and the binary error detection judgements. Moreover, it can be seen that the mass of the distributions is concentrated near the right end of the scale. The simulated confidence data therefore replicated the finding that participants are on average – correctly – assuming that they

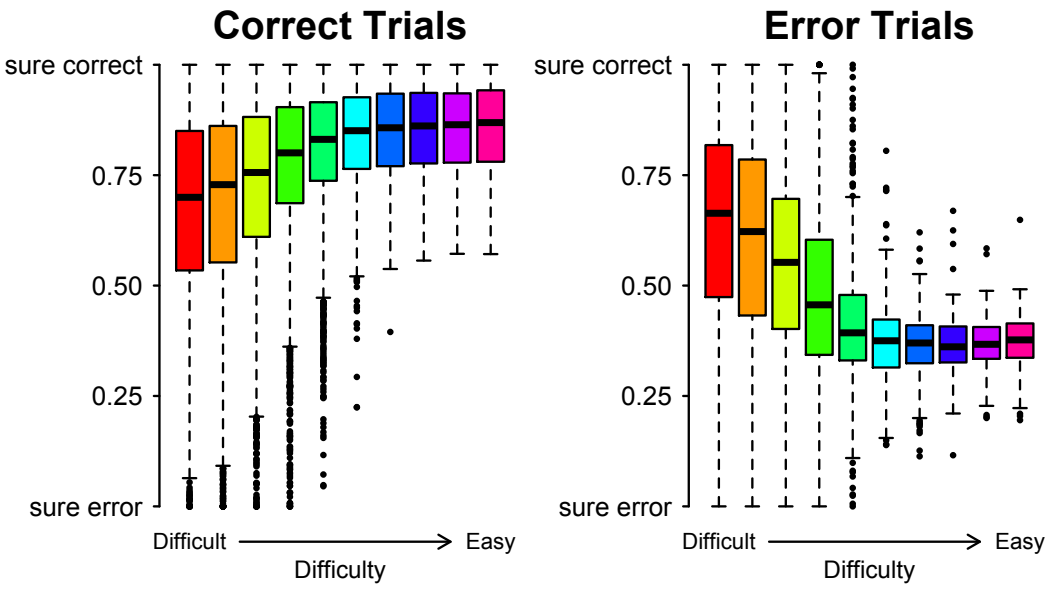

Figure 51: Simulated confidence for the ten difficulty conditions as box plots; 50% of the data lie within the box, the thick black horizontal line within each box indicates the median, the whiskers extend to 1.5 times the interquartile range (IQR); outliers are displayed as dots.

are correct rather than incorrect.

Taken together, it can be concluded from these findings that the confidence values generated by the model simulate good metacognitive resolution, similar to the empirical findings reported previously. Moreover, the simulations demonstrate how both graded confidence judgements and binary error detection judgements could arise from a single mechanism.

3.2.4.2 Visibility manipulations

The simulation results reported in the previous section suggest that a single-route model can explain core features of confidence and error detection. The question addressed here is whether the model can also deal with findings that have previously been taken to support dual-route models in which there are distinct bases for confidence and error judgments. Specifically, I aim to simu-

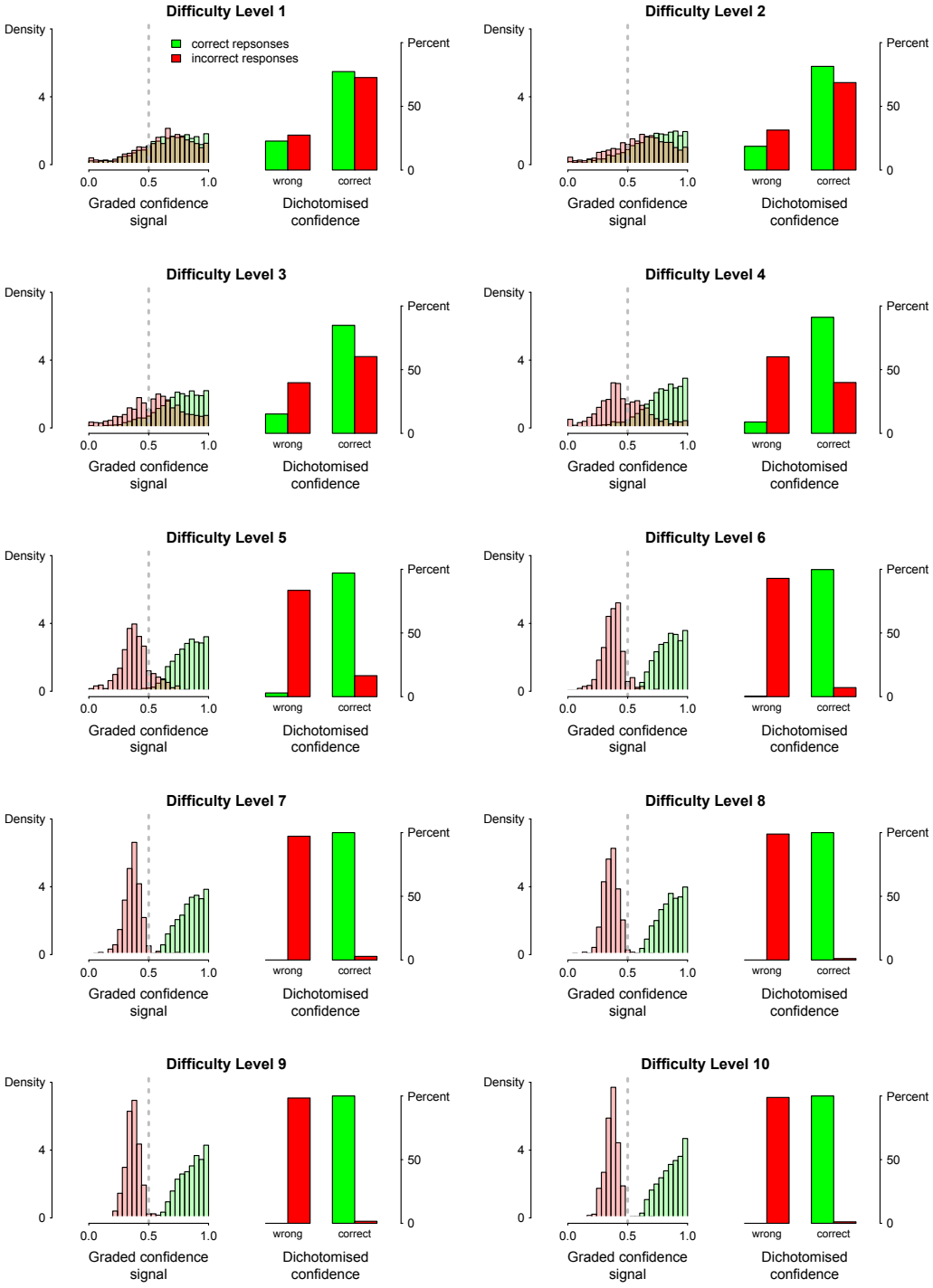

Figure 52: Simulated distributions of confidence responses as a function of difficulty and objective accuracy. The different panels represent difficulty condition, starting at the most difficult condition (2 dots difference) in the topmost, left panel. Within each panel, continuous confidence data is shown as histograms on the left. The same data can by dichotomised by splitting it at a cut off of 0.5, as shown on the right.

late effects caused by the visibility manipulation used by Charles et al. (2013), which resulted in increases in accuracy, visibility rates and metacognitive insight for longer masking times. The challenging finding that the model was set to simulate was the pronounced increase in metacognitive insight for trials subjectively classified as *seen* compared to trials classified as *unseen*. Only one accumulation rate was used for these simulations, $v_c = 5$. Instead, the time during which this visual input was available to the model was varied in eight steps, comprising the following eight SOA values: 16, 32, 48, 64, 80, 96, 112, and 128 ms. As for the previous simulation, 2,000 trials per condition were simulated. Given the eight SOA conditions and the two possible stimulus sides, this resulted in a total of 32,000 simulated trials.

First, it is important to simulate basic, first-order effects to ensure that the model is capable of capturing these, before focusing on the more challenging metacognition findings. It therefore had to be determined whether the model was able to simulate the nonlinear pattern of visibility ratings usually observed with such tasks (see Figure 53B). The right panel of Figure 54 presents proportions of trials classified as *seen* as a function of SOA, that is trials for which one of the counters in the model reached the second-order threshold τ_2 before the deadline has passed. The data pattern indeed mimicked the nonlinear psychometric functions that have been found for human participants (e.g., Charles et al., 2013; Del Cul et al., 2009), with about 20% of trials classified as *seen* for the shortest SOA as opposed to almost 100% for the longest SOA. This makes sense if we consider that for longer SOAs, the evidence sampling is driven by visual stimulus information longer than just by noise (after the mask is shown), and the counters therefore have a larger chance of reaching not just the first-order, response threshold τ_1, but also the second-order threshold τ_2.

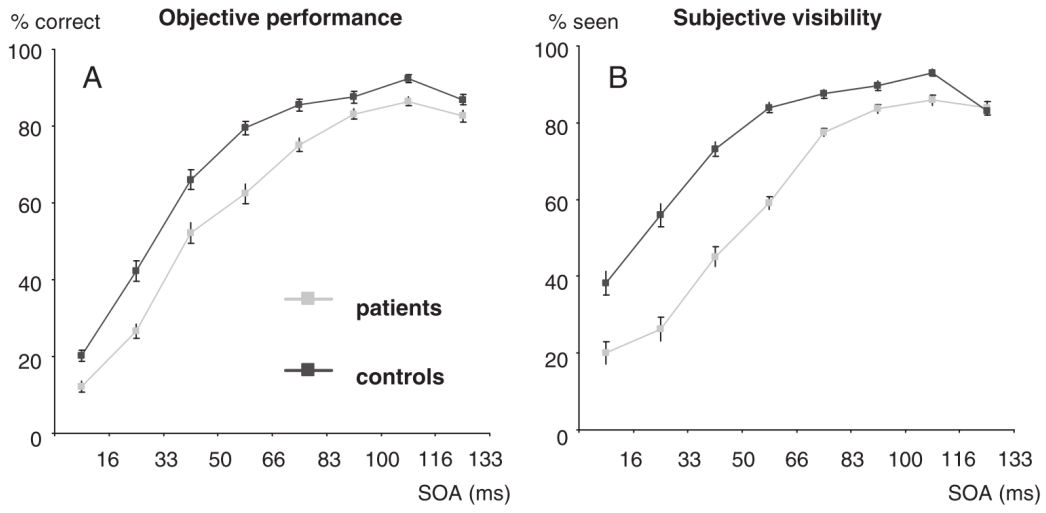

Figure 53: Figure reproduced from Del Cul et al. (2009): accuracy (left panel) and proportion of trials reported as *seen* (right panel) shown as a function of stimulus-onset asynchrony (SOA), ranging from 16 to 133 milliseconds (ms) between presentation of the visual stimulus and the onset of a mask. Data for both a patient and a control group are shown, but the patient group can be ignored here.

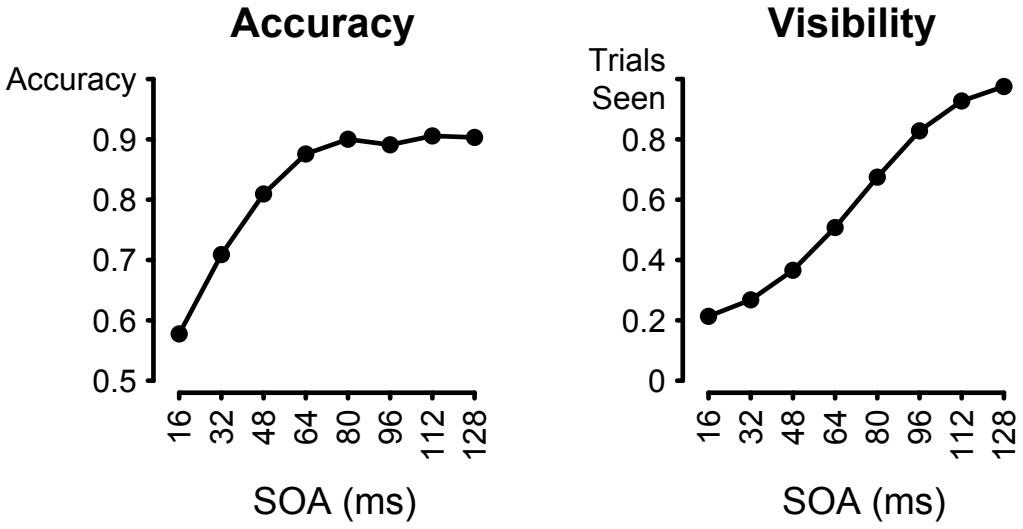

Figure 54: Simulated proportion of simulated accuracy (left panel) and trials reported as *seen* (right panel) shown as a function of stimulus-onset asynchrony (SOA), ranging from 16 to 128 milliseconds (ms) duration of visual input to the model, simulating presentation of the visual stimulus and the onset of a mask.

The next question was whether the model proposed here was able to simulate the relationship between SOA and accuracy as reported by Charles et al. (2013) or Del Cul et al. (2009), who both found accuracy to increase with SOA (see Figure 53A). As shown in the left panel of Figure 54 this was also the case here. This was due to the fact that for longer SOAs, the objectively correct counter had a substantially longer boost and therefore a much higher chance of reaching the response threshold τ_1 first. The shortest SOA used here resulted an accuracy of 57.8% – much higher than the almost 0% reported by Del Cul et al. (2009). However, this was due to the fact that the task simulated here was a 2-alternative forced-choice (2AFC) task, therefore chance rate was 50% as opposed to their task, which had ten response alternatives and therefore a chance rate of 10%.

After showing that the model is capable of simulating basic, first-order effects, I focus on the more challenging metacognition findings. Error detection ratings were compared as a function of SOA, objective accuracy, and subjective visibility ratings. In Figure 55, reproduced from Charles et al. (2013), it can be seen that the overall proportion of misclassified trials decreases with increasing SOA, suggesting that longer masking times result in better metacognitive insight. The confidence model successfully simulated this finding, as can be seen in Figure 56. Moreover, if just the lower half of the figures are considered, that is trials classified as *unseen* by the model and participants, more trials are classified correctly (solid areas) than incorrectly (diagonally shaded areas). This effect was present both in the empirical data (Figure 55), as well as in the simulations (Figure 56). From this finding, it can be concluded that weak metacognitive insight was present even for unconscious stimuli. Moreover, these figures reflect the two previously discussed effects of SOA: For higher SOAs, a larger proportion of trials lies in the upper half of

the figure, meaning the model classified those trials as *seen*. Also, for higher SOAs, the bars are more and more blue, meaning that the model chose the correct response option more often. Second, these figures contain information about the relationship between SOA, visibility and error detection. The solid areas reflect the proportions of trials correctly classified, that is correct trials classified as *correct* (a hit, according to type-II SDT) and errors classified as *errors* (correct rejection). Diagonally shaded areas, on the other hand, reflect proportions of trials that have been misclassified, that is correct trials classified as *errors* (false alarm), and errors classified as *correct* trials (miss). Taken together, visual inspection of the results by Charles et al. (2013) and the simulations suggest that the model mimicked two of the key findings on error detection: Metacognitive insight was present even for unconscious trials and the amount of visual information entering the decision system (i.e., the SOA manipulation) had an effect on this insight.

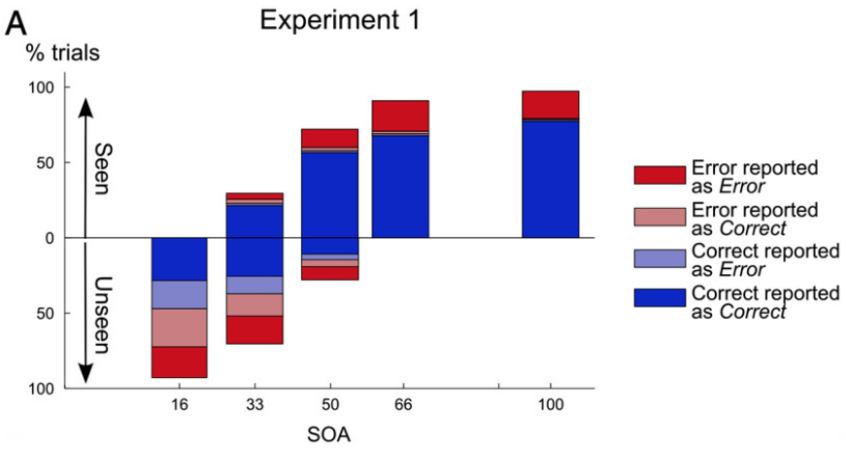

Figure 55: Figure reproduced from Charles et al. (2013): proportions of trials for different stimulus-onset asynchrony (SOA) conditions. Objective accuracy of the trials is reflected in their colour: correct trials in blue and incorrect trials in red. Error signalling is reflected in their shading with trials classified according to their objective category displayed in solid and wrongly classified trials diagonally shaded. Trials reported as *seen* are displayed in the upper part of the figure, while *unseen* trials are shown in the lower part.

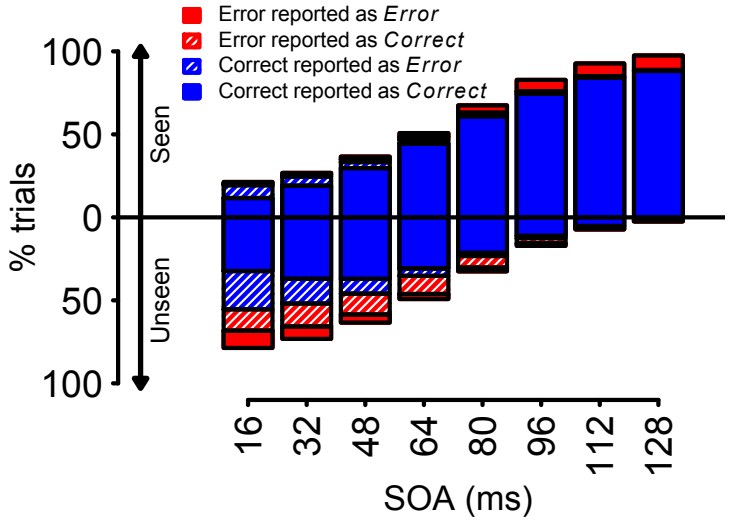

Figure 56: Simulated proportions of trials for different stimulus-onset asynchrony (SOA) conditions. Objective accuracy of the trials is reflected in their colour: correct trials in blue and incorrect trials in red. Error signalling is reflected in their shading with trials classified according to their objective category displayed in solid and wrongly classified trials diagonally shaded. The continuous confidence data was dichotomised here by splitting it at a cut off of 0.5. Trials reported as *seen* are displayed in the upper part of the figure, while *unseen* trials are shown in the lower part; ms: millisecond.

The previous simulations suggested that the confidence model is capable of capturing basic data patterns related to first- and second-order processing. The key question, however, is whether it would also replicate that metacognitive insight is better for trials classified as *seen* by the participant, independent of SOA condition. Figure 56 already suggests this: There are proportionally more misclassified trials (diagonally shaded areas) in the lower half of the figure compared to the upper half. However, if such an effect exists, then it might be easier to see this in more fine-grained, continuous confidence judgements. This cannot be compared with empirical findings from Charles et al. (2013), as only binary error detection rating were collected. I would expect to find a larger difference for correct- and error-trial confidence for trials classified as *seen* compared to trials classified as *unseen*. This is precisely what can be seen in Figure 57. This Figure presents how confidence varies as a function of objective accuracy, SOA and subjective visibility report. First, the effect of SOA on metacognitive insight can be seen as previously discussed: Correct-trial confidence increased and error-trial confidence decreased with longer SOAs. Average confidence increased only slightly for correct trials, presumably because of a ceiling effect. Second, the upper and lower panels of Figure 57 should be visually inspected and compared with regard to the key question of whether subjective visibility judgements have an effect on resolution, independent of SOA. There was indeed a larger overlap between the correct- and error-trial confidence for trials classified as *unseen*. This effect is consistent with the error-detection findings reported by Charles et al. (2013), who found that participants displayed only weak metacognitive performance on sub-threshold trials, mostly driven by higher SOA conditions.

Taken together, the simulations reported here suggest that the confidence model can also mimic participants' behaviour in a task with a manipu-

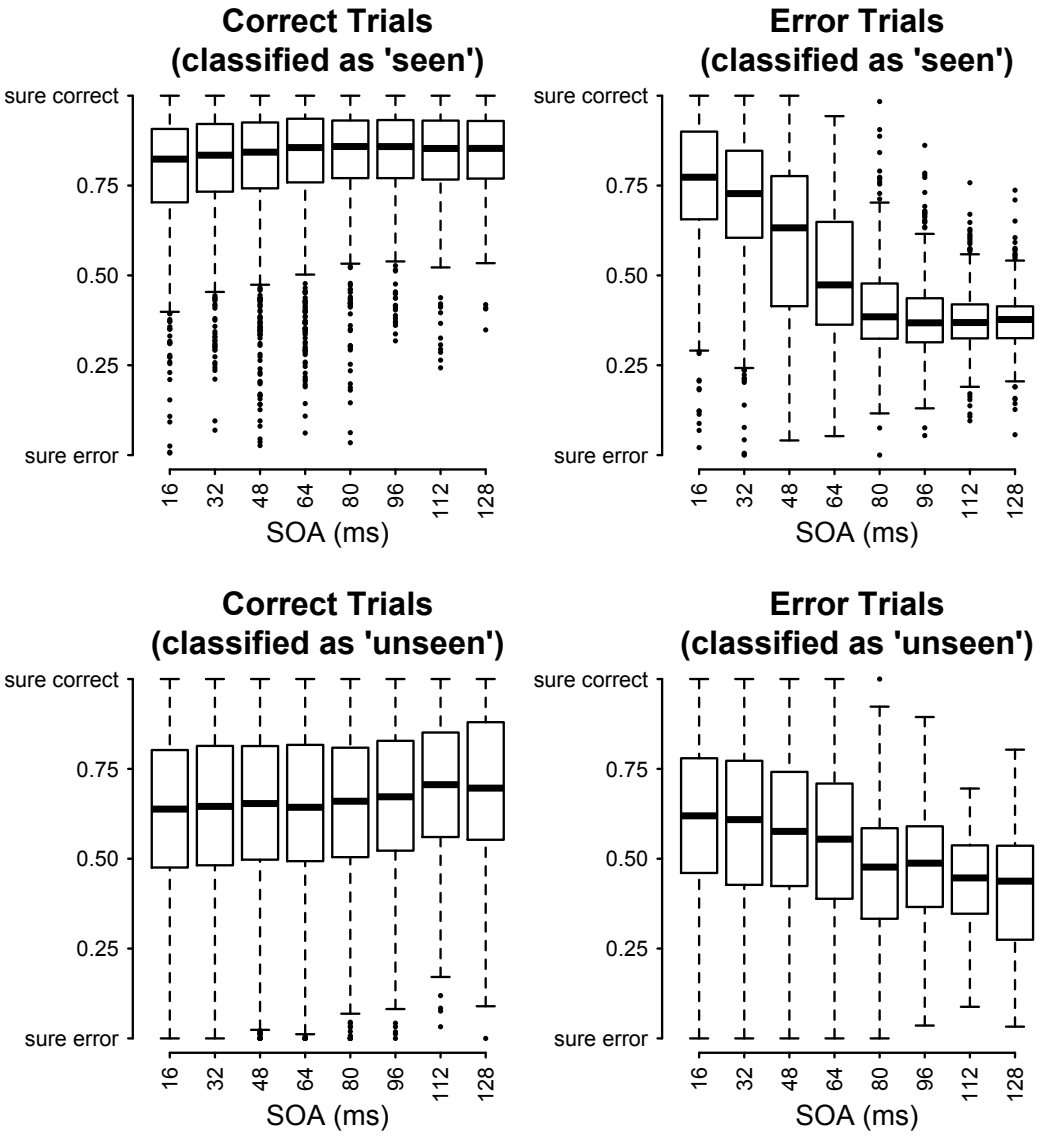

Figure 57: Simulated confidence for the eight SOA conditions as box plots; 50% of the data lie within the box, the thick black horizontal line within each box indicates the median, the whiskers extend to 1.5 times the interquartile range (IQR); outliers are displayed as dots.

lation on visibility. Simulated longer presentation of stimuli before they were masked led to higher proportions of trials classified as *seen* and also increased overall accuracy. Moreover, the model successfully captured the effect of SOA on error detection with more accurate error detection for longer SOAs, as well as for trials subjectively classified as *seen*.

3.2.5 Discussion

The data from Experiment 4 suggested similar neural mechanisms for error detection and decision confidence. However, given that these judgements have been studied largely separately to date, there are no existing models that explicitly account for both. Here, I have proposed a possible hypothesis, by developing a standard model of confidence (balance of evidence) to include post-decision processing allowing error detection and confidence, and show that it can account for a range of findings. First, I showed that the model successfully mimicked key patterns of my own decision and confidence data: The model revealed faster RTs and lower error rates for conditions of high stimulus discriminability, simulated as higher drift rates. Moreover, the simulations also showed that metacognitive insight – operationalised as differences between correct- and error-trial confidence – was higher for easier conditions. All these simulation findings match what has been previously shown for participants in Experiment 1. Second, I showed that the model was capable of replicating recent, seemingly challenging findings reported by Charles et al. (2013): Subjectively-reported visibility modulates error detection, an effect that was true over and above the effect of a common visibility manipulation (masking time) on error detection performance. More specifically, even if masking times (SOAs) were held constant, "meta-performance showed a sudden jump with visibility" (p. 86, Charles et al., 2013). The model successfully reproduced

this data pattern, interpreted by Charles et al. (2013) to suggest the need for a dual-route model of metacognition, in addition to more basic effects of masking time on visibility and accuracy rates.

Assuming that confidence relies on a post-decisional balance-of-evidence mechanism is by no means a novel finding. Indeed, a similar mechanism was already proposed by Van Zandt and Maldonado-Molina (2004), who implemented such a mechanism to explain why participants sometimes changed their mind in a recognition memory task. In their *expanded race model*, the first threshold also elicits the response, while the second threshold elicits the confidence judgement. Their model is different, however, in that it does not make any assumptions over binary error detection processes, or visibility judgements.

The confidence model proposed here assumes a graded metacognitive signal, which is based on the balance of evidence at the end of a post-decision processing stage. This signal can be read out directly and reported as graded confidence, or instead be dichotomised and reported as binary error detection. Such a mechanism is similar to one of the error detection mechanisms proposed by Charles et al. (2013, p. 93), described as "statistical information on the likelihood of having made an error". On the other hand, the authors also assumed a second error detection mechanism, which is described as all-or-none and which they assumed to be present only on trials for which the stimulus had been reported as *seen*. The authors suggest that both mechanisms are necessary to explain why error detection was more accurate for trials reported as *seen*. However, with the present study, I have successfully replicated this effect with a single error detection mechanism.

Moreover, Charles et al. (2013) reported an interesting modulation of the ERN, which is another reason why they have assumed two separate error

detection mechanisms. They found an ERN – that is, a difference in amplitude between correct and error trials at fronto-central electrodes around the time of the response – only for trials reported as *seen*, while there was no such difference for trials reported as *unseen*. In other words, conscious access to the visual stimulus is needed in order for an ERN to be produced. The authors suggest that to simulate such an effect, a dual-route model is needed, proposing the model by Del Cul et al. (2009) as a suitable starting point. As described in Section 3.2.2, their dual-route model consists of two separate sets of racing evidence counters – one for each choice option. One route is fast but noisy (unconscious response route), the other is slow but more precise (higher-level conscious decision route). A decision is triggered if one counter within a route reaches a decision boundary. Trials are classified as *seen* if a counter within the conscious route reaches the threshold within a pre-specified time. If the threshold is not reached, evidence within this route is discarded. An ERN is triggered if there is a mismatch between the two routes, so per definition, there can only ever be an ERN on trials classified as *seen*.

Here, I argue, however, that a dual-route model is not necessary for explaining the modulation of the ERN. Instead, I would like to suggest that my single-route model could account for the effect just as well by assuming that an ERN is triggered by a mismatch between the winners of the two thresholds. On trials classified as *unseen*, no winner of the second-order threshold, τ_2, is registered and a possible mismatch can therefore not be detected and the ERN thus not be triggered. I do not provide an explicit simulation here, because the modulation of the ERN follows intuitively from the mechanism I proposed, but future versions of the model could be aimed at simulating the ERN, not just in whether or not it exists but also with regard to its amplitude and precise timing. Proposing that the ERN is triggered by a simple matching heuristic

rather than a computationally more complex mechanism like the balance of evidence makes sense if we consider that the ERN has often been regarded as an early indication of an error, rather than a reflection of error awareness (Charles et al., 2013; Steinhauser & Yeung, 2010). This also fits well with the findings of Experiment 4, which suggested that the Pe, but not the ERN, reflects differences in decision confidence.

It is important to stress that this single-route model proposed here was intended as an illustration of how metacognitive judgements might be processed internally in a way that combines confidence and error detection. The aim was to outline a plausible model rather than propose a definite account or contrast competing alternatives. Further rigorous, quantitative testing is no doubt necessary to develop this approach into a true computational model of metacognition, which could then be tested against other models using formal model comparison approaches. The simulation results would therefore not only be judged regarding whether or not they fit the empirical data patterns qualitatively, but also quantitatively. This approach would entail exploring the multidimensional parameter space using a fitting algorithm such as simulated annealing (Cerný, 1985; Kirkpatrick, Gelatt & Vecchi, 1983), estimating the goodness of fit of different parameter combinations to find the a quasi-optimal solution. One natural test would, for example, be to compare the single-route approach to a dual-route version of the same model.

A similar, direct comparison of a dual- and a single-route model of decision making has already been reported by Del Cul et al. (2009), who found that the dual-route model outperformed the single-route model in simulating both visibility ratings, and choices (no error detection was simulated in their study). However, it is possible that this reduced goodness of fit was due to several shortcomings of their single-route model, which have been addressed

in my model. First, the here-reported model assumes a second threshold, while the single-route model in Del Cul et al. (2009) implemented visibility as evidence exceeding a first threshold and *unseen* trials as trials on which the choice was forced because this single threshold was not reached. The second threshold proposed in my model has the advantage that the model is made more similar to a comparable dual-route account and that any direct comparisons in the future would be focusing solely on the single/dual-route feature of the model. Second, the model reported in the present chapter implemented fast responses not only by starting point variability and noise in the accumulation of evidence, but also by adding a hand-specific bias, which slightly favoured responses made with the right hand over responses made with the left hand. Such a hand bias is often seen for participants, most of which can be assumed to be right-handed. Presumably, this could have helped in the simulation of error trials. The single-route model proposed by Del Cul et al. (2009) struggled with precisely this point, producing lower error rates than observed in the data.

Taken together, the simulation results reported in this chapter suggest that a single-route model with a post-decisional balance-of-evidence mechanism is able to account for the first- and second-order judgements in the dot-count task. The model assumes that confidence and error detection judgements arise from the same internal mechanism, as also suggested by the findings from Experiment 4. Moreover, a visibility manipulation as used by Charles et al. (2013) and Del Cul et al. (2009) provided another challenging test for the model, in that subjective visibility judgements modulated error monitoring performance over and above the effect of a visibility manipulation. The model also successfully simulated data for this type of task, leading me to conclude that it represents a straightforward and adequate representation of

how metacognitive judgements are generated in a decision process and could therefore serve as a useful starting point for a quantitative model of decision confidence, which could then be formally compared to other theoretical models of metacognition in future research.

3.3 General discussion

The present chapter focused on the question of whether confidence judgements and error detection arise from similar neural mechanisms. Research on both types of judgements has been carried out largely separate, despite using highly similar methodology. Experiment 4 examined neural correlates of metacognition in perceptual decision making. The findings provide evidence that well-characterised neural correlates of error monitoring are predictive of graded changes in decision confidence. I propose that the Pe – an EEG component that has repeatedly been linked to error awareness (Nieuwenhuis et al., 2001; Hester et al., 2005; Steinhauser & Yeung, 2010) – provides a generic, sensitive index of decision confidence, and is not limited to binary error detection, suggesting that shared mechanisms underlie error monitoring and confidence judgements. This was the case both for averaged data, as well as on the individual-trial level. The computational model described in this chapter took the hypothesis of a link between error detection and confidence further, formulating a computational model which assumes that both types of judgements arise from the same internal mechanism. More precisely, it is assumed that metacognitive judgements are based on the balance of evidence at the end of a post-decision processing phase, and which is therefore similar to a model proposed by Van Zandt and Maldonado-Molina (2004). Such a continuous metacognitive signal can either be 'read out' directly as confidence or dichotomised when binary

error judgements are required.

Linking error detection and decision confidence has substantive theoretical implications. First, I would suggest that my findings have an impact on current theories of error monitoring, which have studied error detection largely as an all-or-nothing phenomenon (e.g., the 'conscious' error detection mechanisms proposed by Charles et al., 2013). Here, I have presented data in support of the idea that error detection – in its entire 'range' from *sure errors* to *sure correct* trials – might be a graded signal, which should also be studied as such. Second, many current theories of decision confidence – such as the standard balance-of-evidence model (Vickers & Packer, 1982) – struggle to explain why participants sometimes change their mind and judge responses as *surely incorrect*, therefore ignoring half of this 'range' of confidence scales. The reason for this shortcoming of the standard balance-of-evidence model lies in the implicitly assumed decisional-locus model. Here, I have argued that these standard models have to be extended to also include post-decision processing, given that such error judgements are known to depend on continued processing of stimulus and response information after the initial decision has been made (Yeung & Summerfield, 2012, 2014). This hypothesis is furthermore supported by my findings from Experiment 2, regarding participants tendency to take more time to judge their confidence when the time interval between the response and the onset of the confidence scale is short. Third, another theoretical implication concerns the neural basis of confidence, which is still highly debated (see for instance Fleming & Frith, 2014, for a review). A possible role of the right rostrolateral PFC, especially area BA10, has been suggested (Fleming et al., 2010; De Martino et al., 2013; Yokoyama et al., 2010). The neural basis of error monitoring, on the other hand, is well characterised: It is commonly assumed that error likelihood is communicated by the medial PFC

to lateral PFC, which can then execute cognitive control (Yeung, 2013; Egner & Hirsch, 2005). These findings could therefore provide useful constraints in building theories of the neural bases of metacognition in decision making.

On a related matter, linking confidence and error detection also suggests that more attention should be paid to the possible functions and uses of decision confidence in cognitive control, given that the role error detection plays in such control processes has been extensively studied. This question has largely been ignored in the line of research on decision confidence. Metacognitive judgments play an important role in decision-making because – as shown repeatedly in this thesis and elsewhere – they have a significant relationship with objective performance: When participants report low confidence, they typically also committed an error; when they express high confidence, they have most often been correct. Reflecting on one's own thoughts in this way is highly essential. Without this ability, learning and adapting to an ever-changing environment would be almost impossible. Findings from the error monitoring literature, for example, have reported that participants slow down after detecting an error (Laming, 1979; Dutilh et al., 2012; Notebaert et al., 2009), to prevent further mistakes on subsequent trials. Metamemory on the other hand, that is to say how confident we are that we know something, has been shown to have an influence on how study time is allocated (Nelson & Leonesio, 1988). Measures of confidence are therefore important indices of how participants exert cognitive control (Fernandez-Duque et al., 2000), but assessing them can be difficult due to several reasons. The first reason is a rather practical one: Measuring confidence increases the length of a study, usually doubling or even tripling the amount of time required for a given number of trials. Second, as suggested by findings from Experiment 3, assessing confidence judgements can lead to increased primary RTs and a

more accuracy-focused response strategy, which might not be desirable in some designs. EEG recordings provide a robust, non-disruptive index of confidence that circumvents these problems, enabling researchers to assess subjective confidence without requiring participants to make explicit and overt judgements. This approach will be further tested in Experiment 5. Taken together, in addition to these theoretical implications, there is also a practical implication that follows from a link between error detection and decision confidence: EEG measures of the Pe as studied here promise to provide a useful non-invasive and robust index of metacognitive evaluation that might be leveraged in future research to assess levels of confidence whenever direct measurement is impossible or inconvenient, and hence used to shed further light on the underlying mechanisms of metacognition in decision making.

Finally, another implication for current theories of metacognition in decision making follows from the internal mechanism on which confidence and error detection are based in my model: Many error monitoring theories have assumed that error detection is based on a mismatch between an intended action and the actually executed response (Falkenstein et al., 1991; Gehring et al., 1993; Coles et al., 2001). On the other hand, many confidence theories have focused on balance of evidence as a basis for metacognition (Vickers & Packer, 1982; De Martino et al., 2013; Van Zandt & Maldonado-Molina, 2004; Kepecs et al., 2008). Here, I have proposed a mechanism that indirectly combines both of these mechanisms: post-decisional balance of evidence (cf. Van Zandt & Maldonado-Molina, 2004). In other words, when the losing counter of the first decision threshold reached the second threshold first, this trial would represent a mismatch and therefore a change of mind.

Taken together, the findings reported in this chapter suggest that error detection and confidence represent the same internal, neural mechanism,

thereby combining two previously distinct literatures to provide a more comprehensive model of metacognition in decision making. I have highlighted the value of merging the two literatures that have led to different theories, and different focuses of investigation, by discussing important implications for the underlying theories of decision confidence and error monitoring, as well as practical implications for the indirect measurement of metacognitive signals. As previously proposed by Yeung and Summerfield (2012, 2014), future work in this area should focus on formulating a common theoretical framework for metacognition in decision making, for which this chapter aims to provide a first step.

Chapter 4

Uses of metacognition

Over the course of the previous two chapters, I have highlighted the importance of decision confidence, suggested ways to measure such confidence judgements, and underlined their conceptual overlap with error detection. All of these questions have in common that they focus on how decisions affect metacognitive judgements (see also Chapter 5). Here, I shift my focus towards the functional role of decision confidence, specifically to focus on the involvement of metacognitive information in cognitive control. Consistent with this notion, Shea et al. (2014, p. 186) have defined metacognition as the "use of metacognitive representations (often, but not exclusively, for purposes of cognitive control)". Moreover, theories of metamemory have also acknowledged the key role confidence and other metacognitive cues play in cognitive control. Dunlosky, Serra and Baker (2007), for instance, divided metacognition into the three subareas: knowledge, monitoring and controlling one's own cognitions. The metamemory model by Nelson and Narens (1990; see also Nelson & Narens, 1994) also includes a control component. They assumed three stages of learning: acquisition, retention and retrieval. Throughout these stages, participants monitor their own learning progress and memory and – if necessary – adjust behaviour. This can happen, for example, by means of adapting their

strategies or deciding whether to continue or stop learning an item. Indeed, participants use confidence as a cue in guiding the allocation of study time, spending more time on items they are less confident about (Nelson & Leonesio, 1988).

The proposed feedback loop of the metamemory model by Nelson and Narens (1990) is presented in the upper panel of Figure 58, which has been reproduced from Fernandez-Duque et al. (2000). This model assumes that object-level representations, such as percepts, memories, and decisions, are used bottom-up to inform and update a mental model of the object-level world at the meta-level, therefore constituting a monitoring process. These meta-level representations, in turn, can then be used to exert top-down control over the object level. For example, if the current response is made very slowly, then this information regarding response speed is communicated to the meta-level, where it could subsequently be used to update an estimate of error likelihood for the current trial. If an error is indeed detected, the meta level can signal to the object level the need for increased cognitive control, which could, for instance, result in post-error slowing of the next response.

Fernandez-Duque et al. (2000) have noted that there is conceptual similarity in this proposed feedback loop to theories of executive processing: The lower panel of Figure 58 presents a schematic illustration of the theory on executive attention by Norman and Shallice (1986). Here, top-down control allows participants to adapt flexibly their behaviour according to their current intentions by allowing the executive system activate or inhibit lower-level schemas, which would normally respond automatically according to stimulus input. The state of these schemas is in turn communicated bottom-up to the executive system to allow for a full feedback control loop.

The hypothesis that error detection can serve as an internal cue for the

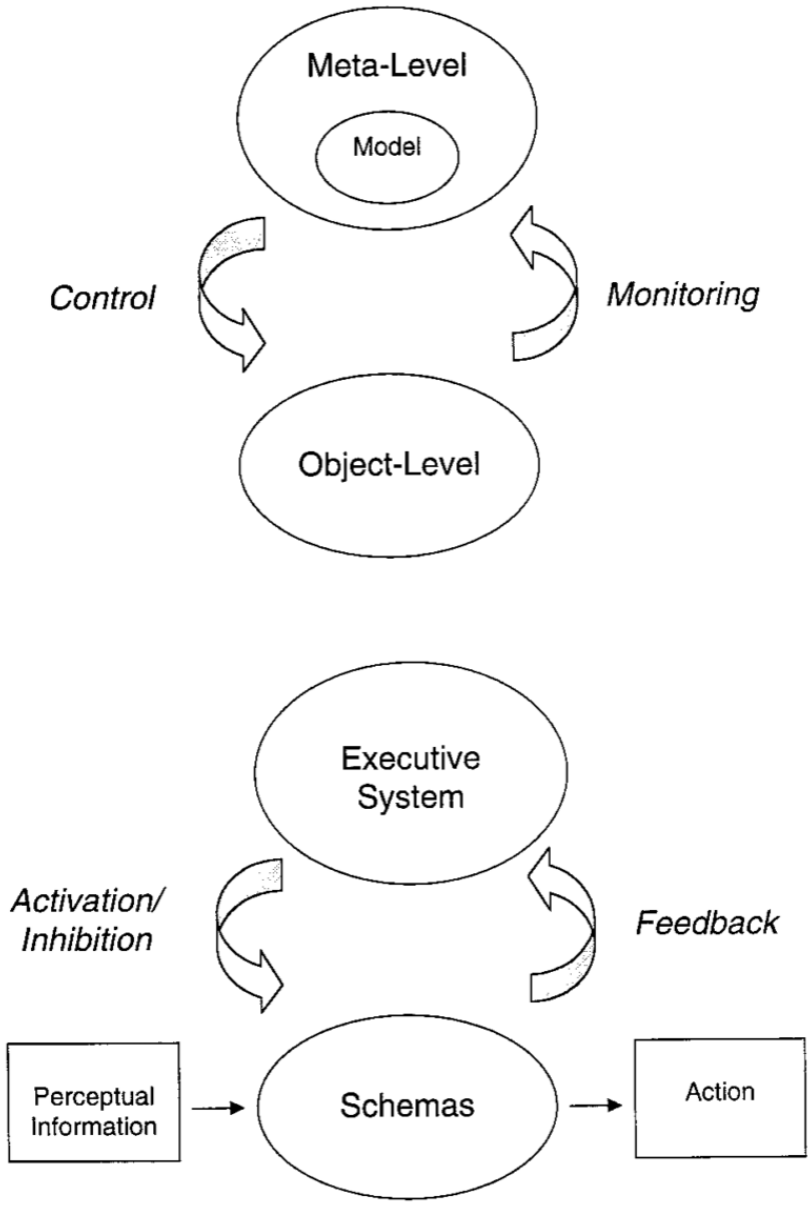

Figure 58: Figure reproduced from Fernandez-Duque et al. (2000): the upper panel shows the proposed feedback loops between the object level (e.g., percepts, memories, etc.) and the meta level (e.g., confidence, judgements of learning, etc.) as suggested by Nelson and Narens (1990). The lower panel shows similar feedback loops for the interaction of the executive system with lower-level schemas according to the theory on executive attention by Norman and Shallice (1986).

need to enhance cognitive control has been previously discussed in this thesis, both in Chapters 1 and 3. Such effects have, for example, been studied in the context of post-error slowing: It has often been observed that people tend to be slower after making an error (Dutilh et al., 2012; Rabbitt & Rodgers, 1977; Rabbitt, 1966), even if they are not aware of such an error (Logan & Crump, 2010). Different interpretations for this effect have been proposed (e.g., Jentzsch & Dudschig, 2009; Danielmeier & Ullsperger, 2011). Dutilh et al. (2013), for example, suggested that following errors, participants adopt a more cautious response mode.

Moreover, Fernandez-Duque et al. (2000) have discussed another example of how metacognitive signals are utilised to improve cognitive control: Cognitive conflict, or response uncertainty arises from competing response tendencies. For instance, in the Stroop paradigm (Stroop, 1935), participants have to name the ink colour of a word, whilst ignoring the colour word itself. Conflict therefore arises from a competition between the automatically triggered response of the colour word and the actual ink colour. It has been suggested that the medial PFC monitors such conflict signals, passing this information on to lateral PFC, which is then in turn able to execute cognitive control (Yeung, 2013; Botvinick, Braver, Barch, Carter & Cohen, 2001). The role of response uncertainty, or conflict in cognitive control has also been highlighted in the recent reviews on metacognition in decision making by Yeung and Summerfield (2014, 2012). This link was also previously made by Davelaar (2009), who proposed a computational model in which metacognitive judgements are a function of cognitive conflict.

Taken together, extensive evidence exists in support of the notion that both cognitive conflict and errors are used as signals to enhance cognitive control (Fernandez-Duque et al., 2000; Yeung & Summerfield, 2014). In the

present chapter, I therefore aim to address the question how decision confidence affects future processing. More precisely, I focus on how confidence modulates attention seeking with regard to the processing of feedback. The general idea is that decision confidence serves as an internal proxy to feedback, especially when external feedback is not available or unreliable. Whenever participants have low certainty (*guessing*), I expect them to pay most attention to feedback, which could presumably give them information on how to avoid mistakes on future trials. On the other hand, I expect that participants will tend to ignore the feedback stimulus when they have a strong belief that they have been correct or incorrect, because they have a low probability of gaining information from the feedback. This hypothesis is also shown in Figure 59. This pattern should be the same for correct and error trials. If this hypothesis held true, it could be interpreted as further support that confidence is used as an internal feedback signal. I aim to test this using a perceptual decision-making paradigm in which participants received feedback after each trial. Using the method introduced in the context of Experiment 4, I aim to measure single-trial Pe amplitude as a proxy of decision confidence. I expect to find attention to feedback to be lowest when Pe amplitude is at the extremes, that is very large or very small.

4.1 Experiment 5: Does decision confidence predict attention to feedback?

The experiment in this chapter focused on how confidence affects the processing of feedback following participants' responses. The general role of feedback has been studied extensively in the context of learning studies. For instance, Kluger and DeNisi (1996) have highlighted that providing feedback does not

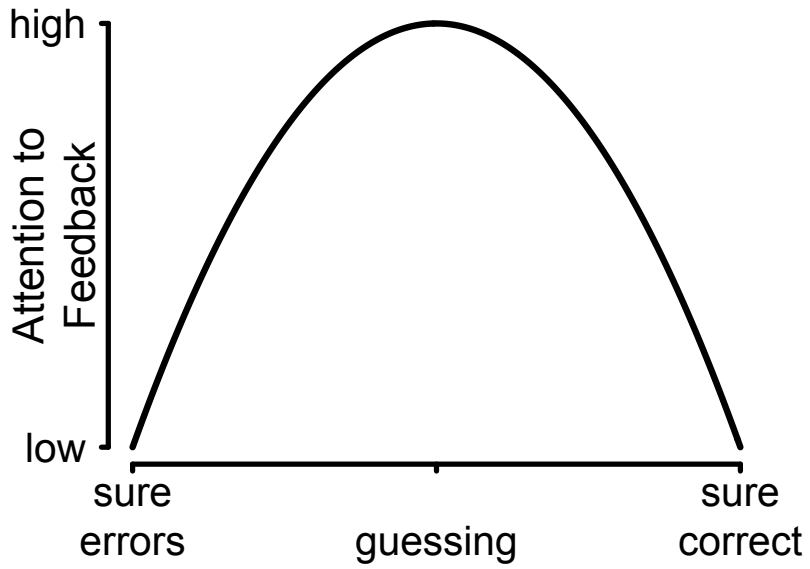

Figure 59: Hypothesis regarding the relationship between confidence (x-axis) and the amount of attention that is paid to a visual feedback stimulus (y-axis).

necessarily improve task performance, and can even lead to decreased performance due to shifts in the locus of attention. Similarly, feedback has been associated with opportunity costs, that is time spent on studying feedback reduces the available time to study the learning material. Indeed, Hays, Kornell and Bjork (2010) showed that performance in a fixed-study-time paradigm could be improved if participants were given the option to skip feedback – an option they used on a large proportion of correct and also some error trials. The trials on which they chose to skip feedback presentation were arguably trials on which they were certain regarding their feedback prediction. This study therefore suggests that confidence might have an effect on how feedback is processed, with sure errors and sure correct trials leading to a higher probability of skipping feedback. Presumably, this effect was due to the fact that confidence represents an internal feedback signal.

The idea that confidence serves as an internal feedback signal also finds support in research on self-regulated learning (SRL). Self-regulated learning

describes situations in which the learner is actively engaged in structuring the learning process, such as students revising course material for an upcoming exam. To do so efficiently, they need to set subgoals internally, structuring the material and identifying sub-areas that need extra attention because they have not yet been mastered. Butler and Winne (1995) have highlighted how for such learning situations, internal feedback – such as confidence – plays an important role alongside external feedback. More specifically, they assumed that participants constantly evaluate their progress, and that this self-generated feedback helps them to efficiently guide their learning process. The authors also cite a study by Zellermayer, Salomon, Globerson and Givon (1991), which introduced *Writing Partner*, a software tool for assisted essay writing, which aims to enhance this internal feedback through triggering metacognitive control processes. For example, the virtual tutor would ask participants questions about the essay they were planning to write and would later on present the students' answers to these questions, presumably helping them at detecting possible mismatches between their own goals and actions.

Whether confidence modulates attention to feedback has also previously been studied in educational psychology, often with the conclusion that feedback has a corrective, rather than a confirmatory function (Timmers & Veldkamp, 2011; Mory, 2004). Timmers and Veldkamp (2011), for example, found that more attention is paid to feedback on error trials. Kulhavy and Stock (1989), on the other hand, found that participants spent more time studying feedback when there was a mismatch between their perceived accuracy (confidence) and their objective accuracy, as conveyed by feedback. In other words, this means they spent longer studying feedback when they had expressed high confidence in an answer but committed an error. However, these results do not provide conclusive evidence that confidence directly affects

attention to feedback, because the study confounds information regarding objective accuracy with the study material itself. In other words, feedback was provided together with information about the correct answer (i.e., elaborated feedback), therefore presumably leading to new information integration.

Taken together, while the literature on educational psychology has provided interesting findings on how feedback is processed and what role internal feedback (i.e., confidence) might play, the precise relationship between confidence and attention directed at feedback remains unclear. This is therefore the goal of the experiment presented in this chapter. Crucially, feedback is taken here to be a visual stimulus providing information regarding the objective accuracy of a previous decision – and therefore not elaborated feedback that contains additional information. The present experiment will therefore address this question, predicting that participants will pay more attention to feedback when they were uncertain in their first-order response – that is when they were guessing – given that they can gain attention from feedback on those trials. This relationship is furthermore shown in Figure 59.

One of the key challenges of this experiment was to measure confidence as well as provide visual feedback to participants: Asking participants to rate their confidence and then providing them with feedback might be not advisable, because asking for metacognitive judgements might change the processing of feedback *per se* because participants are trying to match their second-order judgement against objective feedback. Providing participants first with feedback and then asking them to rate their confidence is, of course, not possible, because it would reveal the true, objective accuracy before giving participants a chance to estimate accuracy themselves. The present experiment therefore makes use of EEG markers of confidence found in Experiment 4. The rationale of this approach is to measure confidence on a single-trial level, without

explicitly asking participants for their confidence rating. This confidence proxy is expected to be predictive of participants' level of attention to explicit feedback.

Attention directed to feedback was measured not through the time spent studying it – which was held constant given that it was just one word – but through EEG correlates of attention, anticipation and expectation. Four ERP components were analysed with this study: the stimulus-preceding negativity (SPN; Van Boxtel & Böcker, 2004), the N2pc (Luck & Hillyard, 1994a), the feedback-related negativity (FRN/fERN; Holroyd & Coles, 2002), and the P3 (Sutton, Braren, Zubin & John, 1965). To foreshadow the results of this experiment, I found that none of these ERP components varied with confidence in the ways predicted here. In the discussion section (Section 4.2), I suggest reasons why this might have been the case.

4.1.1 SPN

The SPN is a slow preparatory EEG component thought to reflect expectation of a task-relevant stimulus, such as visual feedback, therefore being different from the effects of general motor preparation (Brunia, 1988; Damen & Brunia, 1987; for a historical review see Van Boxtel & Böcker, 2004). It is usually observed in experiments in which participants anticipate the presentation of a visual feedback stimulus. The SPN is, as its name says, a negative deflection, which arises over frontal electrodes, often showing right hemisphere dominance (Brunia, Hackley, Van Boxtel, Kotani & Ohgami, 2011, for a review). This slow cortical potential starts to develop up to several seconds prior to the anticipated event and continues to build up until the presentation of this feedback stimulus, often referred to as *knowledge of results* (KR) stimulus.

It has been argued, however, that the SPN does not merely reflect anti-

cipatory attention (Brunia et al., 2011), but also information load. This means a larger (i.e., more negative) SPN occurs under circumstances in which the anticipated feedback or stimulus provides the participant with more information, thereby reducing uncertainty (Fuentemilla et al., 2013; Foti & Hajcak, 2012; Brunia, 1988; Kotani et al., 2003; Ruchkin, Sutton, Mahaffey & Glaser, 1986; Chwilla & Brunia, 1991; see also Brunia et al., 2011, for a review). A recent study by Morís, Luque and Rodríguez-Fornells (2013) provided evidence in favour of this hypothesis. The authors used a paradigm in which participants had to learn which food items would cause an allergy to an imaginary patient. They received informative feedback following each trial. The SPN elicited by this feedback was strongest (i.e., most negative) at the beginning of the experiment and this effect then became weaker over time. Morís et al. (2013) interpreted these findings as support for the learning hypothesis, according to which early on, participants could extract the most information from the feedback. After having learned the relationship between food items and allergic reactions, however, the information load of the feedback decreased and thus also the SPN, reflecting the information load.

For the present experiment, this means that the amplitude of the SPN should be largest in situations in which feedback reduces uncertainty to the largest extent, such as when the participant does not know whether or not he made a mistake. This would therefore affect the levels of medium confidence. On trials in which the participant is certain, however – either certain the just-made decision was correct, or certain that it was an error – the SPN amplitude should be smaller, that means more positive given that this is a negative component. I would therefore expect an inverted U-shaped function for the SPN amplitude over different levels of confidence.

4.1.2 N2pc

The second attention-related EEG component of interest was the N2pc. In contrast to the SPN, however, this is a stimulus-locked component. The N2pc is thought to reflect spatial attention (Eimer, 1996; Luck & Hillyard, 1994a). For instance, the N2pc has been studied in visual search tasks, in which participants had to find a target amongst several distractors (e.g.; Luck & Hillyard, 1994a). Allocation of attention to the left or right side of the search array was reflected in the amplitude of the N2pc. This ERP component has been identified as a negative deflection at around 200 ms after the onset of a visual stimulus. The "p" in this component's name furthermore expresses that the topography of this component is posterior and the "c" stands for contralateral, expressing that the component is lateralised with stronger (i.e., more negative) activity in the brain hemisphere opposite to the side in which the attended stimulus is located (Luck & Hillyard, 1994b). In this experiment, feedback was lateralised. I expected greatest attention to feedback, evident as an enhanced contralateral N2pc, on trials in which participants were uncertain.

4.1.3 FRN

The next analysis focused on the FRN, a fronto-central negativity occurring approximately 200 to 350 ms after a feedback stimulus was presented. Its amplitude is often found to be larger for negative as compared to positive feedback stimuli. It has therefore been argued that the FRN and ERN are generated by the same system (Holroyd & Coles, 2002; Miltner, Braun & Coles, 1997; see also Walsh & Anderson, 2012, for a review). Indeed, the ACC has often been identified as their common generator (Holroyd & Coles, 2002), especially its dorsal part (Hauser et al., 2014; Holroyd et al., 2004).

The FRN is usually understood to reflect reward prediction errors (RPE) processing in the brain, for example in the framework proposed by Holroyd and Coles (2002). This framework links error monitoring to the midbrain dopamine system, assuming that the ACC acts as a control filter and that it elicits the ERN (and also its feedback-related equivalent, the FRN) whenever the outcome of an action was worse than expected (negative RPE). More specifically, they assume that the basal ganglia act as an *adaptive critic*, calculating the value of the outcomes of an action. If these outcomes are worse than expected, there is a phasic decrease in dopamine, which disinhibits the ACC, leading to a more negative ERN or FRN, whereas phasic increases in dopamine lead to a more positive ERN/FRN. They furthermore argued that this signal is used internally to train a motor planning system to produce more adequate actions in the future, using a temporal difference learning framework. This hypothesis was supported by the findings of Cohen and Ranganath (2007), which suggest that the size of the FRN effects predicts behavioural adaptations in future trials.

The RL-ERN model by Holroyd and Coles (2002) suggests that the ERN or FRN reflect a signed RPE (Holroyd & Coles, 2002; Walsh & Anderson, 2012; Nieuwenhuis, Holroyd, Mol & Coles, 2004; Walsh & Anderson, 2011; Pfabigan, Alexopoulos, Bauer & Sailer, 2011). According to this hypothesis, we should expect to find the largest FRN amplitude (i.e., most negative amplitude) for trials in which participants had high confidence but received error feedback. Conversely, the amplitude of the FRN should be smallest whenever the outcome of a trial was much better than expected, that is very low confidence but feedback that the participant had been objectively correct.

However, as opposed to the hypothesis that the FRN reflects a signed RPE, results from recent studies have suggested that it instead reflects an

absolute or unsigned prediction error, such as unexpectedness or a surprise signal (Cavanagh & Frank, 2014; Alexander & Brown, 2011; Talmi, Atkinson & El-Deredy, 2013; Oliveira, McDonald & Goodman, 2007; Hauser et al., 2014; Sallet, Camille & Procyk, 2013; Cavanagh, Figueroa, Cohen & Frank, 2012). According to the idea that the FRN reflects an unsigned prediction error, the amplitude of the FRN should be largest for the case in which correct feedback was expected with high certainty but instead negative feedback was given and the case in which error feedback was expected with high certainty but instead positive feedback was given. The FRN is expected to be similar for these two cases of maximum surprise. In turn, the smallest FRN should be found for conditions in which feedback follows the highly certain expectation. In this study, I tested whether the FRN follows such a surprise signal.

4.1.4 P3

The P3 is a positive deflection in the EEG at around 300 ms after the onset of a task-relevant stimulus. This component, which typically has a centro-parietal scalp topography, has been localised in the delta and theta band (Selimbeyoglu et al., 2012; Başar-Eroglu, Başar, Demiralp & Schürmann, 1992). Studies have furthermore identified two separate subcomponents (Polich, 2007). According to this view, the P3a reflects early attentional processing, while the P3b reflects memory updating. The P3 has been identified as an EEG component that reflects reward processing. The P3 was first reported by Sutton et al. (1965) as a reflection of stimulus probability or surprise (see also Johnson, 1986) – the less probable the stimulus that elicited the P3, the larger its amplitude.

For the present study, we could therefore expect to find that the P3 reflects such stimulus probability, that is the P3 should be largest for error trials in bin 1 (where participants expected positive feedback and instead received

negative feedback), as well as for correct trials in bin 4 (where participants expected negative feedback and instead received positive feedback), similar to the hypothesis for the FRN.

Alternative explanations exist as to what P3 amplitude reflects. Sallet et al. (2013), for instance, recently reported data from a trial-and-error learning paradigm in which the P3 was modulated by a positive RPE only. In the present experiment the rules never changed and learning can therefore not be expected. The data from the present experiment were therefore not expected to follow this hypothesis. Moreover, in the present experiment, feedback was not bound to rewards or losses and the magnitude of such was therefore also not varied. It can therefore not be expected that the P3 varies with the absolute magnitude of these rewards or losses (Yeung & Sanfey, 2004; Sato et al., 2005; Sutton, Tueting, Hammer & Hakerem, 1978).

4.1.5 Methods

4.1.5.1 Participants

I tested 17 participants in total, 6 of whom were female. All participants were right-handed and reported normal or corrected-to-normal vision. Their ages ranged from 18 to 25 years ($M = 19.6$). Nine participants were paid £10 per hour for their participation. The other participants received course credit. All testing was approved by the local ethics committee.

4.1.5.2 Task and procedure

In this experiment, participants completed a number of blocks of a dot-count perceptual decision-making task with visual feedback presented after every trial, but not confidence judgements. The data from this part of the exper-

iment were used to measure attention to feedback. A second part of the experiment then followed in which participants no longer received feedback but instead were asked to rate their confidence after every trial. The aim of this second part was to test whether this study would replicate the findings that confidence varied with Pe amplitude on a single-trial level, similar to the findings from Experiment 4. The task was largely similar to the dot-count task reported in previous experiments, differing only in the fact that a staircase procedure was used to assure that for every participant, so that there would be a sufficient number of both detected and undetected errors. More precisely, during the first four blocks, which were 64 trials long each, participants were asked to respond as accurately as possible and difficulty was adjusted so that participants committed about 15% errors – a substantial proportion of them undetected. In blocks 7 to 12, which were each 32 trials long, participants were asked to respond as quickly as possible, so that error rates were doubled to about 30%, now including both slow data-limitation errors, as well as fast-guess errors (Scheffers & Coles, 2000). This staircase is described in detail in Appendix A.1.

To allow for more fine-grained variations in task difficulty, the dots of this dot-count task were now much smaller: Instead of 7-pixels wide circles, as in Experiment 4, the dots were now 2-pixel wide squares that were arranged in two 64 times 64 large fields. This resulted in 4096 possible locations for the dots. If, for example, difficulty was set to 100 dots, that means there was a difference of a 100 dots between the two fields: One field contained $2048 + 50 = 2098$ dots, the other one $2048 - 50 = 1998$ dots. All other details regarding the task remained the same and will therefore not be described here again. In the following two sections, I describe both main parts of the experiment: the feedback part and the confidence part.

Feedback part. The feedback part began immediately after the staircase had determined a suitable level of difficulty that resulted in both detected and undetected errors. Participants first performed 32 practice trials without a performance judgement but instead with visual feedback after each trial: The letters "COR" for correct, and "ERR" for error were displayed left or right of the fixation cross. The side on which they were displayed was independent of the key pressed or the stimulus and participants were explicitly informed of this independence. Instead, the side was manipulated so that lateralised measures of attention, particularly the N2pc, could be investigated with this paradigm. On the other side, the letters "XXX" were presented in each case. Blocks 14 to 23 were experimental blocks. They differed from the practice block only in being 64 trials long.

A typical stimulus sequence of this part is presented in Figure 60A. There was an 800 ms interval between the participant's response and the onset of the visual feedback. Feedback was then presented for 1 second. Then followed a 200 ms break (with just the fixation cross shown on screen) before the next trial began.

Confidence part. The goal of the last part of the experiment, which followed immediately after the feedback part, was to measure decision confidence and ERPs related to confidence. These ERPs were then used to train the classifier with which confidence could then be 'read out' on the remaining trials of the experiment. Visual feedback was no longer given in this part. Participants first completed 32 practice trials, to familiarise themselves with the 6-point confidence scale (as used in previous studies). After this block, the frequencies of the different confidence categories were presented on screen so that the experimenter could discuss these with them and encourage them to use the

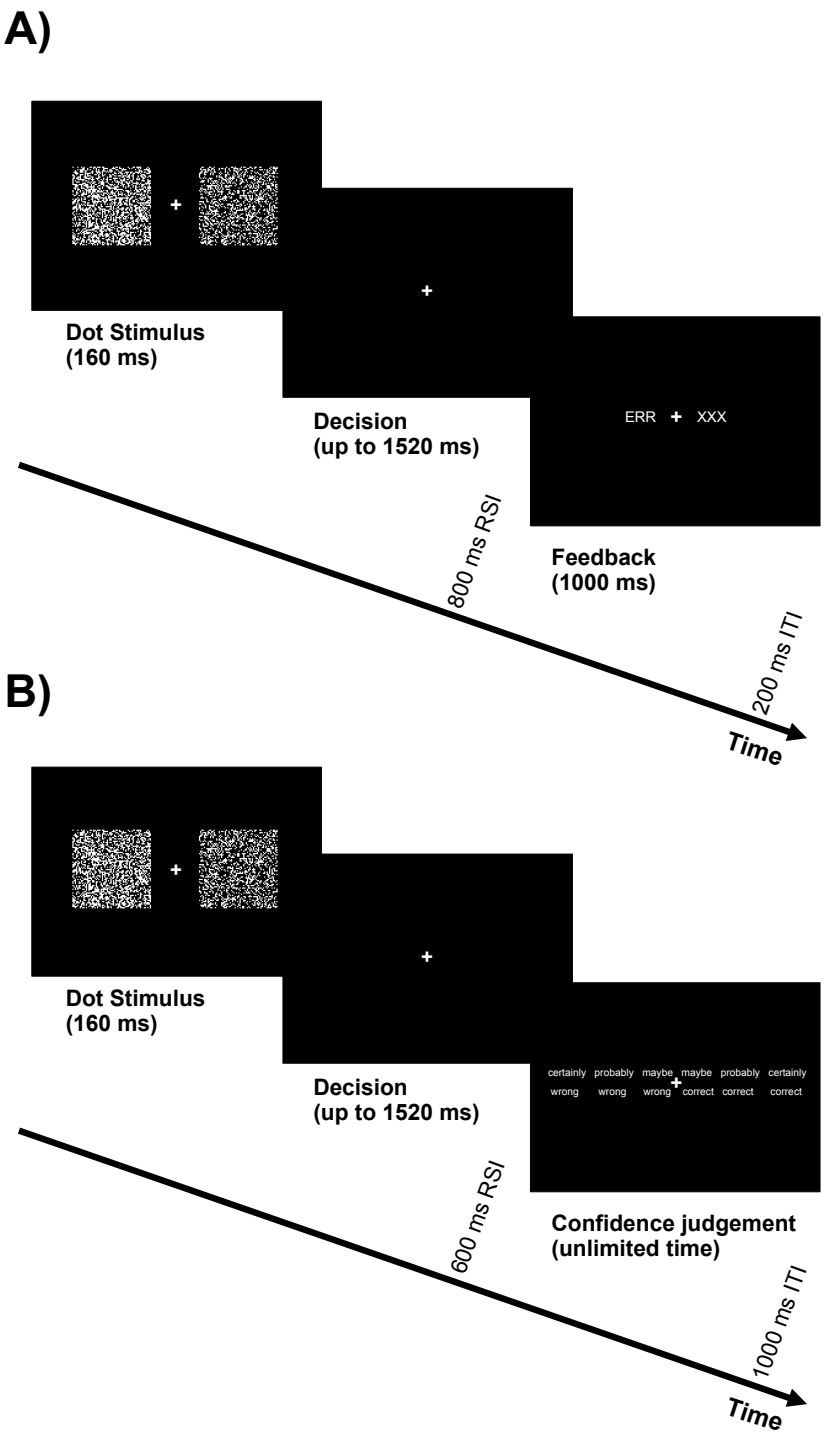

Figure 60: Methods of the dot task. Participants first had to state which of two fields contained more dots by pressing the left or right key. After their response, A) feedback or B) the confidence scales was presented on screen. In the latter case, participants were given unlimited time to choose how confident they were that their last response was correct. RSI: response-stimulus interval; ITI: inter-trial interval; ms: millisecond.

whole scale, if necessary. Blocks 25 and 26, each 64 trials long, then continued with these confidence ratings. As previously reported, the confidence scale – as well as the error detection scale used in blocks 7 to 12 – was counterbalanced over participants.

Figure 60B shows a typical trial sequence for this part – as in previous experiments, there was a 600 ms RSI between the participant's dot decision and the onset of the confidence scale. Participants were given unlimited time for the confidence decision, but were asked to respond according to their first impression. After the confidence judgement, there was a 1 second ITI before the next trial began.

4.1.5.3 EEG recording

As in Experiment 4, EEG data were recorded from 32 scalp locations. The methods for EEG recording were as described in Section 3.1.1.3. For one of the participants, electrode POZ had to interpolated using data from electrodes PZ and OZ. For another participant, electrode F4 had to be interpolated using data from electrodes FP2, F8, FC4, and FZ.

Response-locked EEG data was baselined to -100 to 0 ms pre-response. In the context of Experiment 4, I have presented data in support of the hypothesis that Pe amplitude of the Pe varies with decision confidence. I predict that the same relationship will hold here, but if this was the case then baselining feedback-locked data to -100 to 0 ms pre-feedback could lead to artificial condition differences for the feedback-locked data. Instead, I therefore baselined all feedback-locked EEG epochs -100 to 0 ms pre-stimulus, as illustrated in Figure 61.

In the following four sections, I aim to find suitable electrode locations and time windows for the analyses of three of the ERP components: SPN,

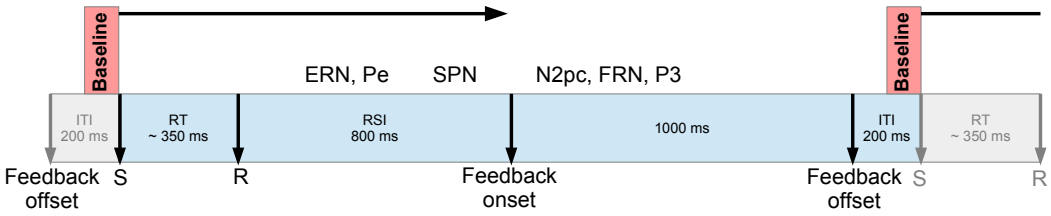

Figure 61: Schematic example of a typical trial, highlighted in blue. The pre-stimulus baseline for each respective trial is shown in red. The event-related potentials that were of key interest in the present experiment are presented on top of time line. ERN: error-related negativity; Pe: error positivity; FRN: feedback-related negativity; S: stimulus; R: reaction; RT: response time; RSI: response-stimulus interval; ITI: inter-trial interval; ms: millisecond.

N2pc, and P3. Subsequent analyses then focused on these electrodes and time windows, as an attempt to reduce complexity of the analyses. Details on how the FRN was calculated are also given.

SPN. Given the slow development of the SPN and its wide distribution over the scalp, my analysis focused on four different time windows: -800 to -600 ms, -600 to -400 ms, -400 to -200 ms, and -200 to 0 ms, with 0 ms referring to the onset of the feedback stimulus. I analysed electrodes F3, FZ, F4, FC3, FCZ, FC4, C3, CZ, C4, CP3, CPZ, CP4, P3, PZ, and P4, divided into three lateral scalp locations (left, right, middle) and five anteroposterior scalp locations (frontal, fronto-central, central, posterio-central, and posterior). The last factor was of course classifier quartile or bin.

All four factors were then submitted to a repeated-measures ANOVA: lateral and anteroposterior scalp locations, time window, and classifier bin. This analysis revealed a reliable main effect of anteroposterior scalp location, $F(1.7,\ 26.4) = 16.8,\ p < 0.001,\ \eta_p^2 = 0.51$, caused by the fact that the SPN was found to be largest at frontal electrodes, $M_F = 0.3\ \mu V$. There was also a reliable effect of lateral location, $F(2,\ 32) = 13.6,\ p < 0.001,\ \eta_p^2 =$

0.46, caused by the fact that the SPN was weakest on midline electrodes, $M_{middle} = 3.1\ \mu V$. Moreover, lateral and anteroposterior scalp location showed a significant interaction, $F(8,\ 128) = 2.7,\ p < 0.01,\ \eta_p^2 = 0.14$: The SPN was strongest (i.e., most negative) at electrode $F3 = -0.1\ \mu V$. The main effect of window was also reliable, $F(1.7,\ 26.9) = 31.8,\ p < 0.001,\ \eta_p^2 = 0.67$, given that the SPN was strongest in the third time window ranging from -400 to -200 ms, $M_3 = 0.3\ \mu V$. Taken together, these results indicate that the SPN was strongest at electrode F3 in the last two time windows. All following analyses will focus on only this electrode in these two time windows. For the sake of completeness, the results for the full four-way ANOVA model are given in Appendix A.2.

N2pc. Pairs of electrodes, one on each side of the scalp, were selected for the analysis of the N2pc, that is P7 and P8, P3 and P4, and O1 and O2. The crucial factor is the side on which the feedback was presented, but as a control factor, the side on which participants had previously pressed the response key was also included into this analysis. The N2pc was furthermore defined as the average voltage in a 60 ms wide time window. Four different time windows were analysed, ranging from 180 to 240 ms, 210 to 270 ms, 240 to 300 ms, and from 270 to 330 ms. The last factor for this analysis was the classifier bin, which is the key factor of interest here.

The purpose of this analysis was to reduce complexity and find the electrode pair and time window at which the N2pc was strongest and to also assess whether response side influenced the N2pc, which it should not. The data were submitted to a five-way repeated-measures ANOVA with the above-described factors as independent variables and mean voltage as dependent variable. There was a reliable main effect of the electrode pair, $F(2,\ 32) =$

17.5, $p < 0.001$, $\eta_p^2 = 0.52$, as well as a marginally reliable interaction between electrode pair and stimulus side, $F(2,\ 32) = 2.7$, $p = 0.08$, $\eta_p^2 = 0.14$. Post-hoc analyses suggested that the most robust and strongest N2pc effect can be measured at the electrode pair P7/P8. There was furthermore a reliable main effect of time window, $F(1.8,\ 29.5) = 4.2$, $p = 0.03$, $\eta_p^2 = 0.21$, as well as a reliable interaction between time window and stimulus side, $F(1.6,\ 25.7) = 11.5$, $p = 0.001$, $\eta_p^2 = 0.42$. Post-hoc tests furthermore revealed that the effect of stimulus side was strongest during the second time window, $M_2 = \eta_p^2 = 0.38$. A final step to reduce the complexity of the ANOVA model was to test whether response side had a reliable effect. It was hypothesised that the influence of whether the response hand used was ipsi- or contralateral to the N2pc measured would have no significant effect. This was indeed the case, $F < 1$. This factor showed no interaction with classifier bin, $F < 1$. Further analyses will therefore exclude this factor and furthermore focus on the stimulus pair P7/P8 during the second time window ranging from 210 to 270 ms post feedback onset. For the sake of completeness, all effects from the six-way ANOVA model are presented in Appendix A.4.

FRN. The FRN was defined as a base-to-peak difference in voltage, as done for example by Yeung and Sanfey (2004). More specifically, the FRN was defined as the difference between the first major positivity (50-300 ms post-stimulus) and the subsequent negative peak that occurred before the large amplitude P3 component: For each participant, time windows were shifted until the early and late P3 peaks were identified. The FRN was then quantified as the most negative point in-between these P3 peaks. To calculate the base-to-peak measure, an average of the P3 peak amplitudes was taken. The FRN effect was then quantified as the difference between this average P3 and the

measured FRN peak.

P3. P3 amplitude was quantified as the average voltage in a 100 ms wide time window. Four different time windows were used in this analysis. The first time window ranged from 100 to 200 ms, the second from 140 to 240 ms, the third from 180 to 280 ms, and the fourth from 220 to 320 ms. The data were submitted to a three-way ANOVA with electrode, bin, and accuracy and independent variables. Five midline electrodes were considered in this analysis: FZ, FCZ, CZ, CPZ, and PZ. The electrode at which the P3 was measured had a reliable effect on the size of the P3, $F(1.8,\ 29.3) = 15.1,\ p < 0.001,\ \eta_p^2 = 0.48$: P3 amplitude was largest at electrode CZ, $M_{CZ} = 6.6\ \mu V$. There was also a reliable main effect of time window, $F(1.5,\ 24.0) = 27.9,\ p < 0.001,\ \eta_p^2 = 0.64$, with the largest P3 amplitude recorded in the third time window. The factors electrode and time window interacted significantly, $F(1.7,\ 27.5) = 6.0,\ p < 0.01,\ \eta_p^2 = 0.27$. The P3 was strongest at electrode CZ during the third time window (180 to 280 ms), $M = 7.9\ \mu V$. All further analyses will focus on this electrode and time window. For the sake of completeness, however, the full four-way ANOVA model is reported in Appendix A.6.

4.1.5.4 Data analysis

I fitted a classifier to the data, trained to differentiate between objectively correct and incorrect responses. The training set for the classifier was composed of half of the error trials from the feedback part of the experiment and a subset of correct responses that was matched in size to the errors. This classifier could then be applied to data from the feedback part of the experiment to read out confidence from trials on which no such confidence judgements were required. Further details regarding this classifier can be found in Section 3.1.1.4.

4.1.6 Results

4.1.6.1 Behavioural data

The staircase procedure was successful at adjusting difficulty so that there would be both a substantial proportion of detected and undetected errors, as outlined in detail in Appendix A.1. By the end of the experiment, there were 20.4% detected errors and 9.4% undetected errors.

RTs were slightly faster in the feedback part of the experiment compared to the confidence part, $M_{feedback} = 347\ ms$ versus $M_{Confidence} = 390\ ms$. This effect was significant, $F(1,\ 16) = 7.4,\ p = 0.02,\ \eta_p^2 = 0.32$. Moreover, RTs were faster for error than correct trials, $M_{cor} = 382\ ms$ versus $M_{err} = 355\ ms$. This difference was also statistically significant, $F(1,\ 16) = 21.1,\ p < 0.001,\ \eta_p^2 = 0.57$. There was no interaction between the two factors, $F < 1$. Error rates were 29.2% in the feedback part and 29.3% in the confidence part. This difference was not reliable, $t < 1$.

Moreover, individual Spearman's rank correlations were calculated to test whether there was a relationship between participants' objective and subjective accuracy. The average error rates for each confidence level are displayed in the left panel of Figure 62. For 15 out of 17 participants this relationship was negative and reliable, $rs <= -0.83,\ ps <= 0.04$, that means high confidence was associated with low error rates. For the remaining two participants this relationship was also negative but the correlation was only marginally significant, $rs <= -0.76,\ ps <= 0.08$. The right-hand panel of Figure 62 presents proportions of confidence judgements for correct and incorrect trials. Once more, some overlap can be seen, with *probably wrong* as the mode for error responses and *probably correct* as the mode for correct responses. Taken together, these findings suggests that participants had good resolution in their

metacognitive judgements in this slightly adapted version of the dot task.

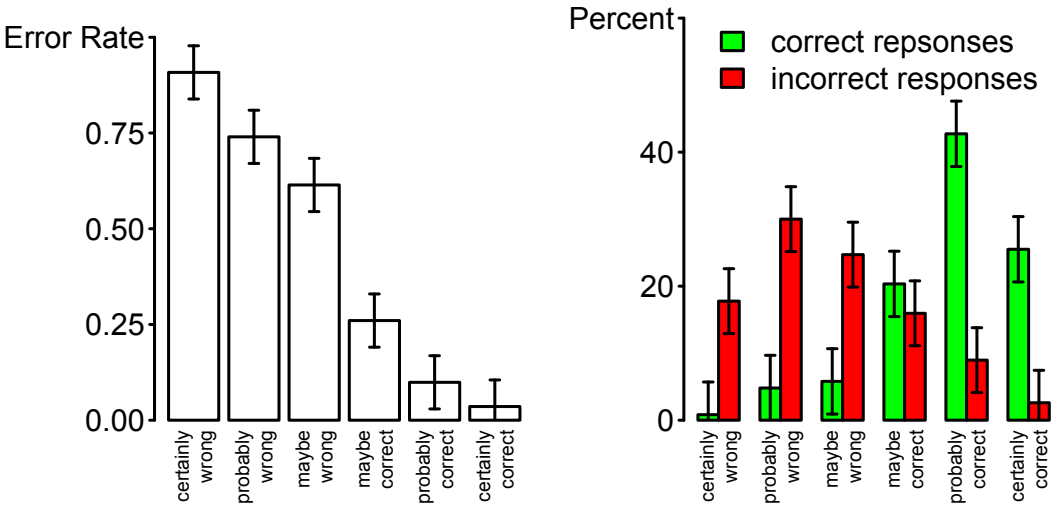

Figure 62: Left panel: Error rates as a function of subjective confidence ratings. Right panel: Distributions of confidence responses as a function of difficulty and objective accuracy. Error bars represent within-subject confidence intervals for the confidence factor within each difficulty condition.

Participants judged error trials with an average confidence of $M_{err} = 2.76$ and correct trials with an average confidence of $M_{cor} = 4.76$. This difference was reliable, $t(16) = 17.8$, $p < 0.001$. The average metacognitive sensitivity was $meta\text{-}d' = 3.39$, while the average first-order sensitivity was $d' = 1.31$. Combined, this resulted in an average metacognitive efficiency of $log(M\text{-}ratio) = 0.45$. These metacognitive efficiency parameters were significantly different from 0, $t(16) = 5.3$, $p < 0.001$. This means participants used more evidence in their metacognitive judgement compared to the dot-decision.

Taken together, the behavioural results reported here suggest that this slightly modified version of the dot-count task led to similar first- and second-order effects: a substantial number of errors of which about one third was undetected, however overall good metacognitive insight.

4.1.6.2 EEG results

Error-related ERPs. The purpose of the confidence part of this experiment was to train a classifier to distinguish between correct and error trials, and to then test whether this classifier could be used to estimate confidence on a single-trial level – as shown in Experiment 4. The confidence part therefore contained such metacognitive judgements to assess whether such single-trial measures of Pe amplitude vary with confidence. The rationale was to then apply this discriminating component to data from the feedback part to 'read out' confidence on a single-trial level and measure how confidence modulates attention to feedback. However, before fitting this classifier, I first had to test whether both the ERN and the Pe vary with not only with objective accuracy, but also decision confidence when averaged over trials, as found for Experiment 4.

First, Figure 63 shows the ERN and Pe at electrode CZ as a function of objective accuracy. This analysis included trials from both the feedback and the confidence part of the experiment. Averaged data from the ERN time window (-40 to 60 ms), measured at the five midline electrodes (FZ, FCZ, CZ, CPZ, and PZ) were then submitted to a repeated-measures ANOVA with location and accuracy as independent variables. Replicating the findings from Experiment 4, there was a reliable main effect of accuracy, $F(1,\ 16) = 43.1,\ p < 0.001,\ \eta_p^2 = 0.73$, with a stronger (i.e., more negative) ERN for error, $M_{err} = 1.0\ \mu V$ as compared to correct trials, $M_{cor} = 2.1\ \mu V$. There was also a reliable effect of anteroposterior scalp location, $F(1.3,\ 20.3) = 18.7,\ p < 0.001,\ \eta_p^2 = 0.54$. The two factors showed a reliable interaction, $F(1.7,\ 27.5) = 8.4,\ p < 0.01,\ \eta_p^2 = 0.34$, reflecting that the strongest difference between correct and error trials was found at electrode FCZ, $M_{FZ} = 1.1\ \mu V$,

$M_{FCZ} = 1.5\ \mu V$, $M_{CZ} = 1.4\ \mu V$, $M_{CPZ} = 1.1\ \mu V$, $M_{PZ} = 0.8\ \mu V$.

The same analysis was repeated for the Pe time window (250 to 350 ms following the response), again for both parts of the experiment. There was a reliable effect of accuracy, echoing findings from EXPERIMENT 4, $F(1,\ 16) = 26.7,\ p < 0.001,\ \eta_p^2 = 0.63$. The Pe being a positive component, this meant that the amplitude was larger for error, $M_{err} = 3.1\ \mu V$, compared to correct trials, $M_{cor} = 0.7\ \mu V$. There was also a reliable main effect of scalp location, $F(1.7,\ 27.6) = 28.6,\ p < 0.001,\ \eta_p^2 = 0.64$. Crucially, the difference between correct and error trials was strongest at posterior electrode PZ, $M_{PZ} = -3.7\ \mu V$, compared to the other electrodes, $M_{FZ} = -0.4\ \mu V$, $M_{FCZ} = -1.7\ \mu V$, $M_{CZ} = -2.8\ \mu V$, $M_{CPZ} = -3.4\ \mu V$. This interaction effect was reliable, $F(1.3,\ 20.7) = 18.2,\ p < 0.001,\ \eta_p^2 = 0.53$.

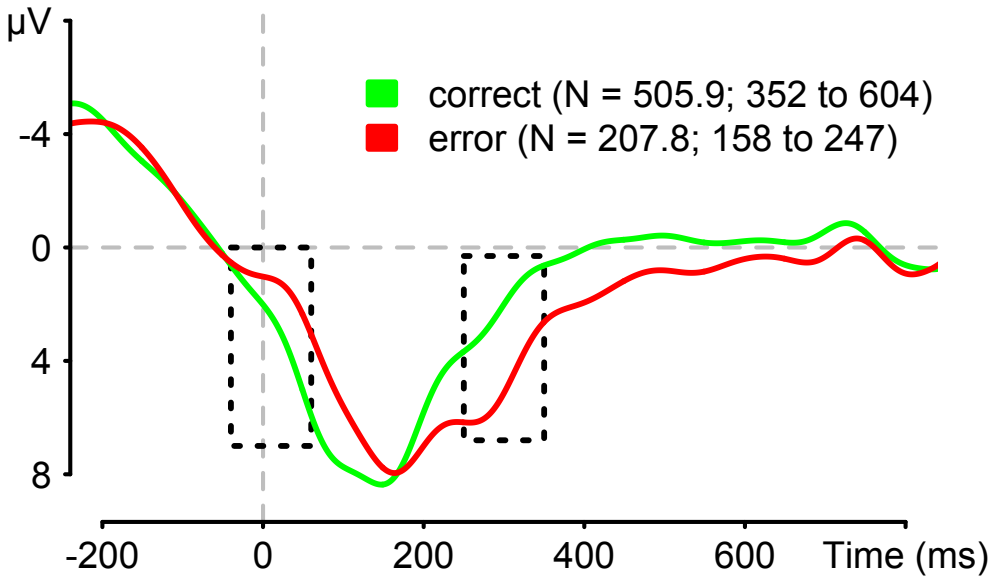

Figure 63: Error-related negativity (ERN) and error positivity (Pe) at electrode CZ, conditioned on objective accuracy; response-locked event-related potential (ERP). The two windows highlight the ERN (-40 to 60 ms) and the Pe (250 to 350 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

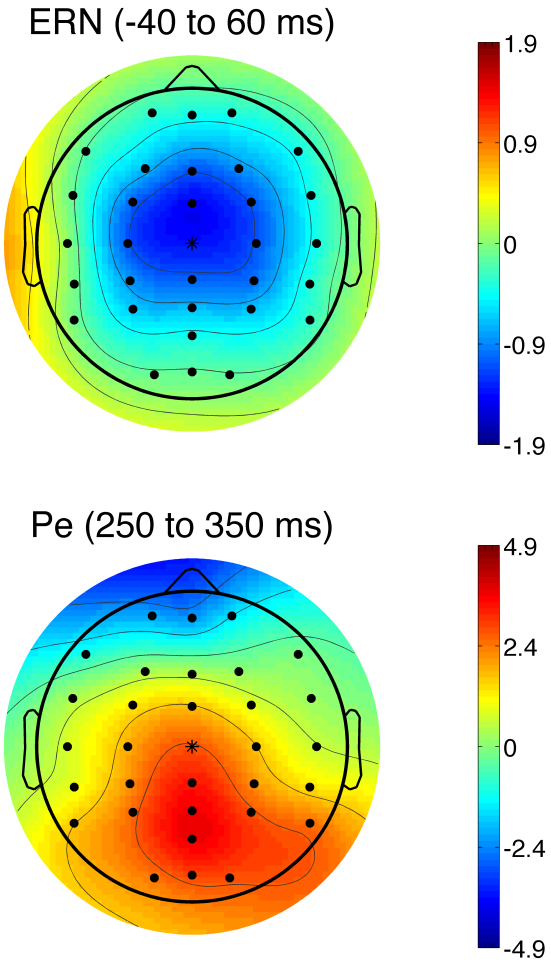

Figure 64: Topographies for the difference between errors and correct trials for both the ERN (top panel) and the Pe (bottom panel). The colours in the topographic plots indicate different values in micro-volt; ms: millisecond.

Figure 65 presents the averaged ERP data collapsed across all trials from the confidence block, across participants and across both correct and error trials. The ERN (-40 to 60 ms; upper panel of Figure 65) was found to be strongest (i.e., most negative) at the most frontal electrode FZ, $M_{FZ} = -0.2\ \mu V$, compared to all other electrodes, $M_{FCZ} = 1.1\ \mu V$, $M_{CZ} = 2.3\ \mu V$, $M_{CPZ} = 2.7\ \mu V$, $M_{PZ} = 2.4\ \mu V$. This main effect of anteroposterior scalp location was found to be reliable, $F(1.3,\ 21.0) = 17.2,\ p < 0.001,\ \eta_p^2 = 0.52$. The main effect of confidence was not reliable, though, $F(2.7,\ 43.3) = 1.7,\ p = 0.19,\ \eta_p^2 = 0.10$. There was, however, a reliable linear trend, $F(1,\ 16) = 6.1,\ p = 0.03,\ \eta_p^2 = 0.28$, reflecting a monotonic increase of the ERN amplitude from the lowest to the highest confidence level, $M_1 = 0.9\ \mu V$, $M_2 = 1.2\ \mu V$, $M_3 = 1.4\ \mu V$, $M_4 = 1.8\ \mu V$, $M_5 = 2.1\ \mu V$, $M_6 = 2.4\ \mu V$. The two factors did not interact reliably, $F(3.5,\ 56.2) = 1.0,\ p = 0.42,\ \eta_p^2 = 0.06$.

The Pe (250 to 350 ms; lower panel of Figure 65) was again modulated by decision confidence, $F(3.2,\ 50.5) = 4.1,\ p = 0.01,\ \eta_p^2 = 0.20$, also with a significant linear trend, $F(1,\ 16) = 10.7,\ p < 0.01,\ \eta_p^2 = 0.40$. There was also a reliable effect of scalp location, $F(1.7,\ 27.7) = 15.6,\ p < 0.001,\ \eta_p^2 = 0.49$. The two factors interacted reliably, $F(3.1,\ 49.7) = 2.8,\ p < 0.05,\ \eta_p^2 = 0.15$. If all five electrodes were analysed separately, the effect of confidence was numerically largest at the most posterior electrode PZ, $\eta^2_{p(PZ)} = 0.29$, compared to all other electrodes, $\eta^2_{p(FZ)} = 0.08$, $\eta^2_{p(FCZ)} = 0.12$, $\eta^2_{p(CZ)} = 0.20$, $\eta^2_{p(CPZ)} = 0.27$. It is furthermore worth noting that these effects were found despite the low trial numbers that contributed to each of these cells, as shown in the legend of Figure 65. For example, there were on average only seven trials contributing to the *certainly wrong* category. The upper panel of Figure 66 shows the topography for the ERN as a difference between the two outermost categories (*certainly wrong* and *certainly correct*). Such a central scalp topography is

typical for the ERN. The lower panel of the same Figure shows the topography of the Pe, which is more posterior, matching what is usually found for the Pe.

Taken together, these effects suggest that – as shown in the previous chapter – both the ERN and the Pe vary with objective accuracy and decision confidence. This finding is crucial as the main analyses were based on a classifier trained on the effect of accuracy and confidence on Pe amplitude.

Single-trial EEG data. The key question of this study was whether confidence modulated how much attention participants paid to feedback. I therefore had to estimate participants' confidence on a trial-by-trial level. In Experiment 4, I presented a method that allows precisely such a 'read out' by estimating Pe amplitude on a trial-by-trial basis. The next analysis will therefore apply the same linear integration method to the data to derive a discriminating component, which maximally distinguishes waveforms on correct and error trials.

I first repeated the analysis from the previous experiment to assess whether this method can be used to read out confidence from trials on which no such confidence judgements were required. The training set for the classifier was composed of all error trials from the feedback part of the experiment and a subset of correct responses that was matched in size to the errors. The trials were also matched with regard to the response hand to avoid the classifier being trained to motor activity. On average, the test sets contained 345 trials, ranging from 276 to 412 trials. Different time windows were again tested and compared with the result that the window from 250 to 350 ms post-response yielded the best results, that is the highest AUC of 0.75 if applied back to the test set (Figure 67). This finding matched those of Experiment 4.

Moreover, a k-fold cross-classification method was applied to test how

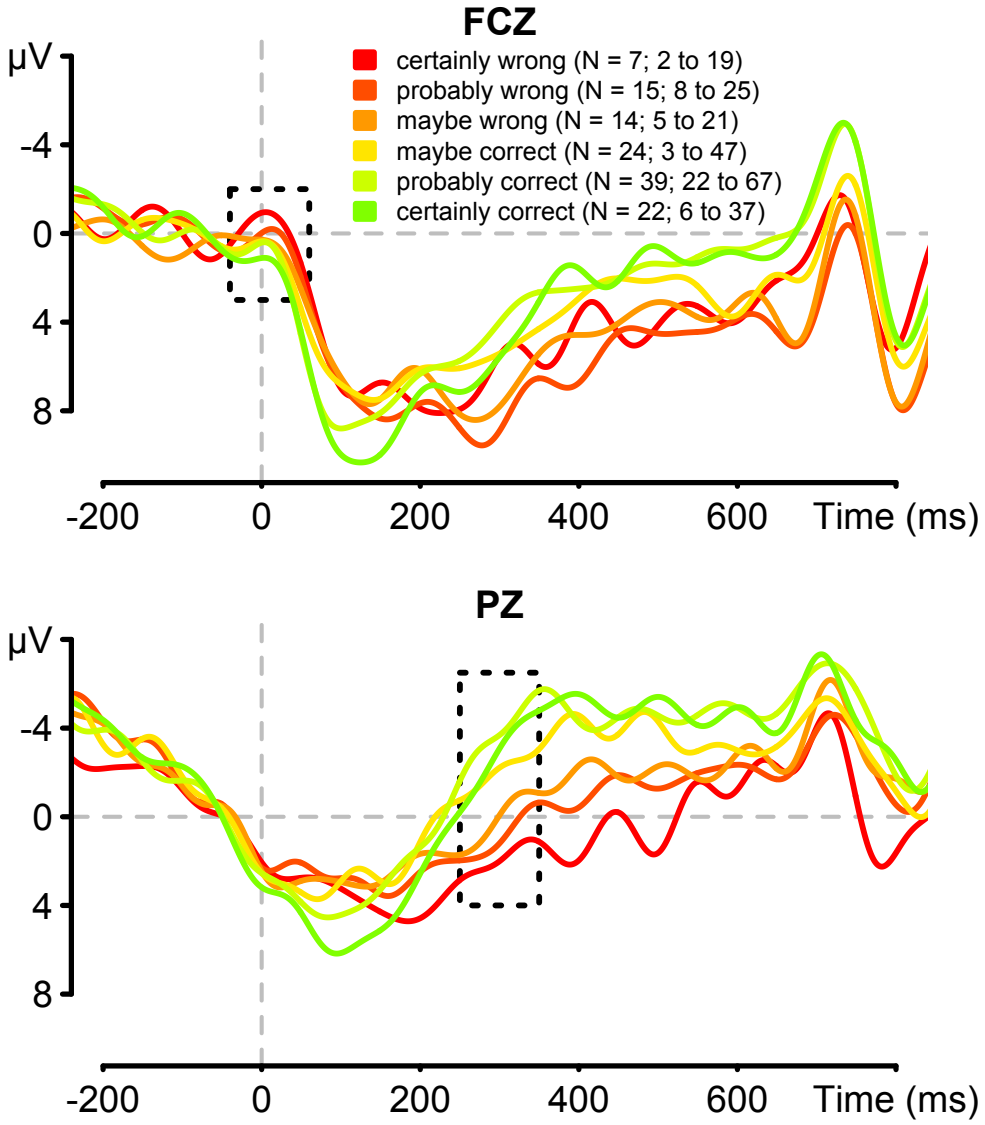

Figure 65: Top panel: Error-related negativity (ERN) at electrode FCZ; bottom panel: error positivity (Pe) at electrode PZ. Both response-locked event-related potential (ERP) were conditioned on subjectively-rated confidence. Plots show data combined for objectively correct and incorrect trials. The two windows highlight the ERN (-40 to 60 ms) and the Pe (250 to 350 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

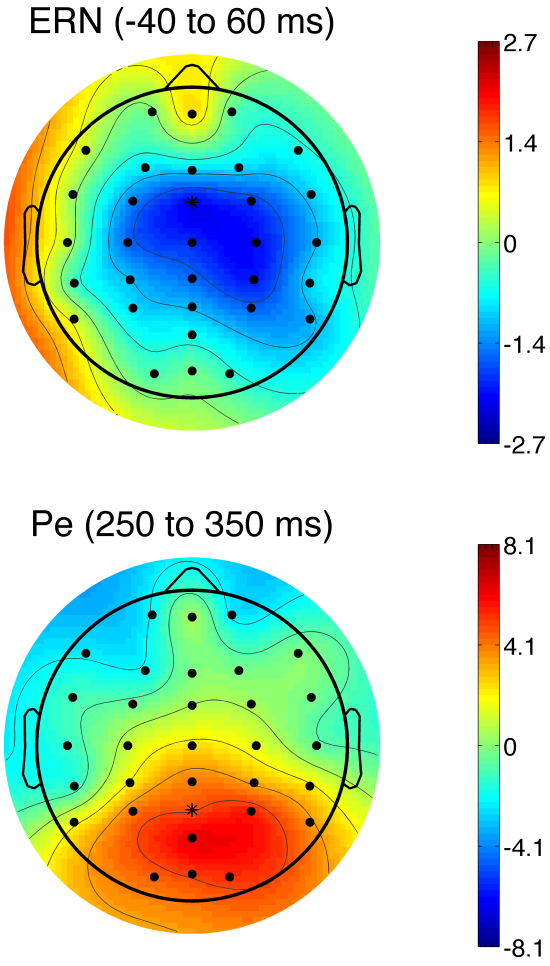

Figure 66: Topographies for the difference between *certainly wrong* and *certainly correct* condition for both the ERN (top panel) and the Pe (bottom panel). The colours in the topographic plots indicate different values in microvolt; ms: millisecond.

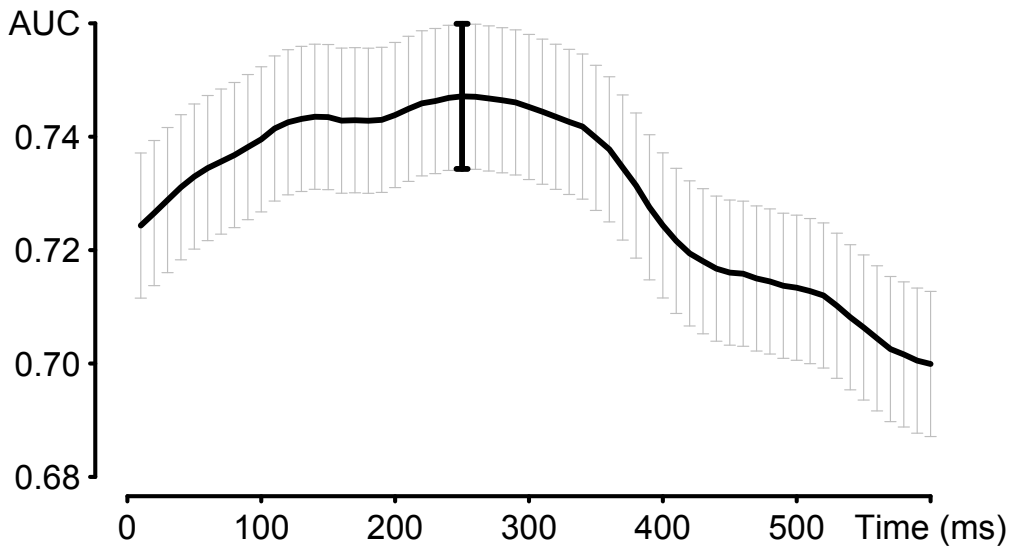

Figure 67: Classification performance (area under curve; AUC) when applied back to the training set for a range different time windows form which the training data was taken. The x-axis specifies the starting point from which onwards a window of 100 milliseconds (ms) width was extracted.

consistent this classifier was. All data from the feedback part of the experiment were divided into four different folds. The classifier was then trained on data from folds 1 to 3 and tested on fold 4, and so forth. The average AUC was 0.64 ($min = 0.55$; $max = 0.73$). This is comparatively low, already hinting at the fact that this classifier might lack the power necessary to detect the predicted data patterns.

Figure 68 furthermore presents classification performance over time using the same shifting-window approach described above (Figure 67), but for average 4-fold AUC. For this analysis, a slightly later time window was identified ranging from 270 to 370 ms. As for Experiment 4, this time course has an almost identical overall morphology if compared to Figure 67. However, I again chose to use the time window ranging from 250 to 350 ms as it is based on a less noisy estimate.

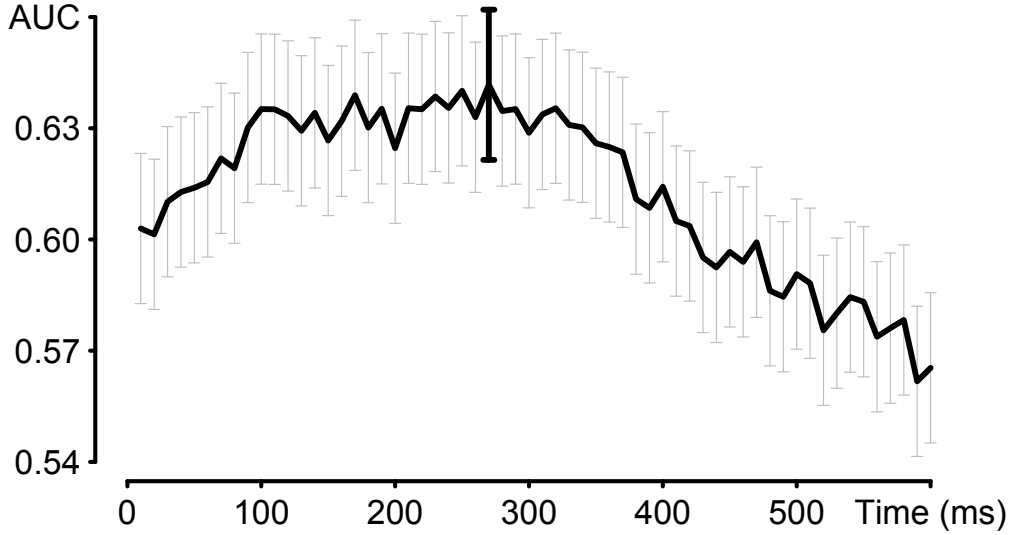

Figure 68: Mean classification performance (area under curve; AUC) using a 4-fold algorithm for a range different time windows form which the training data was taken. The x-axis specifies the starting point from which onwards a window of 100 milliseconds (ms) width was extracted.

Figure 69 shows the time course for the discriminating component. As in Experiment 4, the extracted component corresponds closely to the Pe shown in Figure 63. Indeed, the component differentiated between correct and error trials in the highlighted time window, which was also used for training the classifier ranging from 250 to 350 ms, $t(16) = 14.6,\ p < 0.001$. This also holds for other time windows, for example a window ranging from 350 to 450 ms, $t(16) = 13.0,\ p < 0.001$.

The spatial filter that resulted from this analysis is shown in Figure 70. Once more, the sensor projection corresponded closely to the topography of the Pe (Figures 64 and 66). The classifier can then be applied to all of the correct trials from the confidence part of the experiment. For each participant, there were on average 86 trials ($min = 57$; $max = 107$). The question is again whether this classifier, trained to predict objective accuracy, could also be

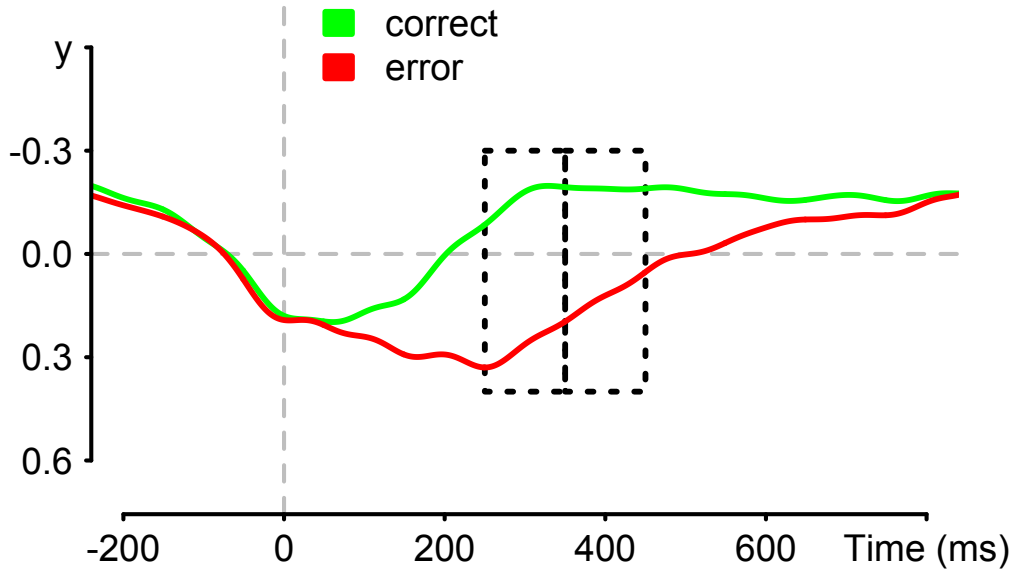

Figure 69: Time course of the discriminating component (y) for the Pe time window, identified by the logistic regression classification analysis of errors versus correct responses, coded in arbitrary units. The window highlight the training window for the classifier (250 to 350 ms) and a second window, assumed to capture later parts of the Pe (350 to 450 ms); ms: millisecond.

used to predict correct-trial confidence. This analysis yields an estimate of Pe amplitude for each time point on each trial. I again averaged these values across a moving time window of 51 ms and then plot it up into quantiles for each time point. I used quartiles instead of quintiles for the present experiment given that there were far fewer trials in each bin for this analysis as opposed to Experiment 4. Quartile 1 is the smallest Pe amplitude and quartile 4 the largest. Average confidence within each quartile can then be calculated, as shown in Figure 71.

Importantly, the finding that confidence varied inversely and monotonically with estimated Pe amplitude on correct trials was replicated here. This effect was not reliable for the classification time window ranging from 250 to 350 ms post-response, $F(3,\ 48) = 2.0,\ p = 0.13,\ \eta_p^2 = 0.11$; although the linear trend was reliable, $F(1,\ 16) = 4.6,\ p < 0.05,\ \eta_p^2 = 0.22$. The effect of quantile

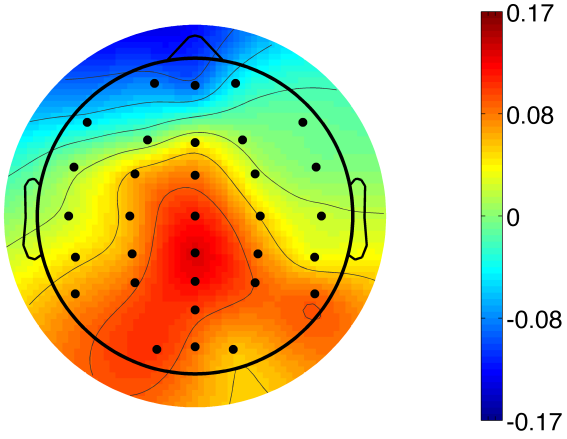

Figure 70: Sensor projection of the discriminating component identified by the logistic regression classification analysis of errors versus correct responses, trained on the Pe time window, coded in arbitrary units.

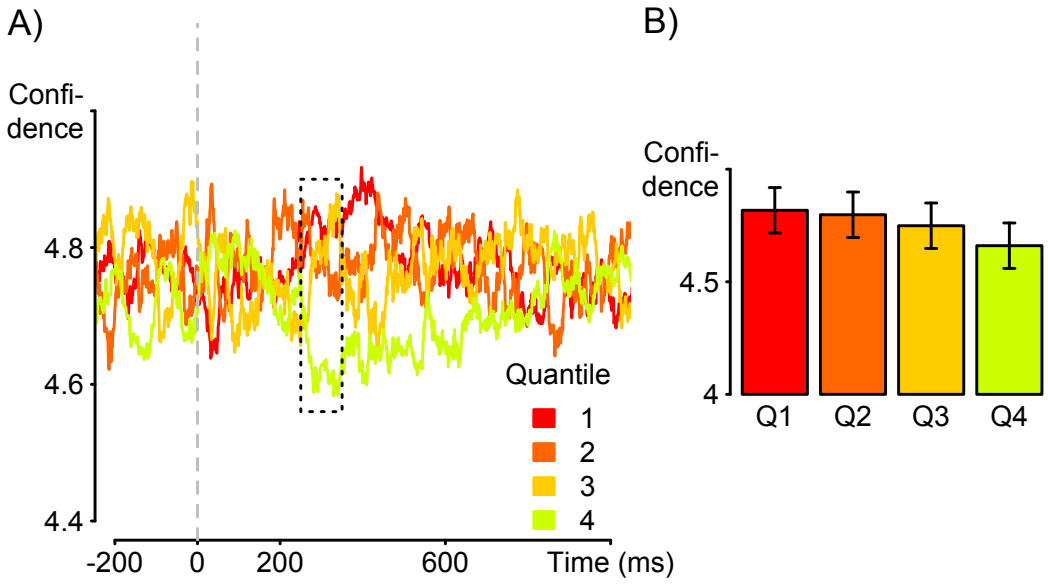

Figure 71: Response-locked, moving average confidence (window width: 51 ms) for quartiles of discriminating component; correct trials only; classifier trained on objective accuracy; ms: millisecond. The window highlight the training window for the classifier (250 to 350 ms). The right panel present averaged confidence over the time window and quartiles.

was present if the time window was shifted forwards by 50 ms, for this time window now both the main effect, $F(2.2,\ 34.9) = 4.3,\ p = 0.02,\ \eta_p^2 = 0.21$, and the linear trend were reliable, $F(1,\ 16) = 6.5,\ p = 0.02,\ \eta_p^2 = 0.29$. Moreover, the gradations of confidence were once more observed around a high average value of almost 5 (*probably correct*), similar to findings from the previous experiment. Taken together, the analyses reported here suggest that Pe amplitude indeed varied with correct-trial confidence – similar to what had previously been suggested in the context of Experiment 4 – even though the relationship seemed less stable than for the previous experiment. Here, however, I attempt to use this classifier to 'read out' confidence from the feedback part; that is, I take classifier bin as an index of confidence on feedback blocks.

The next four sections will focus on each of the four feedback-related components, analysing whether confidence had an effect on how participants processed feedback. The first two components, the SPN and the N2pc, are interpreted as reflections of attention allocated to feedback, while the other two components, the FRN and the P3, are taken to reflect surprise.

Expectation effects (SPN). The first analysis focuses on the SPN, which was expected to vary with certainty, but not confidence. More precisely, I expected to find the SPN to be largest (i.e., most negative) for bins 2 and 3 of the data classified according to the decoded confidence level. The upper left panel of Figure 72 shows the SPN at F3 as a function of classifier bin, which was the electrode at which this potential was found to be strongest (see Section 4.1.5.3). Highlighted in this figure are the two time windows during which this ERP was found to be most pronounced, ranging from -400 to -200 ms, and -200 to 0 ms. Average voltage for each of these time windows and bins is presented in the upper right panel of Figure 72. The topographies for both

time windows and different contrasts are presented in Figure 73.

These data were then submitted to a repeated-measures ANOVA with two independent variables: classifier bin and time window. There were no main effects, $Fs < 1$ and also no reliable interaction, $F < 1$. The key hypothesis regarding the SPN was that participants would direct their attention towards the feedback more in the second and third quartile. This corresponds to a quadratic trend. However, visual inspection of the bar plots in the upper right panel of Figure 72 instead suggests that the effect was the opposite – the SPN amplitude was stronger (i.e., more negative) for the high certainty quartiles, as if participants were paying attention to the feedback to confirm their expectations. This effect was not reliable though, $F(1,\ 16) = 1.7,\ p = 0.21,\ \eta_p^2 = 0.10$. The linear trend was also not reliable, $F < 1$. Taken together, these results indicate that – in contrast to the hypothesis – the amplitude of the SPN did not vary as a function of certainty. Nor did SPN amplitude vary as a function of overall response accuracy, $Fs < 1$, as revealed in an ANOVA with factors accuracy and time-window (see also Appendix A.3 and the lower panels of Figure 73).

Attention effects (N2pc). The next set of analyses focused on the N2pc, an EEG component thought to reflect spatial attention. The upper panels of Figure 74 show difference waves as a function of these factors at electrode pair P7/P8, which was identified as the electrode at which the strongest N2pc effect was measured (see Section 4.1.5.3). Highlighted in this figure is the time window from 210 to 270 ms that was identified as the time during which the N2pc effect was strongest. Moreover, Figure 75 shows the topographies for this time window. The right-hand panel of this figure (contrast error versus correct trials) presumably shows eye-movement artefacts. Even though participants

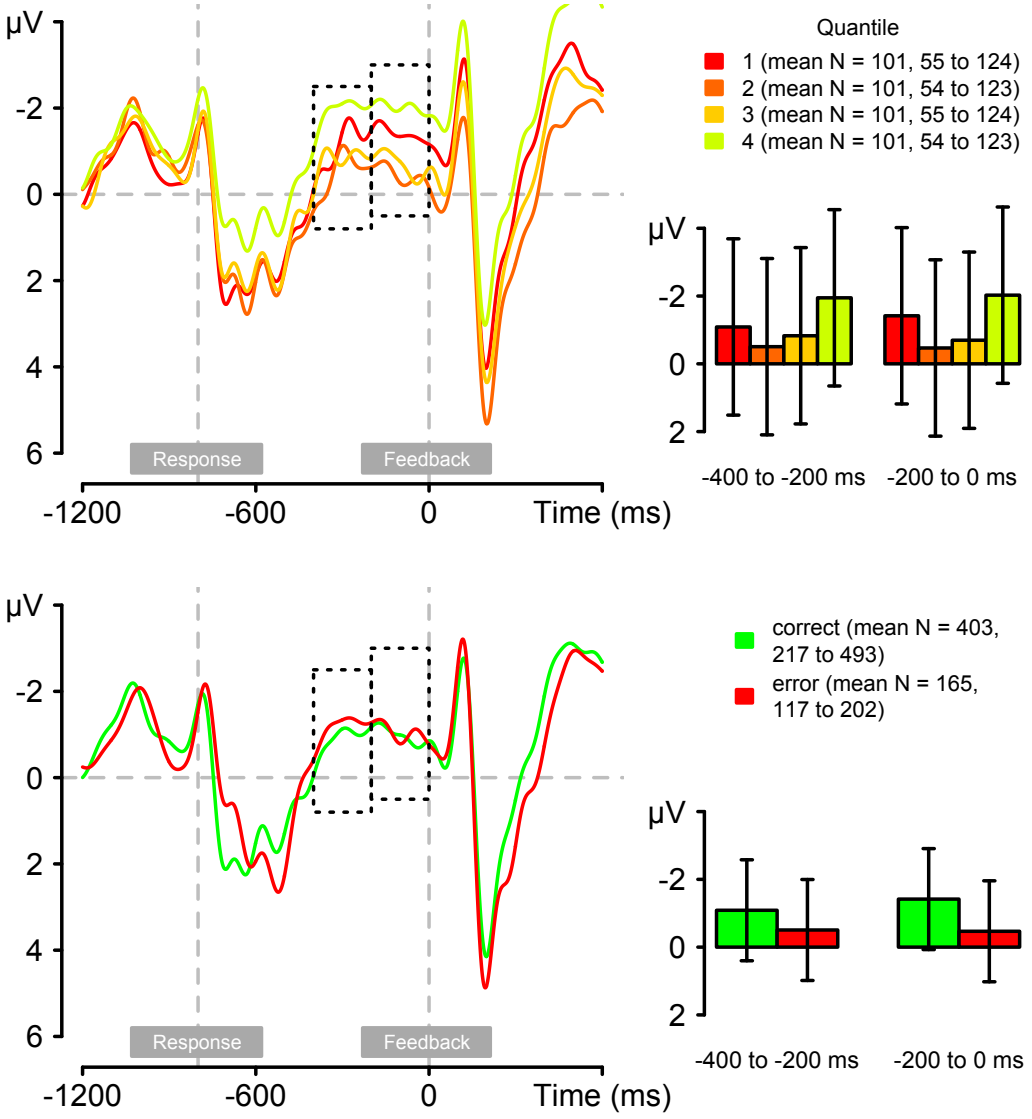

Figure 72: Feedback-locked stimulus-preceding negativity (SPN) at electrode F3, conditioned on single-trial Pe amplitude, reaching from smallest (1) to largest (4) quantile. The data were baselined to -100 to 0 ms pre-stimulus, and therefore to a different time window for each trial depending on RT. The two windows highlight the time during which this component was most pronounced, as identified in previous analyses (-400 to -200 ms and -200 to 0 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

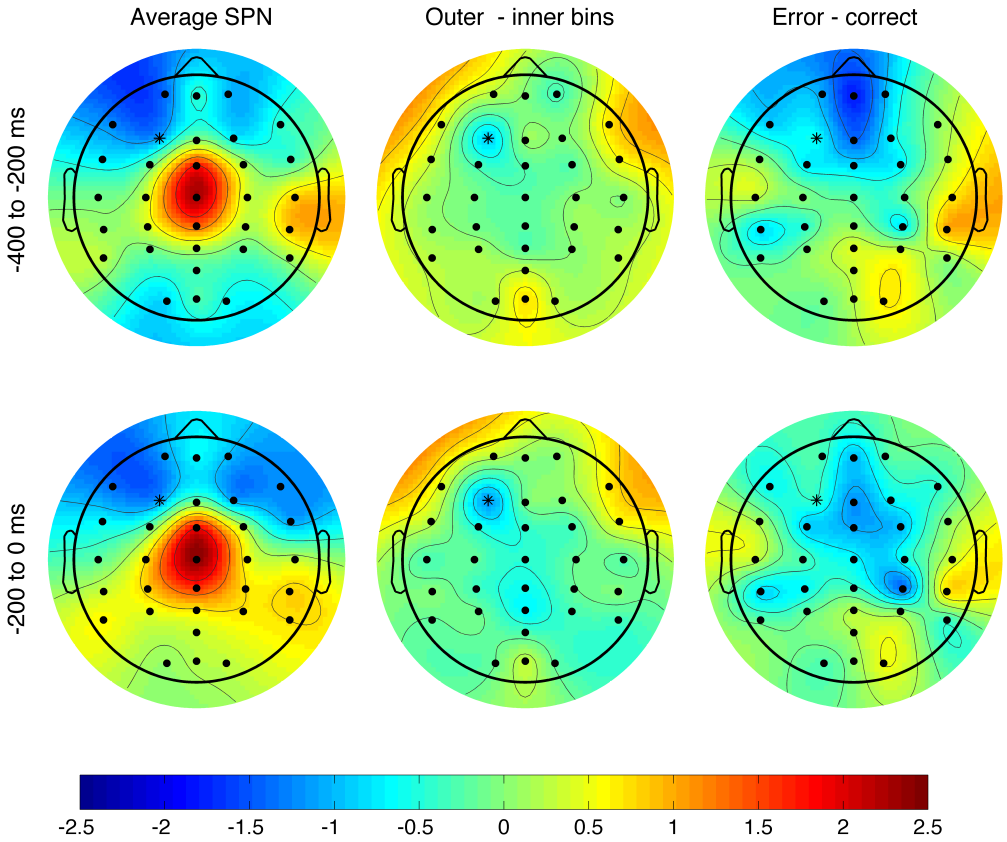

Figure 73: Topographies for the feedback-locked stimulus-preceding negativity (SPN). The upper panels show the data during the earlier time window (-400 to -200 ms); the lower panels during the later time window (-200 to 0 ms). The left panels present the average SPN; the medium panels the difference between the inner and the outer bins; and the right panels the difference between errors and correct trials. The colours in the topographic plots indicate different values in micro-volt; ms: millisecond.

had been instructed to keep their eyes focused on the fixation cross throughout the entire trial, this suggests that they did move their eyes when feedback was presented, and that they differed in this behaviour depending on whether the feedback presented was positive or negative. All these data are locked to the onset of the feedback stimulus.

The data were then submitted to a two-way ANOVA with classifier quartile and stimulus side as independent variables. The general N2pc effect that I expected to find for this data was a main effect of stimulus side: We can expect that the signal is more negative on the side contralateral to the stimulus presentation, which shows that participants directed their attention to this field. In other words, the solid bars in Figure 74 should have more negative values than the shaded bars. This effect was reliable, $F(1,\ 16) = 8.8,\ p < 0.01,\ \eta_p^2 = 0.36$. It can therefore be concluded that allocation of attention to the feedback stimulus was reflected in the N2pc.

I next tested whether this effect was modulated by classifier bin. The interaction between stimulus side and classifier bin was not reliable, though, $F < 1$. The precise hypothesis in this context was that the N2pc effect would be larger for the two middle bins, but the quadratic trend was not reliable, $F < 1$. The same held for the linear trend, $F(1,\ 16) = 1.0,\ p = 0.33,\ \eta_p^2 = 0.06$. There was, however, a marginally reliable effect of classifier bin, $F(1.8,\ 28.9) = 2.6,\ p = 0.09,\ \eta_p^2 = 0.14$, with the smallest deflection for the first quartile, $M_1 = 2.1\ \mu V$, compared to the other bins, $M_2 = 3.8\ \mu V$, $M_3 = 3.1\ \mu V$, $M_1 = 4.2\ \mu V$. There was also a marginally reliable linear trend for this effect, $F(1,\ 16) = 3.8,\ p = 0.07,\ \eta_p^2 = 0.19$. Post-feedback activity will further be analysed in the context of the FRN.

Taken together, these findings suggest that even though there was an N2pc effect, that is participants attended to the side on which the feedback

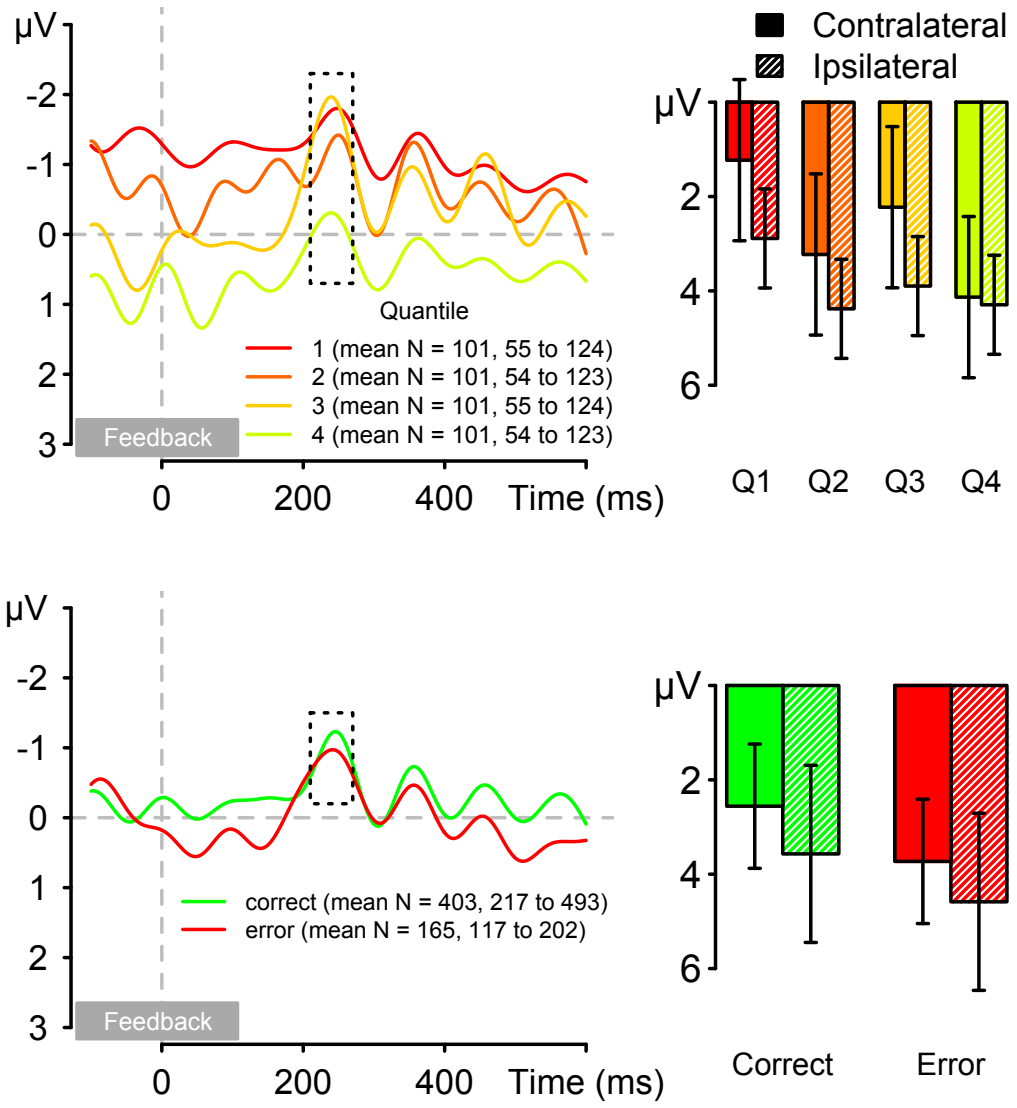

Figure 74: Difference wave (contralateral minus ipsilateral) for the feedback-locked N2pc at electrodes P7 and P8, conditioned on single-trial Pe amplitude, reaching from smallest (1) to largest (4) quantile. The data were baselined to -100 to 0 ms pre-stimulus, and therefore to a different time window for each trial depending on RT. The window highlights the time during which this component was most pronounced, as identified in previous analyses (210 to 270 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

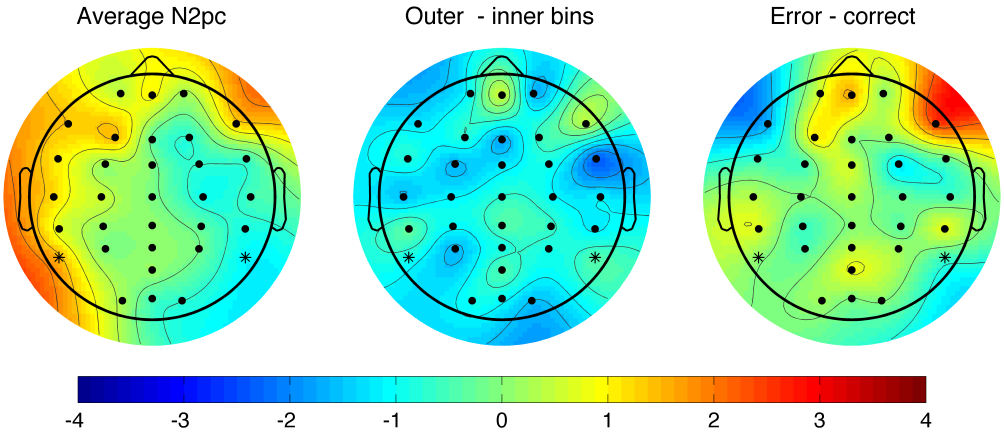

Figure 75: Topographies for the feedback-locked N2pc during the time window from 210 to 270 ms. The left panel presents the average N2pc; the medium panel the difference between the inner and the outer bins; and the right panel the difference between errors and correct trials. The colours in the topographic plots indicate different values in micro-volt; ms: millisecond.

was presented, this effect was not modulated by how confident participants were, as predicted by the classifier. Nor was the N2pc modulated by the valence of the feedback, as shown in the lower panel of Figure 74. In a two-way repeated-measures ANOVA with objective accuracy and stimulus side as dependent variables, there was again a reliable N2pc effect, $F(1,\ 16) = 7.6,\ p = 0.01,\ \eta_p^2 = 0.32$, but there was no effect of objective accuracy, $F(1,\ 16) = 1.1,\ p = 0.31,\ \eta_p^2 = 0.06$, and – critically – no interaction between accuracy and stimulus side, $F < 1$. This suggests that the valence of the feedback had no effect on how much attention was allocated to it. This analysis again focused on just one electrode pair (P7/P8), one time window (210 to 270 ms) and ignored the side on which the response key was pressed. Results for the full five-way repeated-measures ANOVA with objective accuracy, stimulus side, response side, time window and electrode pair as dependent variables can be found in Appendix A.5.

However, Figure 74 suggests that the pre-stimulus baseline chosen for

this analysis led to condition differences prior to the onset of the time windows in question, which makes interpretation of the results difficult, if not impossible. I therefore repeated the N2pc analyses with a baseline ranging from -100 to 0 ms pre-feedback-stimulus. The data for this analysis are presented in Figure 76.

The data were again submitted to a two-way ANOVA with classifier quartile and stimulus side as independent variables. According to the general N2pc effect, we can expect that the signal is more negative on the side contralateral to the stimulus presentation, which shows that participants directed their attention to this field. This effect was reliable, $F(1,\ 16) = 9.7,\ p < 0.01,\ \eta_p^2 = 0.38$, as in the previous analysis. It can therefore be concluded that allocation of attention to the feedback stimulus was reflected in the N2pc.

I next tested whether this effect was modulated by classifier bin. The interaction between stimulus side and classifier bin was indeed reliable, $F(3,\ 48) = 3.8,\ p = 0.02,\ \eta_p^2 = 0.19$. The precise hypothesis in this context was that the N2pc effect would be larger for the two middle bins, however, the quadratic trend was not reliable, $F(1, 16) = 2.6,\ p = 0.13,\ \eta_p^2 = 0.14$. There was also no reliable main effect of classifier bin, $F(3,\ 48) = 1.9,\ p = 0.15,\ \eta_p^2 = 0.11$.

Taken together, even with the improved baseline, these findings suggest that the overall N2pc effect was not modulated by classifier confidence. Moreover, the N2pc was also not modulated by the valence of the feedback, as shown in the lower panel of Figure 76. In a two-way repeated-measures ANOVA with objective accuracy and stimulus side as dependent variables. There was, once more, a reliable main effect of stimulus side, $F(1,\ 16) = 5.0,\ p = 0.04,\ \eta_p^2 = 0.24$, however, there was no reliable interaction effect, $F < 1$. There was also a marginally reliable main effect of accuracy,

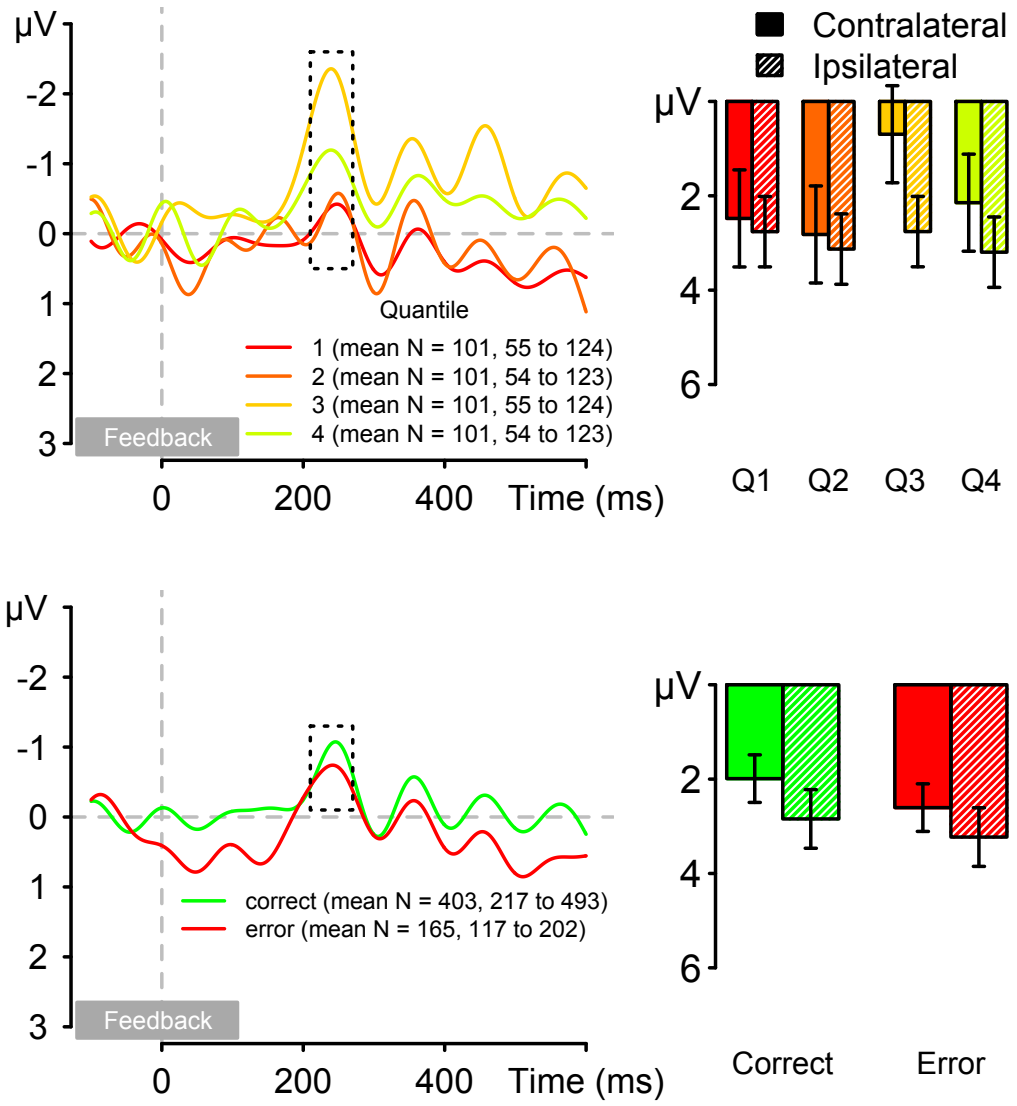

Figure 76: Difference wave (contralateral minus ipsilateral) for the feedback-locked N2pc at electrodes P7 and P8, conditioned on single-trial Pe amplitude, reaching from smallest (1) to largest (4) quantile. The data were baselined to -100 to 0 ms pre-feedback-stimulus. The window highlights the time during which this component was most pronounced, as identified in previous analyses (210 to 270 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

$F(1,\ 16) = 3.5,\ p = 0.08,\ \eta_p^2 = 0.18$, reflecting that mean voltages at P7/P8 were more negative for correct trials, compared to errors, $M_{cor} = 2.4\ \mu V$; $M_{err} = 2.9\ \mu V$.

Surprise effects (FRN and P3). As previously mentioned, I expected the FRN to reflect an unsigned reward prediction error – that is a surprise signal – in response to error feedback. This would mean surprise in cases where feedback deviated from the expected outcome, for instance when the participant expected an error but was correct. The design used here has the advantage that the classifier can predict the expectation participants hold at a given time – if they are low confident, they should be expecting negative feedback, as the feedback was fully contingent on their response. Given the surprise hypothesis, I would therefore expect to find the classifier bin to interact with objective accuracy.

Figure 77 presents the feedback-locked ERPs. The FRN peaks around 250 ms in these figures. The data were submitted to a three-way ANOVA with electrode, bin, and accuracy and independent variables. Only five midline electrodes were considered in this analysis: FZ, FCZ, CZ, CPZ, and PZ. FRN data for electrode CZ are shown in the upper and lower left panels of Figure 77, with the respective topographies shown in Figure 78.

I would have expected the FRN to be largest at fronto-central electrodes, the strongest effect was found at the most frontal electrode FZ, though, $M_{FZ} = 3.5\ \mu V$, compared to the other electrodes, $M_{FCZ} = 4.1\ \mu V$, $M_{CZ} = 4.0\ \mu V$, $M_{CPZ} = 3.7\ \mu V$, $M_{PZ} = 4.0\ \mu V$. However, this effect was not reliable, $F(12.0,\ 31.6) = 2.1,\ p = 0.14,\ \eta_p^2 = 0.12$. I therefore analyse all of these electrodes.

If the FRN reflected an unsigned RPE, we could expect to find a main

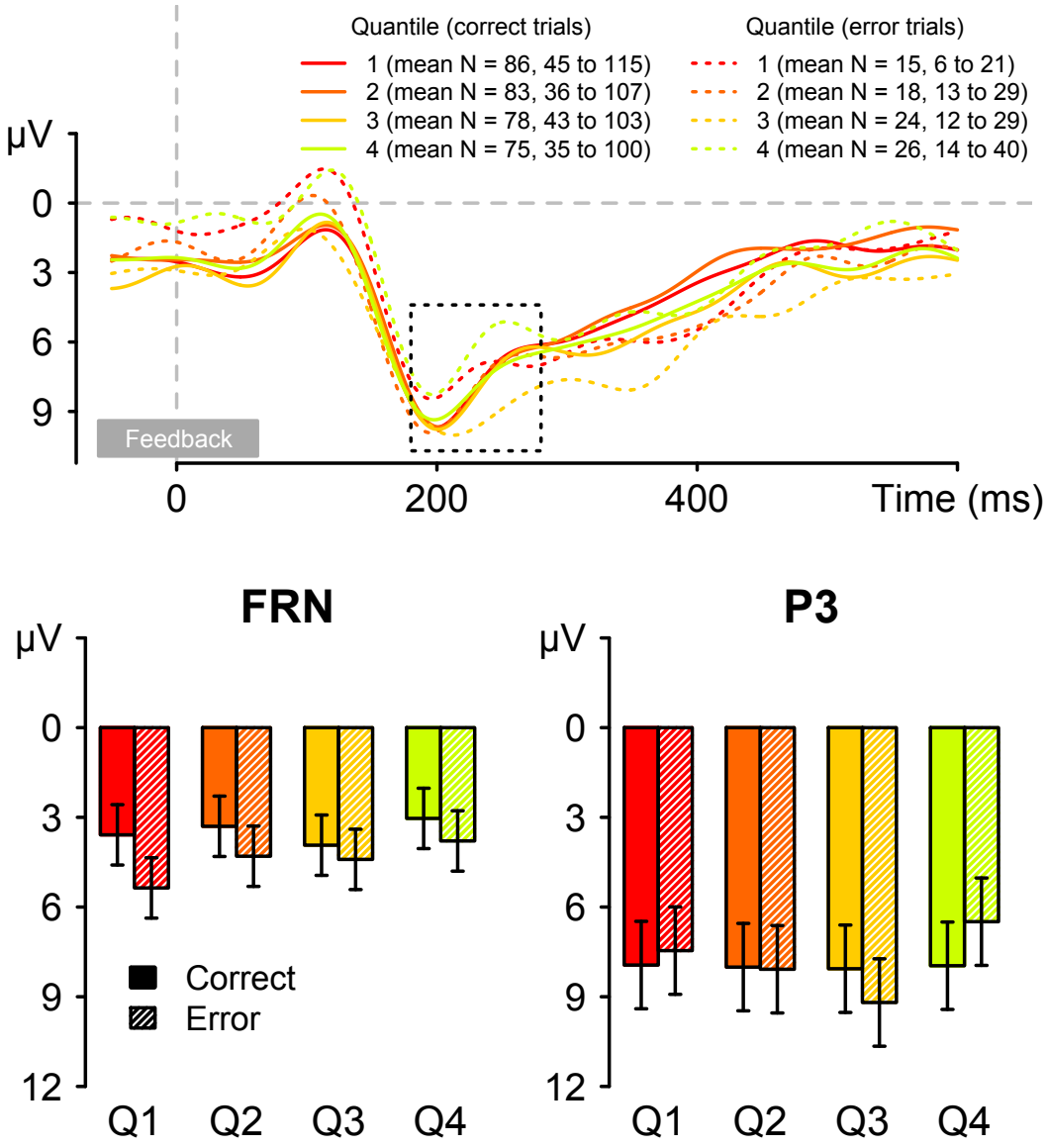

Figure 77: Feedback-related negativity (FRN; fERN) and P3 – both feedback-locked – at electrode CZ, conditioned on single-trial Pe amplitude, reaching from smallest (1) to largest (4) quantile. The data were baselined to -100 to 0 ms pre-stimulus, and therefore to a different time window for each trial depending on RT. The window highlights the time during which the P3 component was most pronounced, as identified in previous analyses (180 to 280 ms). The legend displays the average number of trials across participants, together with the minimum and maximum number of trials; ms: millisecond; μV: micro-volt.

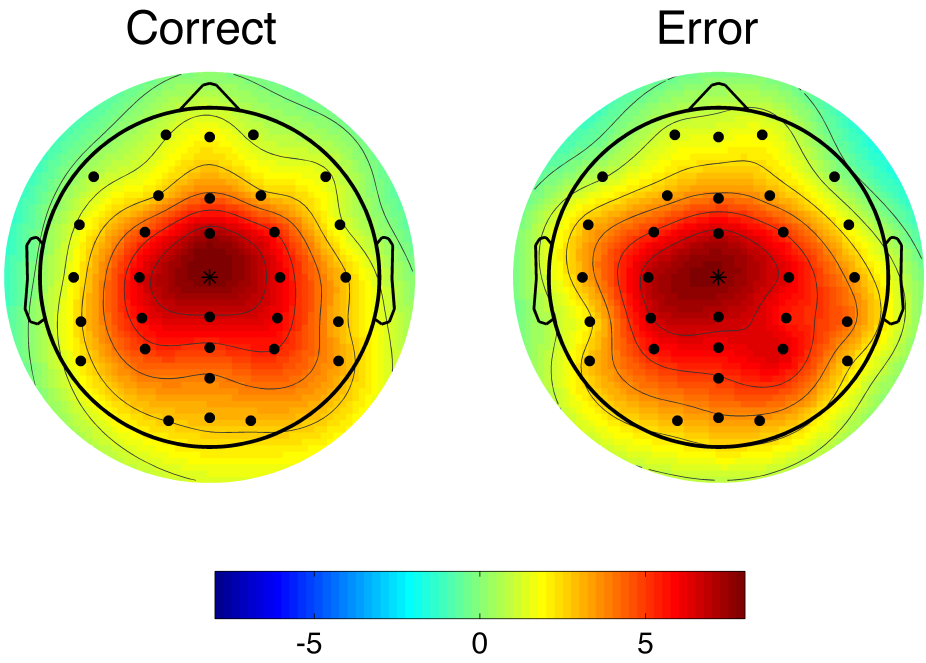

Figure 78: Topographies for the P3 during a time window from 180 to 280 ms. The left panel shows correct trials, the right panel shows error trials. The colours in the topographic plots indicate different values in micro-volt; ms: millisecond.

effect of objective accuracy depending on overall error rates. For instance, if errors are less common than correct trials – as they usually are – then participants would be expected to show signs of surprise at error feedback, as reflected in the FRN. This effect was indeed reliable, $F(1,\ 16) = 27.1,\ p < 0.001,\ \eta_p^2 = 0.63$, with a larger FRN on error, $M = 4.3\ \mu V$, compared to correct trials, $M = 3.4\ \mu V$. These findings therefore support the view that the FRN is a reflection of a signed prediction error, as opposed to a surprise signal. Moreover, according to the surprise hypothesis, I would have expected to find a reliable interaction between classifier bin and accuracy, which was also not found, $F < 1$. There was also no significant main effect of bin, $F(3,\ 48) = 1.2,\ p = 0.32,\ \eta_p^2 = 0.07$. There were no other reliable effects for this ANOVA, $Fs <= 1.6,\ ps >= 0.19$.

Taken together, the results did not follow the above-described surprise hypothesis. Participants did not show a reliably larger FRN when their expectations were violated, such as receiving negative feedback on a high-confidence trial, $M_{highsurprise} = 4.2\ \mu V$, compared to a trial on which feedback was congruent with their expectation, $M_{lowsurprise} = 3.7\ \mu V$. Instead, there was an effect of feedback valence, similar to the findings reported by Yeung and Sanfey (2004; see also Sato et al., 2005).

As argued above, P3 amplitude was expected to reflect surprise, similar to the FRN. The data were therefore analysed for correct and error trials separately, to correctly account for the effects of expectation. The average P3 as a function of bin and accuracy is shown in the lower right panel of Figure 77. Topographies for the average P3 on correct and error trials is presented in Figure 78.

Data from electrode CZ during a time window ranging from 180 to 280 ms had previously been identified to show the largest P3 effect (see Section

4.1.5.3). These data were therefore submitted to an ANOVA with bin and accuracy as independent variables. If the P3 did indeed reflect surprise, then I would have expected to find its amplitude to be largest for correct trials in bin 4 (the participant expected negative feedback but instead received positive feedback) and error trials in bin 1 (the participant expected positive feedback but instead received negative feedback). This would have been reflected in a reliable interaction between accuracy and bin. This effect was not reliable, however, $F(3,\ 48) = 1.4,\ p = 0.24,\ \eta_p^2 = 0.08$. It can therefore not be concluded that P3 amplitude followed surprise. The effect of accuracy was also not reliable, $F < 1$, echoing the findings reported by Yeung and Sanfey (2004). Moreover, the classifier quartile had no significant effect on P3 amplitude, $F(2.1,\ 33.5) = 1.3,\ p = 0.28,\ \eta_p^2 = 0.08$. There was also no reliable linear trend, $F < 1$. It can therefore not be concluded that P3 amplitude reflected a positive prediction error (Sallet et al., 2013).

4.2 General discussion

The present experiment was designed to test whether confidence has an effect on attention paid to feedback. The key hypothesis was that participants would pay more attention to feedback if they were unsure regarding the accuracy of their previous response. The general notion was that participants would be able to gain information from feedback in such situations, whereas feedback would not provide additional information – and could therefore be ignored – when they were certain that their previous response was correct or incorrect. Confidence therefore functions as an internal feedback signal. However, the findings for the ERP components reflecting anticipation (SPN), attention (N2pc), or surprise (FRN/P3) did not provide support this attentional hypo-

thesis.

With the present study, a novel, unobtrusive approach to measure confidence was used that meant that no overt confidence responses had to be collected. Error awareness was instead predicted using the classifier approach that had been introduced in Chapter 3. A similar dot-count decision task was used, which was adaptive in nature, aiming for a substantial amount (one third to a half) of undetected errors. This staircase was indeed successful. In the main experiment, participants received feedback after every trial. For the confidence blocks, resolution mirrored the results from the previous experiments. Moreover, echoing the findings from Experiment 3, participants were slightly faster when no confidence judgements were taken, but a difference in strategy was not reflected in accuracy.

The classifier, which was used to predict trial-by-trial variations in confidence, was once more based on the Pe. The ERN and the Pe not only reflected differences in objective accuracy, but also confidence – despite the low trial number of metacognitive trials – replicating the findings of Experiment 4. Moreover, this effect was also present in correct-trial Pe amplitude on the single-trial level, again replicating the findings from Experiment 4.

The key question in this experiment was then addressed by analysing four feedback-related EEG components, SPN, N2pc, FRN, and P3. For the former two, which were interpreted to reflect allocation of attention to feedback, no effects of classified confidence were found. Whether or not participants were guessing their response did not seem to have an effect on how much attention they allocated to feedback. Presumably, this null effect was not caused by general noise in the EEG data, as for example the N2pc did reflect the general effect of attention to feedback, as would have been expected for this component. For the latter two components, which were thought to reflect

expectation or surprise, there were also no effects of confidence. However, a midline frontal negativity was found that varied with accuracy, similar to the findings reported by Yeung and Sanfey (2004), therefore suggesting an effect of feedback valence, as reflected in the FRN.

One reason why the predicted effects of confidence were not found could be the reduced power of the classifier. As reported above, a k-fold algorithm revealed that the AUC for the classifier used here was only 0.64, which is relatively low. It can therefore be assumed that the predictive performance of the classifier was reduced and the four confidence bins identified were in fact overlapping to a great extent. If we compare the present experiment to the one in the previous chapter, one possible reason for this could be that in the present case, a substantial number of errors remained undetected. This becomes obvious if we compare Figures 30 and 62, which show the distributions of confidence responses as a function of accuracy for the two experiments. There was a larger overlap for the experiment in this chapter, suggesting that the primary task was slightly more difficult, which led to more data-limitation errors (Scheffers & Coles, 2000). The Pe has been repeatedly linked to error awareness (Nieuwenhuis et al., 2001; Steinhauser & Yeung, 2010). Given that the classifier was trained on data taken from the time window of the Pe, we could assume that it worked better for the type of error that was caused by premature responding. This could have led to a more stable classifier for Experiment 4.

Several changes regarding the design of the study could contribute positively to the power of the paradigm. First of all, feedback stimuli could be made more salient, for instance by introducing a reward scheme according to which participants would earn points or money for correct first-order responses. Currently, the feedback served no purpose other than a confirmatory one, that

is participants monitored it merely to compare it to their own error detection. Using a reward paradigm would also have an effect on the strength of the FRN component. This EEG correlate is usually studied with rewards paradigms with feedback signalling whether the last response was a *loss* or a *win*. Using a reward paradigm could also make the SPN stronger, as shown in a study by Kotani et al. (2003).

Another way to make feedback even more salient is to require participants to respond to it. Indeed, it has been found that the FRN is enhanced when a reaction to feedback is required rather than when feedback is passively viewed (Yeung, Holroyd & Cohen, 2005; see also Walsh & Anderson, 2012, for a review). This would be inherently achieved with a learning experiment. For instance, one could design a study similar to the paradigm used by Yu and Dayan (2005) in which participants receive probabilistic feedback from which they have to infer whether the state of the world has changed, and whether they therefore have to change their behaviour accordingly. In such a task, attending to feedback is crucial to notice switches in the environment and to adapt accordingly.

In addition to studying EEG correlates of attention, one could also consider studying the time spent to study feedback, similar to Kulhavy and Stock (1989): After every trial, informative feedback could be given. Participants would be allowed to abort this feedback presentation by pressing a button, and they could do so immediately after they made their response. This design would permit measurement of whether or not participants skipped feedback and in the latter case how long they allow feedback to stay on screen. This could be interpreted as an indirect measure of attention to feedback. This study would be a replication of the design used by Hays et al. (2010) in a perceptual-decision-making context.

Taken together, this study failed to find evidence for modulation of feedback processing by confidence. Instead, these findings suggest that instead of the certainty hypothesis described above, feedback processing was more modulated by feedback valence. Several improvements to increase the power of the design have been suggested. The general problem that remains, however, is that the readout of confidence was not as stable as expected.

Chapter 5

Multiple cues contribute to the formation of metacognitive judgements

In the previous chapters, I have focused on how metacognitive judgements can be measured and how they affect cognitive processing and behaviour. However, precisely which mechanisms lead to the formation of such confidence judgements and how they are linked to actual performance has not yet been touched upon and remains a challenge for most models of confidence. Confidence is often interpreted as a subjectively judged probability of being correct (Moreno-Bote, 2010; see also Fleming & Lau, 2014, for a review), and it has been argued that this representation is task-independent (De Gardelle & Mamassian, 2014). However, research has shown that confidence is by no means a direct, fully reliable index of objective accuracy. Leonesio and Nelson (1990), for instance, have compared different judgements of metamemory. Their findings suggest that correlations between various ratings and actual task performance are moderate, at best. Other studies from the domain of metamemory (Busey et al., 2000), perception (Fleming, Huijgen & Dolan, 2012; Fleming et al., 2010; Kiani et al., 2014), and problem solving (Metcalfe, 1986;

Metcalfe & Wiebe, 1987) have arrived at similar conclusions.

Nevertheless, confidence relates in systematic ways to key factors influencing decision processes. For instance, findings from perceptual decision-making studies suggest that manipulations of difficulty level, such as stimulus intensity, influence how confident participants judge their choices. Evidence suggests that this reduction in confidence with difficulty reflects a change in both metacognitive sensitivity and metacognitive bias. First, participants tend to be more confident overall in easy compared to difficult conditions (Baranski & Petrusic, 1998, p. 936). This follows quite intuitively if we assume a balance-of-evidence mechanism as previously described: Easy conditions are characterised by a larger difference in accumulation rates between the two counters, which means that the correct counter will reach the threshold first on most trials, leading to a high first-order detection sensitivity, d'. The counter representing the incorrect response, on the other hand, will on most trials lose as it will have accumulated only a limited amount of evidence in this time, leading to a larger balance of evidence and therefore higher confidence. On a small proportion of trials, random fluctuations will have led to this counter reaching the threshold first and eliciting an erroneous response. However, the counter representing the correct choice alternative will most likely have accumulated a substantial amount of evidence, leading to smaller balance of evidence and therefore lower confidence. This is an example of a well-known effect, namely the idea that second-order sensitivity, *meta-*d' (Maniscalco & Lau, 2012), covaries with first-order sensitivity – in other words, if a participant was guessing the answer to a difficult perceptual decision task, their confidence judgement following this decision would also be likely to reflect guessing.

Second, difficulty itself can have an effect on the other SDT measure, metacognitive bias; that is participants' tendency to use one or the other end

of the confidence scale. A high bias means that participants tend to classify responses as correct independent of their actual accuracy; they are overconfident. A low bias, on the other hand, means that participants are underconfident, they are more likely to classify responses as incorrect. Findings from numerous studies (Baranski & Petrusic, 1994; Gigerenzer et al., 1991; Juslin et al., 2000; Pleskac & Busemeyer, 2010) suggest that easier conditions are often accompanied by underconfidence whereas more difficult conditions are accompanied by overconfidence, the hard-easy effect. A similar effect has been found for individual differences given that participants' competence in a domain is correlated with how well they are at judging their own competence, in other words, overconfidence increased with incompetence (the "unskilled-and-unaware-of-it" phenomenon; Kruger & Dunning, 1999).

As mentioned briefly in Chapter 1, this work has led to contrasting theories regarding the precise mechanisms of confidence judgements (Koriat, 2012; Schwartz & Metcalfe, 1994), with direct-access and heuristics-based models. Those theories will briefly be reviewed and contrasted here.

First, direct-access theories assume that confidence judgements rely on the same information as the decision itself, or some property of that decision. One example is the theory of probabilistic mental models (PMM) by Gigerenzer et al. (1991). According to this theory, both the decision and the confidence judgement are based on probabilistic cues taken from the environment. Validity of the currently-activated cues then determines the accuracy of both the decision and the confidence judgement. Another popular direct-access model of confidence is type-II SDT (Higham et al., 2009; Maniscalco & Lau, 2012). According to this theory, confidence reflects the quantity of evidence accumulated in favour of the chosen response option.

The term direct-access also refers to views in which other properties

of the decision contributed towards decision confidence. One example is the balance-of-evidence hypothesis, in which the difference in evidence accumulated in each of two racing counters at the time of the decision is thought to provide the level of confidence that accompanies the decision (Kepecs & Mainen, 2012; Van Zandt & Maldonado-Molina, 2004; Vickers & Packer, 1982). This readout does not necessarily have to be noise-free, as highlighted by De Martino et al. (2013). Evidence supporting the unchosen option has also been assumed to play a role in mismatch models, which assume that error detection is based on internally comparing the intended action to the one that was actually performed (Charles, King & Dehaene, 2014; Middlebrooks & Sommer, 2011; Falkenstein et al., 1991; Gehring et al., 1993), as well as Audley's "runs model" (Audley, 1960) and the doubt-scaling model proposed by Baranski and colleagues (Baranski & Petrusic, 1998). These examples are also closely related to models that focus on the role of cognitive conflict in the generation of metacognitive signals (Davelaar, 2009; Yeung, 2013), based on the idea that competing response tendencies can serve as a cue to confidence and error detection.

There are also direct-access models which assign a special role to the quality (or reliability) of evidence. In an accumulation model, such as the ones previously described, evidence quality would be reflected in the drift rate, that is the speed with which the system samples evidence for either of the two responses over time. It has previously been suggested that confidence is derived from a combination of evidence quality and quantity (Peirce's Model; Peirce, 1877; Pleskac & Busemeyer, 2010; see also Yeung & Summerfield, 2014, 2012; Kiani & Shadlen, 2009). The idea here is that evidence quantity alone is not sufficient to determine confidence since the amount of evidence needed to elicit a decision may be fixed within a condition and would therefore result

in the same level of confidence for all trials within this condition. Evidence quality, however, is not directly accessible for the participant (Hanks et al., 2011), otherwise participants would not have to sample evidence to begin with (Yeung & Summerfield, 2012). Instead, the participant could have access to a combination of both parameters, which then determine decision confidence.

In contrast to these various species of direct-access hypothesis, there are models that hypothesise confidence does not depend primarily on direct access to parameters of the decision process, but rather reflects other influences or cues that are external to the first-order decision process. A substantial corpus of research has been conducted on this question in the literature on metamemory. In this context, results suggest that confidence judgements are rarely based directly on the strength of the memory trace of the currently-judged material, but are instead often influenced by heuristics, such as the familiarity (Hertzog, Dunlosky & Sinclair, 2010), fluency (Castel, McCabe & Roediger, 2007), or accessibility (Koriat, 1993) of to-be-retrieved material (see also Schwartz & Metcalfe, 1994). Another heuristics-based model is the self-consistency model (SCM, Koriat, 2012, 2011), according to which participants repeatedly sample the problem or item at hand and judge their confidence depending on how much their chosen alternative differs from those samples (*consensuality principle*). Yet another example of an heuristics-based model in the decision-making domain is the time heuristic. According to this view, RTs are a frugal cue for confidence judgements – the longer it takes us to form a decision, the less certain we should be (Audley, 1960; Zylberberg et al., 2012; Kiani et al., 2014). In fact, it has been suggested that RTs represent a proxy by which participants estimate the reliability of the evidence underlying a decision (Hanks et al., 2011; Kiani & Shadlen, 2009). The notion here is that participants learn over time that information from unreliable sources –

such as reading a road sign without wearing one's glasses – is associated with slower responses, in this case reading the sign takes longer.

Evidence reliability is a factor that has recently been studied intensely in the field of decision making (Beck et al., 2008; Fiser, Berkes, Orbán & Lengyel, 2010; Ma, Beck, Latham & Pouget, 2006). In these studies, evidence reliability is usually operationalised as variability in the stimulus, over time or over a decision-relevant dimension. For example, De Gardelle and Summerfield (2011) presented arrays of coloured shapes, which participants had to classify as on average red or blue. The variability of colour in those shape arrays was critically manipulated (comparing homogenous versus heterogeneous colour arrays) together with the distance of the average colour relative to the decision boundary (i.e., how 'purplish' the shades of blue or red were). On a behavioural level, more variable stimuli led to both increased RTs and error rates (De Gardelle & Summerfield, 2011; Michael, De Gardelle & Summerfield, 2014). In addition to these behavioural effects, other studies have suggested that evidence reliability is represented internally in form of a summary statistic (Alvarez, 2011; Pollard, 1984). These summary statistics can occur at all stages of cognitive processing and can be both internal (interference of other mental representation makes a to-be-retrieved memory more fuzzy) and external (foggy weather conditions hinder clear perception of a road sign; Bach & Dolan, 2012). Representing reliability in such a way makes sense when considered from a Bayesian perspective: If agents were assumed to behave Bayes-optimally they should weigh evidence by its reliability. In line with this idea, studies have suggested that participants adjust the influence of different sources of information according to their reliability (Montgomery & Sorkin, 1996; De Gardelle & Summerfield, 2011). Such information regarding the uncertainty of a neural representation has also been shown to optimise

learning (Behrens et al., 2007). Moreover, a recent study by Michael, De Gardelle, Nevado-Holgado and Summerfield (2015) suggested that uncertainty due to evidence reliability is represented in the dorsomedial prefrontal cortex (dmPFC) separately from uncertainty due to low evidence strength.

The impact of evidence reliability on confidence is not currently understood. Intuitively, one would expect reduced evidence reliability to result in reduced confidence, which has been suggested in recent reviews by Yeung and Summerfield (2012, 2014). A study conducted more than half a century ago by Irwin et al. (1956) found support for this hypothesis. The authors tested influences on confidence in a decision task where mean and variability of a primary task stimulus were varied. For this purpose, participants were shown 20 cards with numbers printed on them out of a deck of 500. They had to judge whether the mean of all these number cards was above or below zero. In addition, they had to judge how confident they were with regard to their response. Interestingly, confidence was inversely related to the standard deviation of the numbers on the cards. However, a recent study by Zylberberg et al. (2014) found precisely the opposite effect in a study of perceptual decision making. In their task, participants had to judge whether an array of line segments was on average oriented clockwise or counterclockwise, rating their confidence together with each response. The authors varied the jitter of these line-segments and found, perhaps surprisingly, that greater evidence variability led to increased confidence. In the present chapter, I follow up on the question of how reliability of evidence influences confidence, asking how these ambiguous findings can be explained and potentially reconciled.

The first experiment in the present chapter, Experiment 6, focused on whether a combination of cues determined how confident participants judge their responses and whether evidence reliability is one of these cues. Parti-

cipants had to judge the average colour of an array of eight coloured shapes, and reliability was operationalised as variability in the information in this multi-element array and critically manipulated as such. This operationalisation of evidence reliability is similar to the one chosen by De Gardelle and Summerfield (2011) and Zylberberg et al. (2014), with the underlying assumption that stimulus variability leads to more noise in the perceptual system and therefore less reliable mental representations of the stimulus values. The experiment compared the effect of evidence strength and evidence reliability on confidence, expecting to find that the effect of evidence reliability on confidence would be stronger because evidence reliability reflects the 'native language' of confidence. In other words, confidence could neurally be represented as the spread of a distribution, in line with a Bayesian perspective on cognitive processing. This spread, or reliability can then be accessed directly, translating it into verbal judgements of decision confidence.

Experiment 7 investigated individual differences in confidence and sensitivity to evidence reliability, specifically using an acute tryptophan depletion (ATD) model of depression. Disturbances in metacognitive processing have previously been reported in a number of clinical groups: Several studies have suggested impaired error detection in patients with schizophrenia (e.g., Alain, McNeely, He, Christensen & West, 2002; Malenka, Angel, Hampton & Berger, 1982). Charles (2013), for instance, found evidence for impaired conscious metacognition in schizophrenia, while unconscious metacognitive processing remained intact. Moreover, obsessive compulsive disorder (OCD) has previously been linked to impaired metamemory (Tolin et al., 2001; Ben Shachar, Lazarov, Goldsmith, Moran & Dar, 2013), resulting, for example, in reduced confidence in the memory trace of switching the gas off. A third example is patients with depression, who have repeatedly been reported to

differ from healthy controls in their metacognitive bias, being either under- (Dunlosky & Metcalfe, 2009, for a review) or overconfident (Dunning & Story, 1991). With Experiment 7, I therefore tested the hypothesis that those changes in metacognitive bias can be attributed to a weaker influence of evidence reliability on confidence, given the often reported increase in focused attention for participants with depression (Ahveninen et al., 2002; Schmitt et al., 2000). Acute tryptophan depletion was used in this study to lower extracellular serotonin levels. There is evidence that serotonin modulates processes related to mood and emotion (Heninger, Charney & Sternberg, 1984). It has furthermore been found in healthy participants that lowering serotonin through ATD leads to lower mood (Young, Smith, Pihl & Ervin, 1985).

Finally, the third experiment in this chapter focused on the question of whether the influence of evidence reliability on confidence was reflected in the same error-related neurophysiological correlates studied previously in this thesis – the ERN and the Pe. EEG activity was therefore recorded while participants performed the colour-judgement task and – following up on the notion that the amplitude of the Pe can be used to predict confidence (Experiments 4 and 5) – I tested whether Pe amplitude reflects yet another internal cue on which confidence judgements are based. It should be highlighted that this hypothesis does not reflect the idea that participants read out their confidence from the Pe, but rather that the Pe reflects influences of other cues not taken into account in the model of confidence.

5.1 Experiment 6: Signal reliability affects meta-cognitive judgements

Using colour and shape judgement tasks, De Gardelle and Summerfield (2011) showed that stimulus variance can affect decision making performance independently of signal strength, see Figure 79. The mean and variance of a signal therefore represent two orthogonal ways to manipulate difficulty, reflecting operationalisations of signal strength and signal reliability, respectively. Using their colour judgement task, I built on these findings to test how stimulus variance affects decision confidence. Colours were sampled from distributions with pre-specified weak versus strong evidence favouring one of the two colours, and low versus high evidence variability. These two experimental factors resulted in four difficulty conditions, of which the two intermediate ones were matched for performance. These two conditions should both have been of intermediate difficulty, but due to different stimulus characteristics: The *low mean, low variance* condition was difficult because the colour information did not clearly favour one option over the other. The *high mean, high variance* condition was difficult because the evidence was noisy. An adaptive procedure was used to match those medium conditions with regards to RTs and error rates. Of interest was whether and how these conditions might differ in their confidence.

Current theories of decision confidence would predict that equally difficult first order tasks will lead to equivalent levels of confidence, given that most of these direct-access models assume that first- and second-order judgements rely on the same information and internal processes. The inferential time heuristic similarly predicts that confidence and first-order performance, specifically RT, are inseparably intertwined. I predicted, however, that confidence should

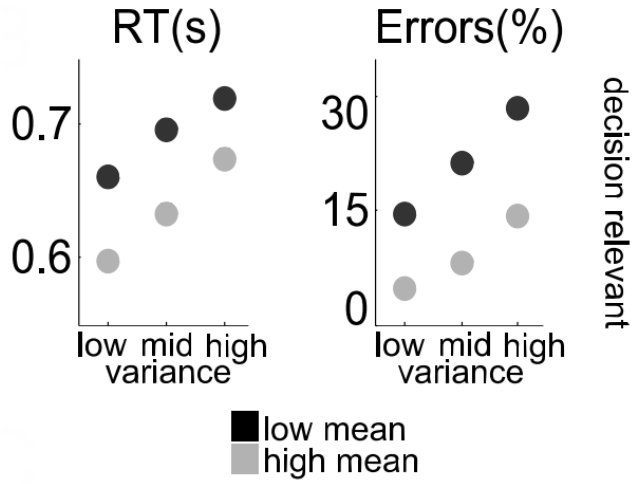

Figure 79: Figure reproduced from De Gardelle and Summerfield (2011), presenting response times (RTs) and error percentages for six difficulty conditions.

be affected by both manipulations of difficulty, but more so by evidence reliability, or colour variance, following from the idea that uncertainty estimates are represented internally in the form of probability distributions (Beck et al., 2008; Fiser et al., 2010; Ma et al., 2006). This hypothesis was tested by directly comparing the two performance-matched conditions, by analysing SDT parameters (metacognitive sensitivity/efficiency and metacognitive bias), and by fitting regression models to predict participants' confidence ratings.

5.1.1 Methods

5.1.1.1 Participants

Twenty-one participants were recruited from a participant database. One participant had to be excluded due to apparently random use of the confidence scale. There were 20 participants in the final sample, 14 of whom were female, with ages ranging from 18 to 25 years. All participants had normal or corrected-to-normal vision and – according to self-report – intact colour vis-

ion. The experiment lasted approximately 90 minutes. Participants received course credit ($N = 6$) or money (£12; $N = 15$) as compensation. All testing was approved by the local ethics committee.

5.1.1.2 Task and procedure

The participants' task was to judge the average colour of eight shapes, determining whether this colour was on average more red or more blue. Each stimulus consisted of eight coloured shapes spaced regularly around a fixation point (radius 2.8° visual arc). This task can be made difficult in two distinct ways: first, by reducing the mean of the distribution (i.e., using colours that are, on average, purple hues rather than clear red or blue) and, second, by increasing variability in the distribution of colours (i.e., using colours that are a heterogenous mix of reds, blues and purples rather than a homogeneous hue). The latter factor was my experimental manipulation of evidence reliability. Factorial crossing of these two experimental factors results in four conditions of varying difficulty (Figure 80).

The task is easy when stimulus mean is high (on average the colour is very red or very blue) and stimulus variability is low (all stimulus elements exhibit this difference). Conversely, the task is challenging when stimulus mean is low (the average colour is 'purplish red' or 'purplish blue') and variability is high. Then there are two conditions which should both be of intermediate difficulty, but due to different stimulus characteristics: The *low mean, low variance* condition is challenging because the average colour is 'purplish red' or 'purplish blue' and the colour information is therefore not clearly favouring one option over the other. The *high mean, high variance* condition, on the other hand, is difficult because the evidence is noisy. Of critical interest is the comparison between these two medium difficulty conditions, which I matched

in terms of primary task performance using a staircase procedure described below.

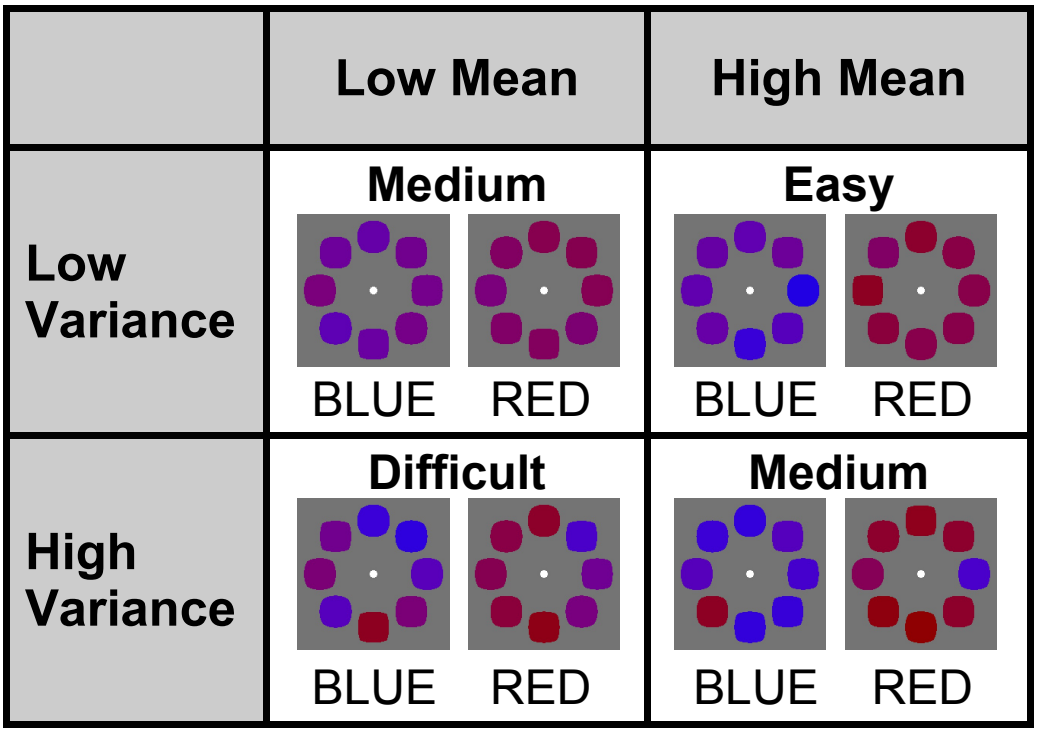

Figure 80: Sample stimuli, showing the four difficulty conditions in the 2 (stimulus mean) x 2 (stimulus variance) design. For each cell, both a red and a blue stimulus are shown. Colour values were made more extreme for illustrative purposes.

A second, decision-irrelevant dimension was furthermore included in the design: Stimuli varied in their shape, being round or rectangular or an intermediate shape (a geometric shape usually referred to as "squircle"), but those variations were not expected to affect task performance. Shape values were varied similar to colour but this stimulus dimension was not relevant for the decision task and therefore not mentioned in the instruction. There was also no staircase to adjust the difficulty of the shape values. Replicating the findings in De Gardelle and Summerfield (2011), mean and variance of shape had no influence on either correct RTs nor error rates, $Fs <= 1.6$, $ps >= 0.23$,

and are therefore not discussed further below.

Stimuli were presented on a 20" CRT monitor with a 75 Hz refresh rate using the MATLAB toolbox Psychtoolbox3 with a 70 cm viewing distance. All responses were made with a USB keyboard. The colour judgements were made with the "c" or "n" key (left or right thumbs). Confidence responses were made with the upper number line (keys "1", "2", "3", "8", "9", and "0") using the index, middle and ring fingers of the two hands.

A typical sequence of trial events is shown in Figure 81: Participants saw the stimulus for 160 ms. They then pressed a key according to whether they thought the average colour of the stimulus was red or blue. They had up to 1,500 ms time to give their response and trials exceeding this time were counted as misses and a warning message would ask them to respond faster. After a 600 ms RSI, a confidence scale was presented and the participants indicated how confident they were about the correctness of their response by pressing one out of six keys. The confidence scale was the same as in previous experiments. Participants were given unlimited time for their judgements. The next stimulus was presented 1000 ms later.

Participants completed extensive training in the task, both with and without confidence judgements (512 to 704 trials), during which an adaptive procedure was used to match the medium conditions with regard to RTs and error rates. The level of stimulus mean in the *low mean, low variance* condition was varied so that performance in this condition was matched with performance in the *high mean, high variance* condition. Matching was done specifically with reference to an efficiency measure, median correct RT divided by accuracy (inverse efficiency score *IES*; Bruyer & Brysbaert, 2011). More precisely, at the beginning of the practice staircase blocks (blocks 2 to 8), the weak evidence condition was adjusted whenever the two medium con-

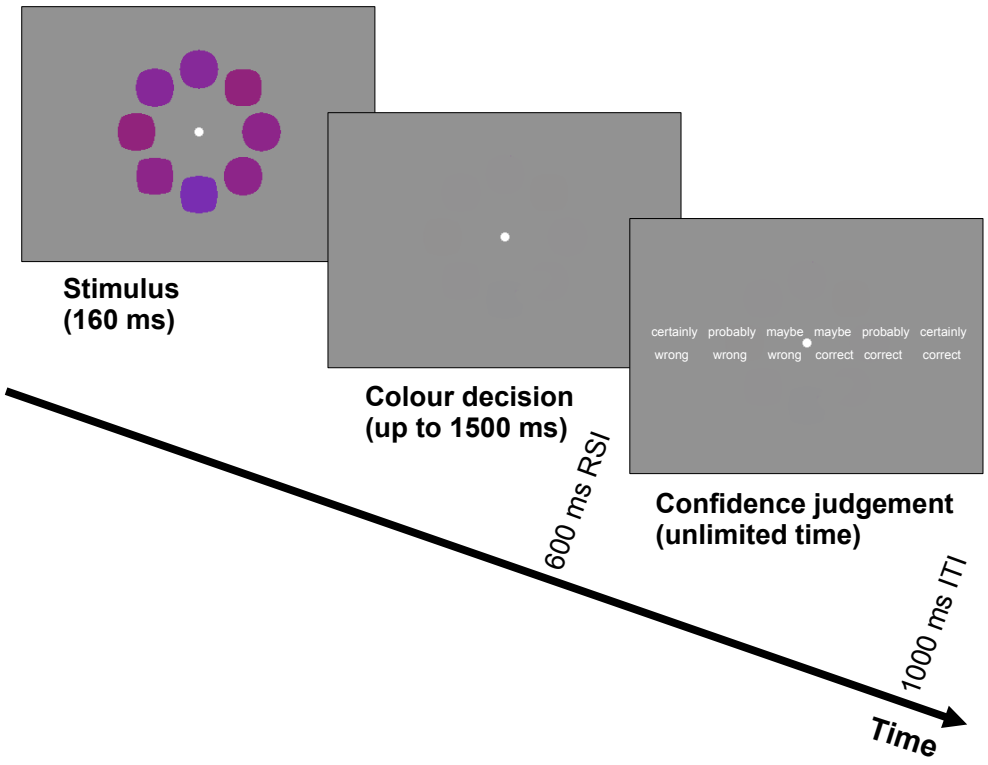

Figure 81: Methods of the colour task. Participants first had to indicate whether an array of eight coloured shapes was on average more red or more blue by pressing the left or right response key. After their response, the confidence scale was presented on screen and participants were given unlimited time to choose how confident they were that their last response was correct. RSI: response-stimulus interval; ITI: inter-trial interval; ms: millisecond.

ditions were not matched with regard to the efficiency measure. Matching was assessed across the preceding block. If there was a difference of at least 10 ms/acc, the weak evidence condition would be increased or decreased by 2% of the initial difference between weak and strong evidence. If the difference was at least 50 ms/acc, the change would be 5%, and for 100 ms/acc or more, the change would be 10%. After these adjustment blocks, the experimenter decided whether additional adjustment blocks were needed, based on visual inspection of RTs and error rates for the two medium conditions, which were shown on screen. For example, if performance in the two conditions was converging but not yet quite matched, the experimenter would decide to include yet another practice block for this particular participant.

Participants completed seven to ten adjustment blocks in total ($M = 7.7$), during which a feedback tone was played every time they committed an error. They then completed one block of task practice in which the confidence scale was presented for the first time. After completion of this block, frequencies for each confidence category were displayed on screen and the experimenter discussed these values with the participants, encouraging them to use the full scale.

Participants then completed 16 experimental blocks of 64 trials each. Prior to each block, they completed 16 colour judgement trials without confidence ratings and instead with auditory feedback to help the participants to maintain a stable colour discrimination criterion throughout the experiment. Feedback was not given during the main part of the blocks in which a confidence rating was required on each trial, however, median correct RTs and error rates for the two colour classes were shown on screen after the completion of each block. Participants were instructed to pay attention to these feedback values and to monitor themselves in that they would not develop a bias towards

one of the colours. The experimenter also monitored these values throughout the entire study, reminding participants of the instruction whenever necessary. Direction of the confidence scale and mapping of colours to response keys were counterbalanced across participants.

5.1.2 Results

5.1.2.1 First-order judgements

Basic performance measures and matching of the medium conditions. The first set of analyses focuses on whether the present experiment replicated findings from De Gardelle and Summerfield (2011) with regard to correct RTs and error rates. The matching of the two conditions of medium difficulty – *high mean, high variance* and *low mean, low variance* will be tested thereafter, both with regard to an efficiency measure and parameters from a diffusion model. The contribution of the different coloured elements to the final decision will then be discussed.

Stimulus mean had a significant effect on both correct RTs, $F(1,\ 19) = 79.3,\ p < 0.001,\ \eta_p^2 = 0.81$, and error rates, $F(1,\ 19) = 203.7,\ p < 0.001,\ \eta_p^2 = 0.92$, with a higher mean leading to faster RTs and lower error rates. Stimulus variance also had a reliable effect on correct RTs, $F(1,\ 19) = 72.5,\ p < 0.001,\ \eta_p^2 = 0.79$, and error rates, $F(1,\ 19) = 92.8,\ p < 0.001,\ \eta_p^2 = 0.83$, with lower stimulus variance leading to faster RTs and lower error rates. These results replicate the findings reported by De Gardelle and Summerfield (2011). There was no interaction between the two factor for correct RTs, $F < 1$, but there was for error rates, $F(1,\ 19) = 11.9,\ p < 0.01,\ \eta_p^2 = 0.37$. These effects are presented in the left (correct RTs) and right (error rates) panel of Figure 82.

Furthermore, matching of the medium conditions was successful, using a staircase based on efficiency, with 722 ms for the *low mean, low variance* condition and 735 ms for the *high mean, high variance* condition, $t < 1$. This interpretation was supported by a Bayes factor of $BF_{NULL} = 3.18$. However, a slight speed-accuracy tradeoff was observed, that is the *low mean, low variance* condition triggered 31 ms faster correct RTs, $t(19) = 3.9,\ p = 0.001$, whereas the *high mean, high variance* condition had 3.0% lower error rates, $t(19) = 2.2,\ p = 0.04$. This apparent speed-accuracy trade-off will be further analysed below.

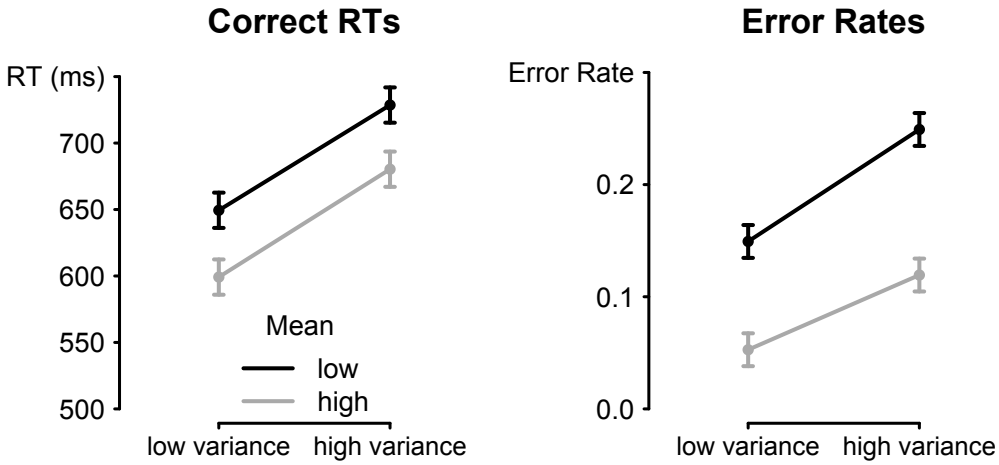

Figure 82: Mean response times (RTs) for the four difficulty conditions; correct trials only. Mean error rates for the four difficulty conditions; ms: millisecond.

Another measure that combines response speed and accuracy was taken into account to establish further evidence that the medium conditions were matched for performance. Such a measure is the drift rate, v, in a diffusion model. There was a reliable main effect of both stimulus mean, $F(1,\ 19) = 96.4,\ p < 0.001,\ \eta_p^2 = 0.84$, and stimulus variance, $F(1,\ 19) = 80.1,\ p < 0.001,\ \eta_p^2 = 0.81$, on drift rate. These factors also showed an interaction, $F(1,\ 19) = 11.6,\ p < 0.01,\ \eta_p^2 = 0.38$. Importantly, drift rate did not differ

between the two medium conditions, $M_{highhigh} = -0.32$, $M_{lowlow} = -0.33$, $t < 1$, the performance, that is the rate of evidence accumulation was therefore matched in these conditions. The hypothesis that these two conditions are equal was in fact supported by a Bayes Factor of $BF = 2.88$. We can therefore assume that the differences in RTs and error rates reported above was caused by a difference in strategy, with participants being slightly more cautious in the *high mean, high variance* condition as compared to the *low mean, low variance* condition. This was indeed reflected in a significant difference in boundary separation, a, between these two conditions, $M_{highhigh} = 0.15$, $M_{lowlow} = 0.14$, $t(19) = 3.9,\ p < 0.001$. Moreover, there was a reliable effect of stimulus mean, $F(1,\ 19) = 8.5,\ p < 0.01,\ \eta_p^2 = 0.31$, and stimulus variance, $F(1,\ 19) = 7.4,\ p = 0.01,\ \eta_p^2 = 0.28$, on boundary separation, and again an interaction between the two factors, $F(1,\ 19) = 8.8,\ p < 0.01,\ \eta_p^2 = 0.32$.

Taken together, the here reported effects suggest that evidence strength (mean colour) and evidence reliability (colour variability) have an effect on both correct RTs and error rates, thereby replicating the findings reported by De Gardelle and Summerfield (2011). Moreover, the medium conditions were found to be matched for task difficulty, albeit with apparent differences in speed-accuracy trade-off.

Contribution of each element to the colour decision. The next analysis focused on the impact each element had on the final colour decision. De Gardelle and Summerfield (2011) have previously shown that participants tend to ignore the most extremely coloured parts of the stimulus array (i.e., the most clearly red and clearly blue elements) when variability is high. Such down-weighting of outliers could be considered a good strategy as it artificially reduces the variability of the colours in the stimulus array. I therefore test

whether such an effect was replicated in my dataset, and later I test whether the extent to which participants chose such a down-weighting strategy correlated with the effect variability had on confidence. The eight coloured elements of the stimulus were therefore sorted according to their colour value, with clear red and blue elements at the extremes. A logistic regression was then fitted to estimate the decision weight for each element, which were then averaged for both outlying (positions 1, 2, 7, and 8; clearly red or blue stimuli) and inlying elements (positions 3 to 6). The weights of the coloured elements were normalised by dividing them by their root mean square (RMS), to minimise the influence unreliable estimates for individual participants.

These normalised regression weights are presented in Figure 83. De Gardelle and Summerfield (2011) previously found that these "weighting functions" revealed that inlying elements contributed more towards the final decision for the *high variance* conditions. This was also the case in the present experiment: Stimulus variance and position interacted reliably, $F(1,\ 19) = 12.1,\ p < 0.01,\ \eta_p^2 = 0.39$, whereas there was no such interaction for stimulus mean and position, $F < 1$. There was also a reliable main effect of stimulus mean, $F(1,\ 19) = 17.9,\ p < 0.001,\ \eta_p^2 = 0.48$, with relatively more influence of each element in a *low mean* compared to a *high mean* condition. The main effect of stimulus variance was also significant, $F(1,\ 19) = 119.2,\ p < 0.001,\ \eta_p^2 = 0.86$, caused by a high relative influence of elements in the *high variance* conditions. Finally, position also showed a reliable main effect with more influence of inlying, compared to outlying elements, $F(1,\ 19) = 8.5,\ p < 0.01,\ \eta_p^2 = 0.31$. Both the main effects of stimulus variance and position were presumably driven by the above mentioned interaction. There was neither a reliable interaction of stimulus mean and variance, $F(1,\ 19) = 1.9,\ p = 0.18,\ \eta_p^2 = 0.09$, nor a reliable three-way interaction,

$F(1,\ 19) = 2.7,\ p = 0.12,\ \eta_p^2 = 0.13$. To briefly summarise these findings – when variance was low, participants took all elements into account roughly equally when making their decision. When variance was high, however, they were found to down-weight the elements of the most extreme colours (very red and very blue elements), therefore taking the inlying elements (more purple elements) into account more than the outlying ones (more extreme elements). In Section 5.1.2.2, these effects will furthermore be analysed, asking how confidence relates to processing selectivity in the primary task.

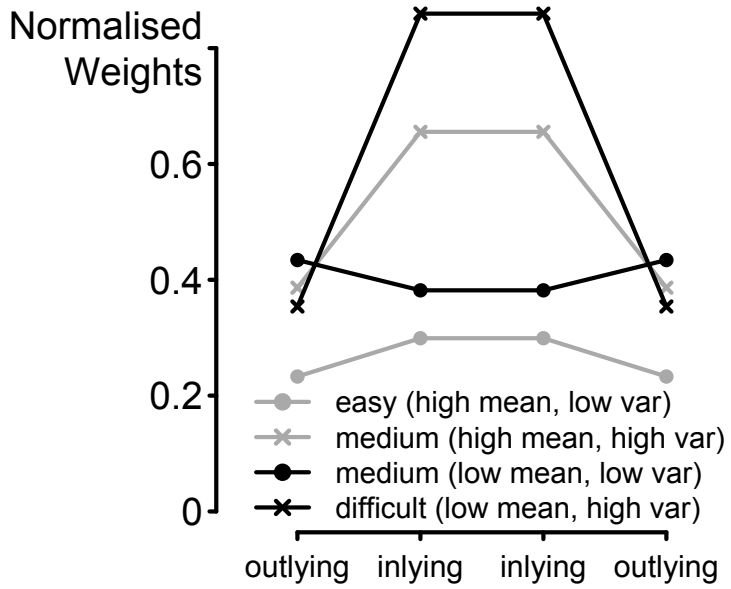

Figure 83: Contribution of coloured shapes to a final decision, sorted by their colour value and aggregated over inlying and outlying position; regression weights were divided by their root mean square (RMS). The values are mirrored so that the figure can be interpreted more intuitively. No error bars are shown because they are somewhat overlapping and hinder interpretation of the figure; var: variance.

5.1.2.2 Confidence judgements

Confidence resolution. Given that stimulus mean and variance affected decisions as predicted, it was of crucial interest how stimulus mean and variance affected confidence responses. This analysis therefore aims to test the overall stability of people's confidence judgements in the colour-judgement task. Given this, I can then look at how mean and variance affected confidence. There was a monotonic decrease in error rates with level of confidence, with the highest error rates for trials reported as *certainly wrong*, $M = 88.9\%$, and the lowest error rate for the trials reported as *certainly correct*, $M = 3.0\%$. Across participants, confidence varied with accuracy, as expressed in Spearman rank correlations, which were found to be significantly different from zero, $rs >= -0.94,\ ps <= 0.01$, except for one participant, $r = -0.21,\ ps <= 0.74$, who did not have enough trials in the two lowest categories of the confidence scale to obtain a stable correlation estimate.

Average confidence. Mean levels of confidence for each of the four conditions, separately for correct and error trials, are presented in Figure 84. First, there was a reliable effect of accuracy, with more confident responses for correct, compared to error trials, $F(1,\ 19) = 153.0,\ p < 0.001,\ \eta_p^2 = 0.89$. This replicates the good resolution of confidence for this task, as established in the previous analysis. Second, the pattern found here matches what would have been expected judging from the previously reported findings regarding the relationship between difficulty and confidence: Participants were more confident, and better able to distinguish their correct and incorrect responses, when the task was easy than when it was hard. This was expressed in the fact that both stimulus mean and variance had a significant effect on confidence, as revealed by a repeated-measures ANOVA on mean confidence values: The higher the

mean colour of a stimulus array, the higher the confidence reported by participants, $F(1,\ 19) = 22.1,\ p < 0.001,\ \eta_p^2 = 0.54$. The higher the variability of a stimulus, however, the lower the confidence rating that followed the response to this stimulus, $F(1,\ 19) = 5.6,\ p = 0.03,\ \eta_p^2 = 0.23$. These two factors did interact marginally, $F(1,\ 19) = 4.3,\ p = 0.05,\ \eta_p^2 = 0.19$. Moreover, accuracy interacted reliably with both stimulus mean, $F(1,\ 19) = 80.2,\ p < 0.001,\ \eta_p^2 = 0.81$, and stimulus variance, $F(1,\ 19) = 35.1,\ p < 0.001,\ \eta_p^2 = 0.65$. As previously stated, these effects were caused by the fact that people were more accurate in the easier conditions (*high mean or low variance*) for corrects and less confident for error trials, which reflects better metacognitive insight for easier conditions. There was no reliable three-way interaction, $F(1,\ 19) = 2.4,\ p = 0.14,\ \eta_p^2 = 0.11$.

One key analysis in this context investigates whether stimulus mean and variance have equivalent effects on confidence. It was therefore tested whether confidence differed in the two medium conditions that were matched for overall difficulty: According to my hypothesis, *high variability* should decrease confidence. Consistent with this prediction, participants were reliably less confident in the *high mean, high variability* condition than the *low mean, low variability* condition, both for correct, $t(19) = 4.0,\ p < 0.001$, and error trials, $t(19) = 3.0,\ p < 0.01$. However, it is crucial to test whether this difference in confidence could have been caused by insufficient matching in the first-order performance. But participants were less confident in the *high mean, high variance* condition despite being slightly more accurate. The difference in confidence did therefore not follow from the slight difference in accuracy.

One could, however, argue that the difference in confidence between the two medium conditions reflected participants' use of the time heuristic: Participants were slightly slower in the *high mean, high variance* condition,

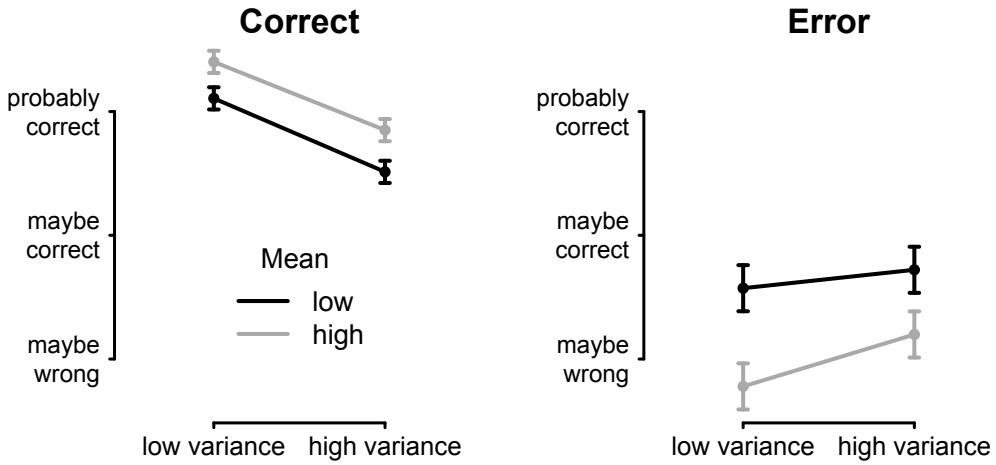

Figure 84: Confidence for correct and error trials separately.

which – according to the time heuristic – should translate into lower levels of confidence. There are, however, two arguments that speak against such an interpretation. First, if only error trials are analysed, the effect remains the same yet those error trials do not show an RT difference, $t < 1$, $BF_{NULL} = 3.71$. Second, we see the same effect in confidence in a median-split analysis for a group of 10 participants that showed no difference in correct RTs between the medium conditions, $t(9) = 1.1$, $p = 0.30$, $BF_{NULL} = 2.00$. Despite not showing a difference in RTs (but in error rates, see below), there was a significant difference in confidence for both correct, $t(9) = 2.5$, $p = 0.03$ (more confident in the *low mean, low variance* condition), and error trials, $t(9) = 2.8$, $p = 0.02$ (again, more confident in the *low mean, low variance* condition). For this median-split group, error rates in the two medium conditions were reliably different, $t(9) = 2.9$, $p = 0.02$, but notably, participants were less accurate in the *low mean, low variance* condition, $M_{lowlow} = 10.7\%$ versus $M_{highhigh} = 15.4\%$. The full factorial analysis of both median-split groups is presented in Appendices B.1 and B.2.

Taken together, stimulus mean and variance both had a reliable effect

on average levels of confidence. Critically, for the two conditions matched for first-order task performance, confidence was reduced for the *high variance* condition. Whether this reflects that stimulus variance affected confidence above and beyond its effect on first-order performance will be tested in the following section.

Comparing the influence of stimulus mean and variance on confidence. The previous analyses suggest that unreliable evidence leads to lower confidence, and that this effect is present despite matching two conditions with regard to their difficulty and therefore task performance. The question arises whether this suggests that *stimulus variance* had a more substantial impact on participants' confidence judgements as opposed to the *stimulus mean* factor. To test this question, several regression models were fitted to each participants' data and then compared. Each of these models was fitted to eight data points, that is the four conditions crossed with the two colours. Figure 85 presents the results from this model-comparison approach. Several cues will be considered as predictors for these models. The first of these is accuracy – a critical factor given that confidence is expressed as a subjective estimate of accuracy. Another cue to confidence was RT, given the above-discussed relationship between response speed and decision confidence. Mean and variance of the stimulus will form the other two predictors that will be considered here. Of these predictors, accuracy, stimulus mean, and stimulus variance are direct-access cues to confidence, the latter two reflecting evidence quality. If participants based their decision confidence on RT, on the other hand, this would constitute a case of an heuristics-based cue, given that RTs are composed of more than just the time it takes participants to decide (Ratcliff, 1979).

The null model is presented on the left of each of the three panels

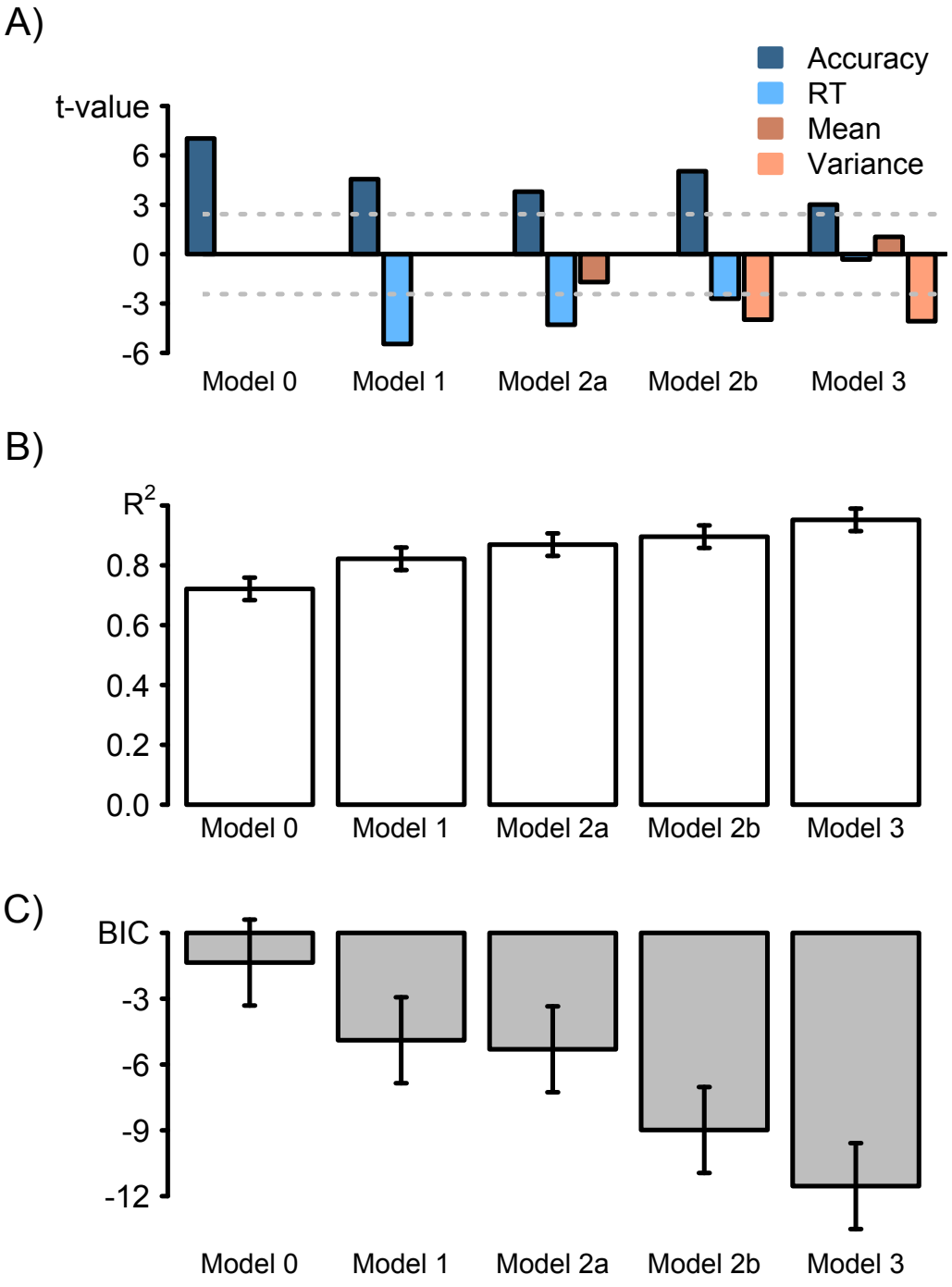

Figure 85: A) Signed t-values for the different models to predict confidence; model parameters above or below the dashed, horizontal lines are significantly different from zero. The horizontal lines highlight the critical t-values. B) Mean R^2 and C) Mean BICs for these models.

of Figure 85. This model is the normative model because it assumes that confidence is only influenced by how accurate the participant actually was. The regression weights for each participant were tested against zero using a one-sample t-test. They were found to be reliably different from zero, $t(19) = 7.030,\ p < 0.001$, also reflected in the fact that they cross the dashed grey, horizontal line in the upper panel of Figure 85. This panel presents the signed t-values. In this case, the value was positive, meaning that the more accurate a participant was, the more confident he or she would be. The explained variance of the model, as expressed in R^2, is shown in the middle panel, $R^2 = 0.72$. R^2 cannot be used, however, to compare different models, because this parameter does not take into account the number of degrees of freedom, thereby testing whether the additional parameter's cost was worth the increase in explained variance. The lowest panel therefore shows the Bayesian Information Criterion (BIC) for this model, $BIC = -1.35$, which will subsequently be compared to more complex models.

As already discussed above, there was at least a mild correlation between reported level of confidence and RT. Model 1 therefore includes RT as a second predictor. The regression weights for accuracy were again reliably different from zero, $t(19) = 4.6,\ p < 0.001$. The influence of RT was also significant, $t(19) = 5.5,\ p < 0.001$ – the slower the RT, the less confident the participants were. The R^2 was higher compared to the null model, $R^2 = 0.82$, while the BIC was now -4.89. Adding RT as an additional regressor in model 1 resulted in a significant decrease in BIC, $t(19) = 3.0,\ p < 0.01$.

Do changes in stimulus mean explain the confidence data over and above basic performance measures? This question was tested with model 2a. Critically, only the already-reported influences of accuracy, $t(19) = 3.8,\ p = 0.001$, and RT, $t(19) = 4.3,\ p < 0.001$, were found to be reliable, while the

participants' regression weights for stimulus mean were not reliably different from zero, $t(19) = 1.7,\ p = 0.11$. The model still explained a slightly larger amount of variance, $R^2 = 0.87$, but the decrease in BIC values to -5.31 was not large enough to be reliably different if compared to model 1, $t < 1$. Though it has to be noted that a non-significant result should be interpreted as the fact that the newly added parameter explains enough variance without hurting the overall explanatory value of the entire model and that the two compared models cannot be distinguished with regard to their goodness of fit.

The same question was then asked for stimulus variance, as shown in model 2b. Importantly, for this model, all three predictors were reliably different from zero, $ts >= 2.7,\ ps <= 0.01$. The R^2 for this model was 0.90, and the BIC was -8.98. The decrement in BIC from the model with just accuracy and RT (model 1) to the model including also stimulus variance (model 2b) was indeed reliable, $t(19) = 3.2,\ p < 0.01$, suggesting that stimulus variance explains between-condition differences in confidence over and above basic performance measures, consistent with the hypothesis.

Finally a model was fitted that included all four predictors (model 3). Only accuracy, $t(19) < 3.0,\ p < 0.01$ and stimulus variance, $t(19) = 4.1,\ p < 0.001$ were reliable predictors, but not stimulus mean, $t(19) = 1.0,\ p = 0.31$, and also no longer RT, $t < 1$. This model explained more variance in absolute terms, $R^2 = 0.95$. This increment in explained variance was 'worth' the additional parameter. The average BIC was -11.54 for this model, reliably different only to the model without stimulus variance (model 2a), $t(19) = 4.6,\ p < 0.001$, but only marginally significant for the model already including stimulus variance as a predictor (model 2b), $t(19) = 2.1,\ p = 0.05$. This finding speaks again for the interpretation that stimulus variance contributes to confidence over and above the general effect of first-order performance.

Taken together, these analyses show that evidence mean and evidence reliability, both affect first-order difficulty and through that also confidence. This replicates what has already been shown in the previous sections. Critically though, only evidence reliability showed an effect on subjectively-rated confidence, and this effect existed over and above the effects reliability had on first-order performance.

SDT model fits. While the previous analyses suggested that evidence reliability directly affects confidence judgements, it is not entirely clear how exactly this influence operated. One possibility is that stimulus variability, but not stimulus strength, has an effect on how well people discriminate between their own correct and error responses. The other possibility is that evidence reliability has a larger effect on metacognitive bias, if compared to evidence strength. These two hypotheses will be tested using an SDT approach.

Figure 86 presents SDT measures for the four difficulty conditions. The left panel presents metacognitive efficiency, that is how accurate participants were in distinguishing between correct and error trials while taking first-order performance into account. More precisely, metacognitive efficiency is the ratio of first- and second-order sensitivity (not shown in Figure 86). A first set of analyses was therefore focused on these sensitivity measures: Both stimulus mean and variance had a significant influence on metacognitive sensitivity – the higher the mean of the stimulus, the better participants were at discriminating their own correct from their error responses, $F(1,\ 19) = 48.4,\ p < 0.001,\ \eta_p^2 = 0.72$, whereas the opposite effect held for stimulus variance, $F(1,\ 19) = 34.3,\ p < 0.001,\ \eta_p^2 = 0.64$. These two factors did not significantly interact, $F < 1$. There was no significant difference between the two medium conditions, $t < 1,\ BF_{NULL} = 3.58$. The question arises as to

whether the main effects on metacognitive sensitivity were caused by first-order performance. With regard to first-order sensitivity, there was again a reliable effect of stimulus mean, $F(1,\ 19) = 140.0,\ p < 0.001,\ \eta_p^2 = 0.88$, and stimulus variance, $F(1,\ 19) = 96.9,\ p < 0.001,\ \eta_p^2 = 0.84$, but no interaction, $F(1,\ 19) = 1.0,\ p = 0.32,\ \eta_p^2 = 0.05$. Finally, first- and second-order sensitivity can then be combined to form metacognitive efficiency. For this parameter, however, there was no reliable influence of neither stimulus mean nor variance, $Fs < 1$, nor was there an interaction between the two factors, $F(1,\ 19) = 1.9,\ p = 0.19,\ \eta_p^2 = 0.09$. The two medium conditions were matched, $t < 1,\ BF_{NULL} = 4.16$, suggesting that participants were equally good at detecting their own errors in those performance-matched conditions. Taken together, these results suggest that neither stimulus mean nor variance have an influence on how well participants distinguished between correct and error responses over and above the effect of these factors on the colour decision itself.

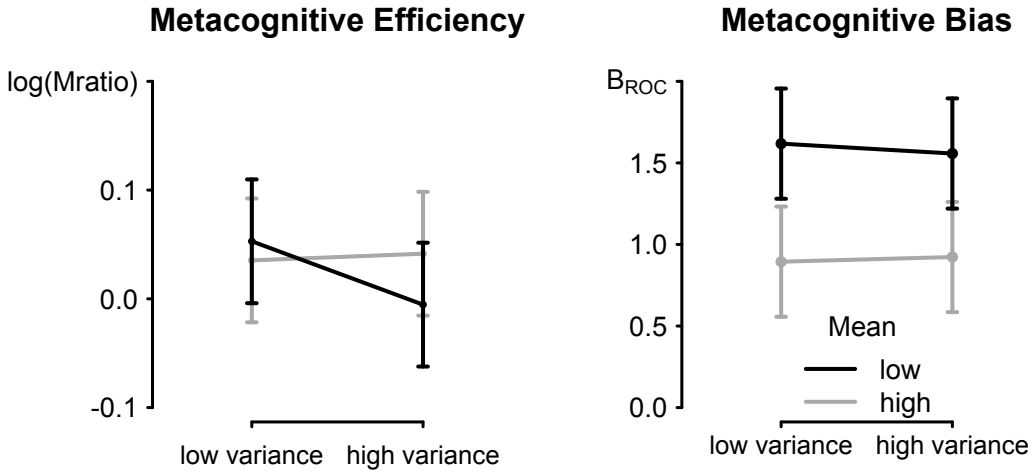

Figure 86: Type-II SDT parameters as a function of difficulty condition. The left panel presents metacognitive efficiency ($log(M\text{-}ratio)$); the right panel metacognitive bias (B_{ROC}).

The second analysis focused on metacognitive bias, that is the overall tendency to classify responses as correct or incorrect independent of their actual accuracy. If participants rate their responses more likely to be correct than they are objectively, they are overconfident. If, however, they rate them as less likely to be correct than they are, then they are underconfident. In this context, a hard-easy effect has often been observed, that is participants tend to be overconfident in the more difficult condition and well calibrated or underconfident in the easy conditions. The question arises as to how stimulus mean and variance affect such metacognitive bias. The right panel of Figure 86 shows the metacognitive bias for the four conditions. Only stimulus mean had a significant effect on metacognitive bias, B_{ROC}: Participants tended to rate confidence as higher when the stimulus mean was low, regardless of objective accuracy, $F(1,\ 19) = 26.7,\ p < 0.001,\ \eta_p^2 = 0.58$, consistent with the hard-easy effect. Surprisingly, however, given the pervasiveness of the hard-easy effect, stimulus variability showed no such effect on metacognitive bias, $F < 1$. The two factors also did not interact significantly, $F < 1$. The results from this analysis were further supported by the fact that there was most evidence in favour of a Bayesian model with just the main effect of stimulus mean, $BF = 125.56$. The two medium conditions were found to be significantly different, $t(19) = 4.2,\ p < 0.001$.

Taken together, these findings suggest that there was a hard-easy effect only for stimulus mean. In other words, participants were more overconfident in their *low mean* responses, compared to their confidence judgements following *high mean* trials. This could be interpreted as evidence that people fail to scale their confidence appropriately to the difficulty of the task if this difficulty is caused by changes in stimulus mean. For stimulus variance, on the other hand, there was no such effect, and participants' confidence responses

followed objective accuracy, correctly taking into account the changes in difficulty caused by changes in stimulus variance. These findings support the hypothesis that evidence reliability can be more directly translated into confidence, given its direct relationship to the inherent uncertainty of a mental representation.

Individual difference in the influence of stimulus variance on metacognitive processing. In the previous section, I have shown evidence in support of the hypothesis that stimulus reliability has an effect on confidence. An interesting question is whether there are systematic individual differences with regard to this effect. For example, regarding my evidence reliability hypothesis, some participants may be more accurate than others in translating the inherent spread of an internal representation into a subjective rating of confidence. Participants who show more sensitivity to changes in evidence reliability are presumably also more likely to react to high variability stimuli by adopting the previously discussed selective sampling strategy, as analysed above in Section 5.1.2.1: When colour variance is high, participants have been observed to down-weight the influence of the most extremely coloured elements ('clearly red' or 'clearly blue' elements).

I therefore tested whether there is a correlation between the extent to which participants down-weight outliers and the extent to which they adjust their confidence in response to evidence reliability as opposed to evidence mean. Such a correlation would support the interpretation that participants who score high on both measures are more sensitive to evidence reliability both in first- and in second-order processing. Figure 87 presents a correlation between the degree to which participants selectively processed a subset of the eight items present in the stimulus display and the overconfidence effect. All

data points show the difference between the two medium conditions. Consistent with the hypothesis, this analysis showed that participants who accounted for the influence of stimulus variance on confidence more than for the influence of stimulus mean (as reflected in a difference in metacognitive bias) also sampled information more selectively in the *high variance* condition. This relationship was not significant, though, $r = 0.37,\ p = 0.11$. However, if outliers – highlighted in Figure 87 as crosses – were excluded, this relationship was reliable, $r = 0.48,\ p = 0.04$. Outliers were identified using a Grubbs test for two opposite outliers (Grubbs, 1950), $G = 4.5,\ U = 0.4,\ p = 0.04$, testing against the hypothesis that the two most extreme bias effects are outliers. This test identified only the two most extreme points (highlighted as crosses in Figure 87) as outliers for the bias effects, there were no outliers for the down-weighting effects.

Taken together, these findings provide tentative support that the overconfidence effect in the two performance-matched conditions correlated with the effect of down-weighting the most extreme elements of the colour stimulus, suggesting that some participants are more prone to the influence of stimulus variance on performance – both on first- and on second-order processing.

5.1.3 Discussion

The present experiment asked whether multiple cues contribute to the generation of confidence judgements. Recent findings suggest that both evidence strength and evidence reliability affect first-order performance (De Gardelle & Summerfield, 2011). The question here was whether these variables show similar effects on second-order performance, specifically assessing the prediction that lower evidence reliability should make participants less confident. Replicating De Gardelle and Summerfield (2011), task performance was worse when

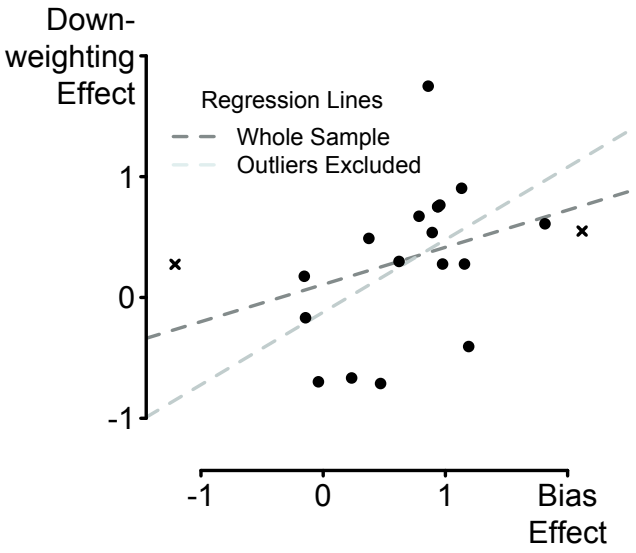

Figure 87: Scatter plot of the difference in metacognitive bias for the medium-difficulty conditions against the difference in down-weighting for these conditions; each symbol represents data from one participant. For the bias effect, but not the down-weighting effect, two outliers were identified, highlighted as crosses. The regression line for all participants is shown in dark grey. The light grey line resulted from the same analysis after exclusion of the two outlying participants.

the stimulus mean was lower or when the stimulus variance was higher. Confidence followed these effects, meaning that participants were less confident when the stimulus mean was lower or when the stimulus variance was higher. Crucially, the two performance-matched medium-difficulty conditions showed a reliable difference in confidence, with participants being less confident in the condition with higher stimulus variability: Evidence variance had an effect on confidence over and above basic task performance. This hypothesis was furthermore supported with a model-comparison approach which compared the explanatory value of different regression models aimed at explaining confidence.

Critically, regression models also revealed that evidence strength of a stimulus did not contribute to confidence in the same way as evidence reliability: Only evidence reliability but not evidence strength predicted confidence over and above first-order performance. Indeed, this interpretation was supported by the fact that only for evidence strength was there a hard-easy effect, meaning participants were overconfident on *low mean*, that is more difficult, trials, as opposed to *high variance* trials, for which they successfully adjusted their confidence. Neither stimulus variance nor mean, on the other hand, had a reliable effect on participants' ability to distinguish between their correct and error responses if first order performance is taken into account.

One intriguing hypothesis is that participants underestimate the detrimental effect of stimulus mean on accuracy while being able to correctly adjust their confidence judgements for the different stimulus variance conditions due to the fact that differences in stimulus variance reflect the 'native language' of confidence. Confidence or uncertainty judgements could be represented internally as the reliability or spread of a mental representation. Participants can very easily translate this reliability into decision confidence, correctly ac-

counting for changes in difficulty due to evidence reliability. The influence of stimulus mean, on the other hand, is much more difficult to estimate. The experimental factor representing evidence strength (stimulus mean) is a relative factor, meaning that to say whether a stimulus has a *low* or *high mean* requires them to judge the mean colour against an internally represented decision boundary, which can be noisy too. According to this hypothesis, participants should have a higher metacognitive bias in the *low mean, low variance* compared to the *high mean, high variance* condition. If someone is particularly good at accounting for the effect of stimulus variance on difficulty, then they should also show a larger influence of stimulus variance on element weighting. Evidence in support of this hypothesis was indeed found, suggesting that enhanced sensitivity to evidence reliability shows effects both with regard to first- and to second-order processing. In other words, participants who were more efficient at taking into account the effect stimulus variance had on their performance, also showed a larger effect of down-weighting outlying elements in the colour stimulus.

Evidence reliability had an effect on decision confidence over and above basic task performance. The question arises as to why this was not found to be the case for evidence mean. One interpretation which has been suggested here is that signal reliability is the 'native language' of confidence. Shea et al. (2014) have recently suggested that metacognitive representations can be processed both fast, automatically, but non-conscious (*system-I* metacognition) or slow, effortful, but conscious (*system-II* metacognition). Such system-I metacognitive information might, for example, be represented as "variance in the firing rate of a population of neurons" (p. 186; Shea et al., 2014). Using such a precision or uncertainty estimate of a respective mental representation to calculate confidence judgements (system-II metacognition) is presumably

straightforward, whereas taking into account the signal strength of a stimulus requires an additional step of comparing the mean colour of a stimulus to an internal decision boundary. It should be noted, though, that this might only apply to categorisation tasks such as the one studied here.

A recent study by Zylberberg et al. (2014) investigated the effect of variability of evidence on confidence but found the opposite effect to that reported here with noisier stimuli leading participants to more confident responses. However, there is a crucial difference between the paradigm used in that study and the paradigm used here: The conditions used in the present study were created by sampling stimuli from a pre-specified distribution, but rejecting those that exceeded a narrowly-defined tolerance level (cf. De Gardelle & Summerfield, 2011). The resulting conditions did therefore not overlap with regard to their stimulus means and variances. In Zylberberg et al. (2014), on the other hand, such trimming of the distributions was not performed, and could therefore have been more difficult for participants to tell apart stimuli from different difficulty conditions. The authors suggest that participants used the same decision criteria for conditions with different internal noise, presumably failing to adopt their criteria. One could therefore argue that due to the factorial nature of our tasks, participants were more likely to tell conditions apart and it was thus easier for them to adjust their criteria between trials.

Several different precursors and causes for confidence have been suggested over the years, resulting in different models of how confidence judgements are formed. The present experiment suggests a multi-cue model of confidence with different internal sources and signals that contribute to a final confidence judgement (see Nelson, Gerler & Narens, 1984; Koriat & Levy-Sadot, 2001, for a similar suggestion regarding feeling-of-knowing judgements; also see p. 71, Dunlosky & Metcalfe, 2009, for a review of these approaches). The predictors

that successfully predicted confidence came from both types of models that have been suggested in the past: direct-access and heuristics-based cues. An example for such direct access would be the influence of stimulus mean and variance on confidence, both reflecting evidence quality. At the same time, confidence was also inferred from RTs, in accordance with the time heuristic. Future research will have to closely examine how much each of these cues contributes to the final confidence judgement and in what role context, such as a stress on speed or accuracy, might play in this process.

5.2 Experiment 7: Effects of serotonin on decision confidence

The results of Experiment 6 indicate that decision confidence is influenced by different cues, such as RTs, as well as properties of the stimulus that affect task difficulty. The question arises as to what influences the contribution of each of these cues, thereby influencing metacognitive efficiency and bias both in different situations and across different participant groups. One clinical group that has shown interesting effects on confidence are patients with depression. The "depressive realism hypothesis" (Moore & Fresco, 2012) states that people with depression are more likely to be correct in judgements regarding for example contingencies between stimuli and responses, as well as judging their own performance. According to this view, participants with depression are less confident because they do not suffer from cognitive distortions such as overconfidence, due to an illusion of control, or neglecting a base rate. Others (Hancock, 1996; Fu, Koutstaal, Fu, Poon & Cleare, 2005; Szu-Ting Fu, Koutstaal, Poon & Cleare, 2012; see also Dunlosky & Metcalfe, 2009, for a review of these findings) have suggested that patients with depression are

less confident because of a tendency towards judging things more negatively, the "selective processing hypothesis". This view was supported by the finding that in situations where controls show good metacognitive calibration or are even underconfident, patients with depression are even more underconfident. Such an effect was shown both for overall performance estimates (Fu et al., 2005), as well as for trial-by-trial confidence judgements (Szu-Ting Fu et al., 2012). Moreover, participants with depression seem to exhibit underconfidence mainly in their correct responses (Hancock, 1996; Wood, Moffoot & O'Carroll, 1998).

Interestingly, however, the finding that participants with depression have a tendency to rate responses as low confident has not always been observed. In fact, Dunning and Story (1991) have shown the exact opposite: They let participants make judgements about future events and express their confidence about those judgements. Both students who suffered from subclinical or mild depression, as well as students who rated themselves as not depressed were overconfident in this task, rating their future predictions as more accurate as they really were. The degree to which they were overconfident, however, was stronger for the group with depression. This finding was therefore the exact opposite to the one reported above, suggesting that individuals who were depressed display differences in metacognitive processing, but the direction of these effects is by no means clearly established.

The present study tested how precisely metacognitive processing is affected by serotonin. Serotonin, or 5-hydroxytryptamine (5-HT) is a neurotransmitter which has repeatedly been linked to negative mood and depression (Asberg, Thoren, Traskman, Bertilsson & Ringberger, 2003). Indeed, antidepressants such as selective serotonin reuptake inhibitors (SSRIs) focus on inhibiting absorption of serotonin, thereby increasing the levels of serotonin

available in the brain. Studies investigating this relationship of serotonin and cognitive processing have typically used a methodology known as acute tryptophan depletion (ATD). L-Tryptophan – here only referred to as *tryptophan* – is an essential amino acid that humans obtain through their regular diet and that is subsequently synthesised into serotonin. ATD makes use of this process by asking participants to observe a special low-protein diet prior to an experiment. This diet prevents participants from obtaining this crucial precursor to serotonin, leading to a decrease in serotonin in the brain similar to that observed in participants with depression (Young et al., 1985). ATD therefore mimics clinical depression in healthy individuals. For example, tryptophan-depleted participants often rate their mood lower compared to controls (Young et al., 1985). However, it has been argued that this effect is not as stable as initially thought and that it depends on other variables, such as gender and genetic predispositions for mood disorders (Young & Leyton, 2002). In healthy participants, these effects of lowered mood are usually found in the subclinical range (Young & Leyton, 2002). Moreover, the effects of this method are reversible, meaning that once participants consume food containing tryptophan, neurotransmitters revert back to their normal levels (Young, 2013).

The present experiment used ATD as a method to transiently decrease serotonin levels in the brain. Apart from trying to replicate findings from Experiment 6, the key question then becomes whether the contribution of different cues to confidence, such as signal strength and signal reliability, was altered in tryptophan-depleted individuals, which would explain how precisely the above-described effects of over- and underconfidence are caused.

The first set of analyses focused on the effect of ignoring outliers as occurring in highly variable stimulus arrays. It has previously been shown that ATD increases people's ability to attend to relevant information while

ignoring outliers (Ahveninen et al., 2002; Schmitt et al., 2000). One could therefore expect the tryptophan-depleted group to be less influenced by the effect of stimulus variance on RTs and error rates.

The second set of analyses focused on confidence judgements. First, I tested whether the above-described effects of confidence could be replicated, in particular with regard to the finding that evidence reliability drives confidence judgements over and above its effect on basic task performance. However, given that I expect the tryptophan-depleted group to be affected less by evidence reliability in their first-order performance, I should expect them to also show a less pronounced effect of reliability on confidence.

I furthermore predicted that mean and variance of the stimulus would have no effect on metacognitive efficiency, but that stimulus mean would cause changes in metacognitive bias as observed in Experiment 6, meaning that participants display a tendency to be overconfident in the *low mean* condition. When comparing the groups, I test whether tryptophan-depleted participants display an effect of underconfidence when compared to the group whose tryptophan levels were set back to normal levels after an amino acid drink containing tryptophan. As described above, effects in both directions have been found, which both speak against the hypothesis that tryptophan-depleted people are better calibrated than controls. Both over- and underconfidence affect the bias parameter in an SDT framework. I should therefore expect to find the two groups to differ with regard to their metacognitive bias, but not their metacognitive efficiency.

5.2.1 Methods

5.2.1.1 Participants

The sample comprised 53 participants in total, 27 of whom were female. Five participants had to be excluded due to technical problems. Two further participants were excluded prior to the data analysis because their performance was almost at chance (45.1% and 46.5% errors) and an apparently random use of the confidence scale (average confidence difference between errors and corrects was close to zero on an arbitrary scale from -6 to 6; 0.28 and -0.37). This resulted in a final sample of 46 participants, of whom 23 were female and 6 were left-handed. The participants' ages ranged from 18 to 43 years (M = 24.2) and all people had self-reported normal or corrected-to-normal vision. All participants gave informed consent and all procedures were approved by the local ethics committee.

5.2.1.2 Acute tryptophan depletion (ATD) and testing procedures

In an initial session prior to the study, participants were screened for mood and addictive disorders, as well as physical or psychiatric illnesses prior to the study using a semi-structured interview from the Structured Clinical Interview for DSM-IV-TR Axis I Disorders (First, Spitzer, Gibbon & Williams, 2002). This interview was conducted by a trained psychiatrist or psychologist. In the same session, participants also completed questionnaires of cognitive ability (Raven's Standard Progressive Matrices Sets A, B, C, D, & E; Raven, 1996) and impulsivity (Barratt Impulsiveness Scale; Patton, Stanford & Barratt, 1995). Participants also completed questionnaires of state affect using the

Positive and Negative Affect Schedule (PANAS; Watson, Clark & Tellegen, 1988) as a baseline measure.

Participants were asked to follow a low-protein diet: They were allowed a maximum of 20 g of protein the day before the study. Upon arrival in the lab at 8.00 am on the day of the study, measures of state affect were again taken. Blood samples of 6 ml were taken from each participant to measure the plasma tryptophan at baseline. Over a period of approximately 30 to 40 minutes, participants then consumed an amino-acid drink: Twenty-four participants were administered an amino-acid drink containing tryptophan (T+ group), the other 22 participants received an amino-acid drink without tryptophan (T- group). The precise composition of the amino-acid drinks for male and female participants, respectively, was: L-alanine (5.5 g; 4.58 g), L-arganine (4.9 g; 4.08 g), L-cystine (2.7 g; 2.25 g), glycine (3.2 g; 2.67 g), L-isoleucine (8.0 g; 6.67 g), L-leucine (13.5 g; 11.25 g), L-lysine monohydrochloride (11.0 g; 9.17 g), L-methionine (3.0 g; 2.5 g), histidine (3.2 g; 2.67 g); L-phenylalinine (5.7 g; 4.75 g), L-proline (12.2 g; 10.17 g), L-serine (6.9 g; 5.75 g), L-threonine (6.5 g; 5.42 g), L-tyrosine (6.9 g; 5.75 g), L-valine (8.9 g; 7.42 g). For the T+ group, the drink furthermore contained L-tryptophan (2.3 g; 1.92 g). The taste of the drink was masked with 5 g (15 calories; 1.3 g carbohydrates) of citric (or malic) acid (cherry-and-vanilla or grapefruit) and artificial sweetener. Participants reported transient nausea as a side effect.

Participants then occupied themselves with reading or watching television. At lunchtime, they were given a low-protein lunch (2 g of proteins). Five hours after administration of the amino-acid drink, another blood sample was taken to measure the reduction in plasma tryptophan. Follow-up measures of state affect were also collected. Participants then completed the perceptual decision-making task described in the following section as part of a larger

battery of tests. Participants were discharged at around 5:00 pm. They were remunerated for their participation in the study (approximate £120, depending on wins and losses in another task).

All testing was carried out double-blind and placebo-controlled. This meant that the experimenter who interacted with participants during the day was not aware of their assignment to one of the two intervention groups. I was unblinded and therefore not involved in data collection, except for providing the experimental software with which participants were being tested.

5.2.1.3 Decision task

The present experiment was very similar to Experiment 6. I therefore only report where it differed from the methods already described in Section 5.1.1.2. Participants first practised the colour task (one block of 64 trials) and then the confidence judgements (another block 64 trials). The main experiment consisted of four blocks, of which 16 trials with auditory feedback were followed by 64 trials without feedback and instead with confidence judgements.

As opposed to the previous experiment, colour levels were fixed in this study. Due to time constraints on the days of the study, the experiment had to be short and could therefore not include a staircase to set the level of difficulty to provide perfect matching. I therefore took the average values for the *low mean* condition that resulted from the adaptive procedure used in Experiment 6.

5.2.2 Results

The first set of analyses focused on the question of whether the two groups (after excluding outliers as discussed above) differed with regard to variables

such as age or intelligence. If the two groups turned out not to be matched, this would lead to difficulties in the interpretation of any possible effects given that those would not necessarily have been caused by ATD but instead by the confounding variable.

First, the two samples were perfectly matched for gender: There were 12 men and 12 women in the T+ group and 11 men and 11 women in the T- group. Participants were slightly younger in the T+ group, $M_{T+} = 23.3$, compared to the T- group, $M_{T-} = 25.2$. However, this difference was not reliable, $t(37.9) = 1.2,\ p = 0.25,\ BF_{NULL} = 1.94$. Participants were also matched with regard to their intellectual abilities as expressed in their score in the Raven's matrices, $M_{T+} = 55.0$, $M_{T-} = 53.9$, $t(42.3) = 1.1,\ p = 0.27,\ BF_{NULL} = 2.04$. The T+ group showed numerically slightly higher scores on the impulsivity scale, $M_{T+} = 56.9$, compared to the T- group, $M_{T-} = 53.1$. However, this difference was not reliable, $t(36.6) = 1.6,\ p = 0.11,\ BF_{NULL} = 1.15$.

Moreover, I compared the two groups with regard to their plasma tryptophan levels in the morning as well as 5 hours after administration of the amino-acid drink. Plasma tryptophan should be reduced in the T- group but increased in the T+ group. Those different patterns in plasma tryptophan were reflected in a reliable interaction between time at which the blood sample was taken and group, $F(1,\ 37) = 82.3,\ p < 0.001,\ \eta_p^2 = 0.69$, as data from both groups and time points were submitted to a mixed-model ANOVA. The pattern was found in the hypothesised direction with total plasma tryptophan reliably reduced in the T- group over time, $M_{T-,t1} = 51.6\mu g/ml$ versus $M_{T-,t2} = 19.6\mu g/ml$, $t(18) = 6.0,\ p < 0.001$, but increased in the T+ group, $M_{T+,t1} = 49.4\mu g/ml$ versus $M_{T-,t2} = 96.2\mu g/ml$, $t(19) = 6.9,\ p < 0.001$. This mixed-model ANOVA furthermore revealed a reliable difference in plasma tryptophan between groups, $F(1,\ 37) = 62.1,\ p < 0.001,\ \eta_p^2 = 0.63$, as well as a marginally

significant effect of the time at which the measurement was taken, $F(1,\ 37) = 3.5,\ p = 0.07,\ \eta_p^2 = 0.09$. For the analysis of plasma tryptophan, three baseline blood samples (1 T+, 2 T-), three afternoon samples (2 T+, 1 T-) and one pair of both baseline and afternoon samples (T+) were unavailable. All values are furthermore presented in Table 1, together with standard errors of the mean.

Table 1: Demographic data (mean and standard error of the mean) of participants who consumed an amino-acid drink with (T+) and without (T-) tryptophan (TRP) in the morning and 5 hours later.

Group	T+	T-
Male/Female	12/12	11/11
Age (years)	23.3 ± 0.9	25.2 ± 1.3
Raven's matrices	55.0 ± 0.7	53.9 ± 0.8
Impulsivity score	56.9 ± 1.9	53.1 ± 1.3
Plasma total TRP at 0h (μg/ml)	49.4 ± 1.9	51.6 ± 2.2
Plasma total TRP at 5h (μg/ml)	96.2 ± 6.7	19.6 ± 4.0

I furthermore tested whether ATD affected participants' mood ratings as measured with the positive and negative PANAS scales (which were unavailable for one T+ participant). First, ratings made on the negative scale were compared for the three time points at which they were measured (baseline, morning, and afternoon) as well as the two groups. The two groups did not show any difference overall, $F < 1$, but there was a reliable effect of time, $F(1.6,\ 67.1) = 3.6,\ p = 0.04,\ \eta_p^2 = 0.08$, with participants' negative mood ratings highest at baseline, $M_{t0} = 12.3$, lowest in the morning $M_{t1} = 11.2$, and intermediate in the afternoon, $M_{t2} = 11.7$. Critically, the two factors did not show a reliable interaction, $F < 1$. For the positive scale, none of the effects were reliable, $Fs <= 1.7,\ ps >= 0.19,\ \eta_p^2 s <= 0.04$. The averaged results for the two scales, groups, and the three measurement points are furthermore presented in Table 2.

In the following section, I test whether key effects found for Exper-

Table 2: PANAS ratings for negative and positive affect (mean and standard error of the mean) of participants who consumed an amino-acid drink with (T+) and without (T-) tryptophan at baseline, in the morning and 5 hours later.

Group		T+	T-
PANAS negative	baseline	12.2 ± 0.4	12.5 ± 0.6
	morning	11.0 ± 0.3	11.5 ± 0.6
	afternoon	11.4 ± 0.5	12.0 ± 0.8
PANAS positive	baseline	28.0 ± 1.7	27.6 ± 1.8
	morning	27.8 ± 1.8	27.7 ± 1.7
	afternoon	26.5 ± 1.8	26.4 ± 1.9

IMENT 6 were replicated in the present study. I furthermore focus on the hypotheses aimed at comparing the two groups, but only with regard to the most robust analyses whilst avoiding tests that are likely to be underpowered for such a between-subject comparison.

5.2.2.1 First-order performance

Basic performance measures. The effects of stimulus mean, stimulus variance, and group were tested using two mixed-design ANOVAs with correct RTs and error rates and dependent variables, as shown in Figure 88. Similar to the results from EXPERIMENT 6, stimulus mean had a reliable effect on both correct RTs, $F(1,\ 44) = 13.6,\ p < 0.001,\ \eta_p^2 = 0.24$, and error rates, $F(1,\ 44) = 266.9,\ p < 0.001,\ \eta_p^2 = 0.86$, with faster RTs and lower error rates for the *high mean* compared to the *low mean* condition. The main effect of stimulus variance was also replicated for both correct RTs, $F(1,\ 44) = 85.9,\ p < 0.001,\ \eta_p^2 = 0.66$, and error rates, $F(1,\ 44) = 71.0,\ p < 0.001,\ \eta_p^2 = 0.62$. Once more, the two factors showed an interaction for error rates, $F(1,\ 44) = 6.8,\ p = 0.01,\ \eta_p^2 = 0.13$, but not correct RTs, $F(1,\ 44) = 1.1,\ p = 0.30,\ \eta_p^2 = 0.02$. Moreover, a comparison of

the two groups revealed that the T- group was faster than the T+ group, $M_{T-} = 631\ ms$ versus $M_{T+} = 685\ ms$. This main effect of group was reliable, $F(1,\ 44) = 4.9,\ p = 0.03,\ \eta_p^2 = 0.10$. Numerically, the T- group committed more mistakes than the T+ group, $M_{T-} = 25.8\%$ versus $M_{T+} = 23.5\%$, but this effect was not reliable, $F < 1$.

Given the hypothesis that ATD causes a more focused attentional style (Ahveninen et al., 2002; Schmitt et al., 2000), I predicted that stimulus variance would affect the T+ groups more. Indeed, consistent with this hypothesis, the difference between the *high* and the *low variance* condition was larger for the T+ group for both correct RTs, $M_{T+} = 207\ ms$ versus $M_{T-} = 97\ ms$, and error rates, $M_{T+} = 11.9\%$ versus $M_{T-} = 5.5\%$, as expressed in the flatter slopes of the T- group presented in Figure 88. These effects were statistically significant as expressed in reliable interactions between group and stimulus variance for both correct RTs, $F(1,\ 44) = 5.4,\ p = 0.03,\ \eta_p^2 = 0.11$, and error rates, $F(1,\ 44) = 9.1,\ p < 0.01,\ \eta_p^2 = 0.17$. There was no interaction between group and stimulus mean, $Fs < 1$, and also no three-way interactions, $Fs < 1$. These findings are consistent with the hypothesis that tryptophan-depleted participants have a stronger tendency to ignore outlying elements in a variable stimulus array.

Matching of the medium conditions. No staircase was used in this experiment and the question therefore arises as to whether the two medium conditions were still matched for performance. This was indeed the case for the T- group if efficiency was considered, $t < 1,\ BF_{NULL} = 3.71$, but not for the T+ group, $t(23) = 2.2,\ p = 0.04$, who exhibited better efficiency scores in the *low mean, low variance* condition, $M_{lowlow} = 860\ ms$, compared to the *high mean, high variance* condition, $M_{highhigh} = 927\ ms$. Moreover, in a factorial ANOVA

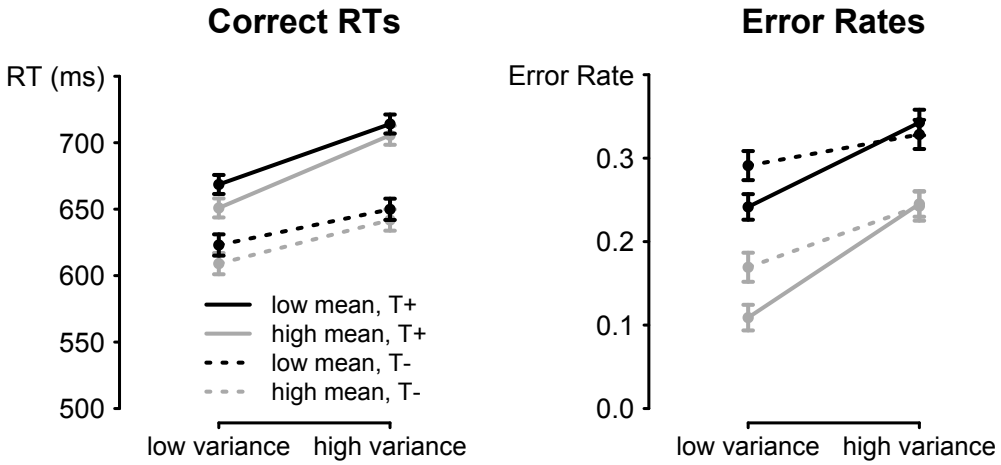

Figure 88: Mean correct-trial response times (RTs) and mean error rates as a function of condition and treatment group. T+: participants who consumed an amino-acid drink with tryptophan; T-: participants who consumed an amino-acid drink without tryptophan; ms: millisecond.

including all four experimental conditions, there were reliable main effects of stimulus mean, $F(1,\ 44) = 133.3,\ p < 0.001,\ \eta_p^2 = 0.75$, and stimulus variance, $F(1,\ 44) = 90.3,\ p < 0.001,\ \eta_p^2 = 0.67$, on efficiency, as well as the above reported interaction of stimulus variance and group, reflecting the reduced effect of stimulus variance for the T- group, $F(1,\ 44) = 11.3,\ p < 0.01,\ \eta_p^2 = 0.20$. There was also a marginally significant interaction between stimulus mean and variance, $F(1,\ 44) = 2.8,\ p < 0.10,\ \eta_p^2 = 0.06$. There was no main effect of group, $F(1,\ 44) = 1.1,\ p = 0.30,\ \eta_p^2 = 0.02$. There were also no further reliable interactions, $Fs < 1$.

However, analysing RTs and error rates separately revealed a speed-accuracy tradeoff: For correct RTs, the two medium conditions were reliably different for both the T+, $t(23) = 5.6,\ p < 0.001$, and the T- group, $t(21) = 4.0,\ p < 0.001$, with faster RTs for the *low mean, low variance* condition. For the T- group, there was also a reliable difference in error rates, with higher error rates for the *low mean, low variance* condition, $t(21) = 2.3,\ p = 0.03$,

suggesting a similar speed-accuracy tradeoff as was observed in EXPERIMENT 6. Such an effect was not found for the T+ group, however, $t < 1$.

I furthermore tested whether this matching of conditions was also evident in other measures that combine accuracy and response speed. One such measure is drift rate in a diffusion model. There was no reliable difference between the two medium conditions for this measure, neither for the T+, $t(22) = 1.4,\ p = 0.19,\ BF_{NULL} = 2.03$, nor the T- group, $t(21) = 1.4,\ p = 0.18,\ BF_{NULL} = 1.95$. This contradicts the findings for the efficiency measure, according to which there was a reliable difference for those conditions in the T+ group.

In conclusion, analyses of first-order effects replicated the effect of evidence strength and reliability observed in EXPERIMENT 6. Tryptophan-depleted participants were responding faster, but were not reliably less accurate. This difference in response speed is congruent with the finding that ATD makes people respond more impulsively in a continuous performance test (Walderhaug et al., 2002; see also Worbe, Savulich, Voon, Fernandez-Egea & Robbins, 2014). Tryptophan depletion has furthermore been linked to a more efficient suppression of outliers (Ahveninen et al., 2002; Schmitt et al., 2000). Congruent with this notion, there was a reduced effect of stimulus variance for the T- group. Moreover, the two medium conditions of this design were not matched with a staircase, which resulted in reliable differences in performance for these two conditions for at least some of my measures of difficulty. Comparison between these two conditions with regard to confidence therefore must be treated with caution.

5.2.2.2 Second-order performance

The next set of analyses compares the two groups with regard to their metacognitive performance. Wherever applicable, I test whether the present study replicates findings from Experiment 6.

Average confidence. Figure 89 presents mean confidence as a function of stimulus mean, stimulus variance, objective accuracy and group. As expected, there was a reliable main effect of accuracy with higher confidence for correct compared to error trials, $F(1,\ 44) = 196.5,\ p < 0.001,\ \eta_p^2 = 0.82$. There was also a reliable main effect of stimulus mean, $F(1,\ 44) = 24.6,\ p < 0.001,\ \eta_p^2 = 0.36$, which was qualified by a significant interaction with objective accuracy, $F(1,\ 44) = 100.2,\ p < 0.001,\ \eta_p^2 = 0.69$. Subjects expressed higher confidence for the *high mean* condition than the *low mean* condition for correct trials but showed the opposite pattern for error trials. In contrast to the previous experiment, there was no main effect of stimulus variance, $F(1,\ 44) = 2.0,\ p = 0.16,\ \eta_p^2 = 0.04$. There was, however, a replication of the interaction between stimulus variance and accuracy, $F(1,\ 44) = 45.2,\ p < 0.001,\ \eta_p^2 = 0.51$, as can also be seen in the slopes being negative for correct trials and positive for error trials in Figure 89. Stimulus mean and variance did not reliably interact, $F(1,\ 44) = 2.3,\ p = 0.14,\ \eta_p^2 = 0.05$. Taken together, the findings reported here suggest that the effects stimulus mean and variance have on confidence are stable.

Interestingly, there was a main effect of group on mean confidence, $F(1,\ 44) = 4.1,\ p < 0.05,\ \eta_p^2 = 0.08$. This reflected the fact that participants in the T- group were more confident in both their correct responses, $M_{T+} = 4.7$ versus $M_{T-} = 4.9$, as well as their errors, $M_{T+} = 3.3$ versus $M_{T-} = 3.6$. When tested separately, however, this difference was only mar-

ginally reliable for correct trials, $t(43.9) = 1.7,\ p = 0.09$, and not reliable for errors, $t(39.0) = 1.3,\ p = 0.19$. The group factor did not interact with stimulus mean, $F(1,\ 44) = 1.7,\ p = 0.21,\ \eta_p^2 = 0.04$, nor stimulus variance, $F(1,\ 44) = 1.7,\ p = 0.20,\ \eta_p^2 = 0.04$, nor objective accuracy, $F < 1$. There were also no interactions between group, objective accuracy, and stimulus mean or variance, $Fs < 1$. There was, moreover, a marginally significant interaction between group, stimulus mean, and variance, $F(1,\ 44) = 2.9,\ p = 0.09,\ \eta_p^2 = 0.06$, indicating that there was an interaction between stimulus mean and variance for the T+ group, $F(1,\ 23) = 4.4,\ p < 0.05,\ \eta_p^2 = 0.16$, but not the T- group, $F < 1$. There was furthermore a reliable interaction between stimulus mean, stimulus variance, and objective accuracy, $F(1,\ 44) = 9.9,\ p < 0.01,\ \eta_p^2 = 0.18$, as well as a significant four-way interaction, $F(1,\ 44) = 5.6,\ p = 0.02,\ \eta_p^2 = 0.11$. However, in both cases, the effects were not systematic and I will therefore not discuss them further. Taken together, tryptophan-depleted participants were found to be more confident compared to the T+ group.

I next compared the two medium-difficulty conditions with regard to average confidence, although this analysis is complicated by the fact that the two groups were not perfectly matched in primary task performance for at least the T+ group – participants were faster in the *low mean, low variance* condition. For the T+ group, there was a reliable difference in confidence on correct trials, $t(23) = 5.6,\ p < 0.001$, while on error trials this difference was only marginally significant, $t(23) = 1.9,\ p = 0.07$. In both cases, the *high mean, high variance* condition was less confident than the *low mean, low variance* condition: $M_{highhigh,cor} = 4.49$, $M_{lowlow,cor} = 4.79$, $M_{highhigh,err} = 3.35$, $M_{lowlow,err} = 3.58$. For the T- group, the findings were similar, with a reliable difference between condition only for correct, $t(21) = 2.5,\ p = 0.02$, but not for error trials, $t(21) = 1.4,\ p = 0.17$. The *high mean, high vari-*

ance condition was again the one with lower average confidence in both cases: $M_{highhigh,cor} = 4.78$, $M_{lowlow,cor} = 4.95$, $M_{highhigh,err} = 3.48$, $M_{lowlow,err} = 3.68$. Those findings therefore replicate what has previously been reported for EXPERIMENT 6 – lower evidence reliability leads to decreases in confidence – an effect observed in both T+ and T- groups despite somewhat different patterns of primary task performance across the two conditions.

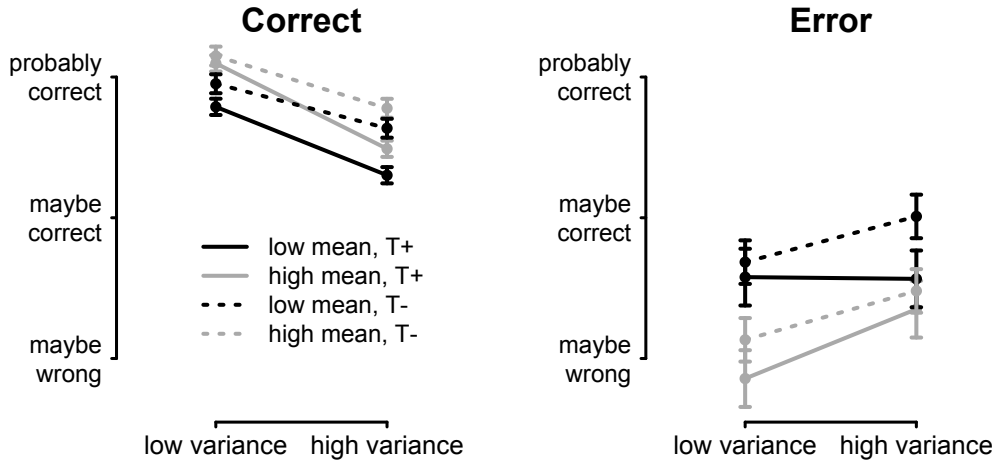

Figure 89: Confidence as a function of objective accuracy, condition and treatment group. Only part of the 6-point confidence scale is used here to improve visibility. T+: participants who consumed an amino-acid drink with tryptophan; T-: participants who consumed an amino-acid drink without tryptophan.

Comparing the influence of stimulus mean and variance on confidence. As was done for EXPERIMENT 6, several regression models were fitted to test which variables contributed to a participant's confidence judgements. The question here was whether this influence of variables differed between the two groups, especially with regard to the variance of the stimuli. Figure 90 presents the results from this model-comparison approach. The solid bars represent data from the T+ group and the dashed bars represent data from the

T- group.

The normative model in which only accuracy predicted confidence is shown on the left of each of the three panels. This predictor was reliably different from zero for both the T+, $t(23) = 16.0,\ p < 0.001$, and the T- group, $t(21) = 9.2,\ p < 0.001$. The t-values are shown in the upper panel of Figure 90. The R^2 value, as shown in the medium panel, was larger for the T+, $M_{T+} = 0.69$, compared to the T- group, $M_{T-} = 0.50$, $t(34.3) = 2.9,\ p < 0.01$. There was, however, no reliable difference between BIC scores, $t < 1$, and the BIC of the T- group was actually smaller compared to the T+ group, $M_{T+} = 2.97$ versus $M_{T-} = 1.43$. The BIC scores are shown in the lower panel of Figure 90.

RT was then included as an additional predictor in model 1. The regression weights for accuracy were again reliably different from zero for both the T+, $t(23) = 7.3,\ p < 0.001$, as well as for the T- group, $t(21) = 6.0,\ p < 0.001$. The RT predictor was also reliably different from zero for both the T+, $t(23) = 5.2,\ p < 0.001$, and the T- group, $t(21) = 3.8,\ p = 0.001$. As in the previous experiment, this predictor was negative, meaning that RT and confidence were negatively correlated. This regression model explained slightly more variance for the T+ group, $R^2 = 0.78$, if compared to the T- group, $R^2 = 0.67$, $t(36.0) = 2.0,\ p < 0.05$. However, as mentioned above, the R^2 value does not take into account the number of degrees of freedom. Ultimately, the BIC values of the two models have to be compared. These values were not reliably different, though, $t < 1$. The question arises whether the increase in explanatory power was worth adding the additional parameter. This was tested separately for the two groups. For both the T+, $t(23) = 2.2,\ p = 0.04$, and the T- group, $t(21) = 2.3,\ p = 0.03$, the difference in BIC scores was indeed reliable, which can be interpreted as evidence for the superiority of the

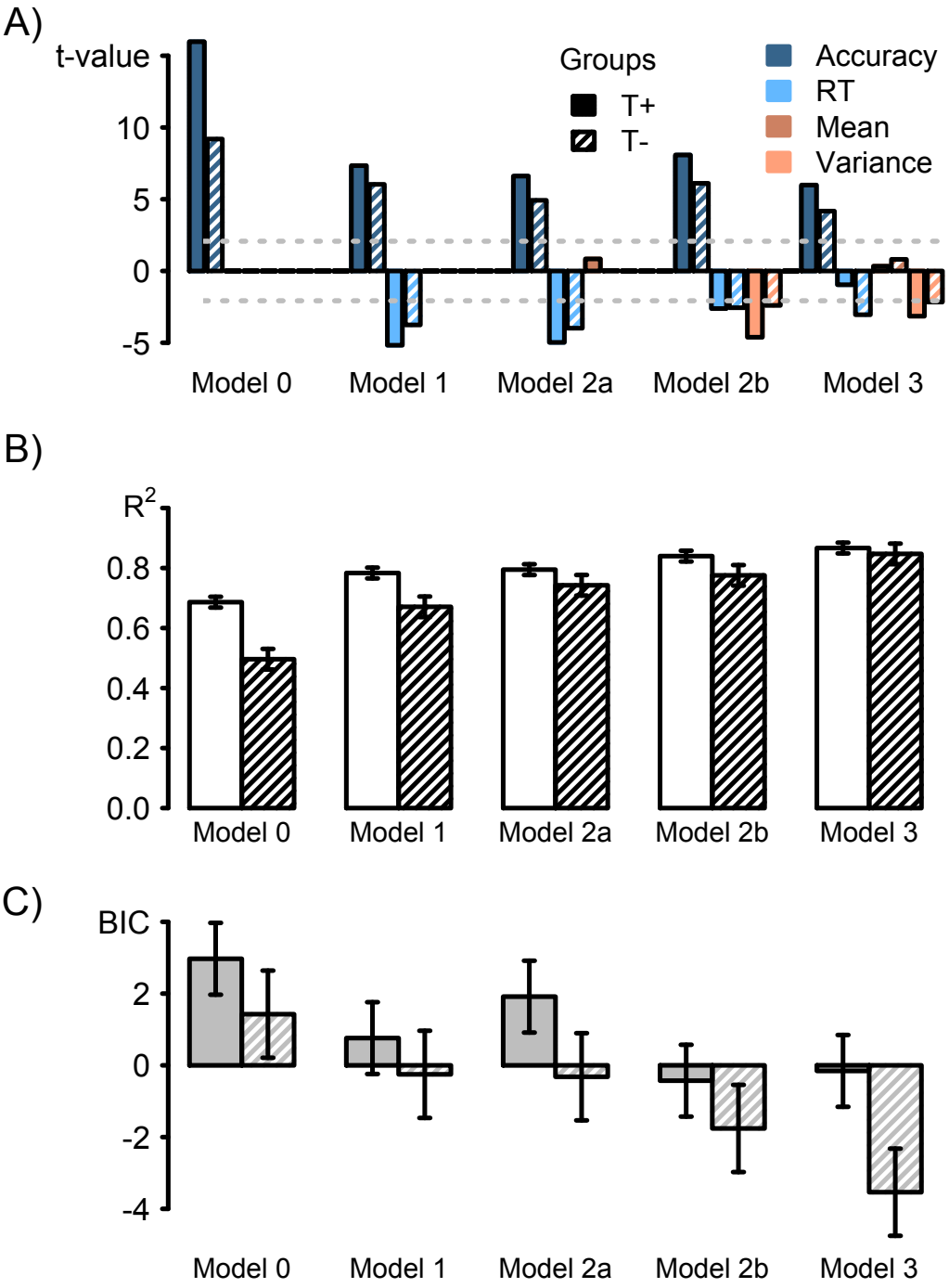

Figure 90: A) Signed t-values for the different models to predict confidence; model parameters above or below the dashed, horizontal lines are significantly different from zero. The horizontal lines highlight the critical t-values. These values are very similar for the two groups and are therefore overlapping. B) Mean R^2 and C) Mean BICs for these models. T+: participants who consumed an amino-acid drink with tryptophan; T-: participants who consumed an amino-acid drink without tryptophan.

more complex model, model 1, over the normative model 0.

Does stimulus mean predict confidence over and above its effect on basic task performance? This question was addressed with model 2a. For both groups, only accuracy and RT were reliable predictors of decision confidence, $ts >= 4.0,\ ps < 0.001$. The mean of the stimulus did not have such an effect, $t < 1$. The two groups did not differ with regard to the variance explained by the model, R^2, $t(41.3) = 1.1,\ p = 0.27$, nor the model fit, BIC, $t < 1$. Adding stimulus mean as an additional parameter did not improve the model fit for the T- group, $t < 1$. For the T+ group, it actually decreased the fit, $t(23) = 3.5,\ p < 0.01$. It can therefore be concluded, that stimulus mean is not a good predictor of confidence over and above the effect stimulus mean has on accuracy and RT.

The results from the previous experiment suggested that stimulus variance, on the other hand, had a reliable effect on confidence over and above basic task performance. This hypothesis was tested with model 2b. Echoing earlier findings, all predictors were significantly different from zero, $ts >= 2.4,\ ps <= 0.03$. Stimulus variance was negatively correlated with confidence – the more variable a stimulus was, the less confident the participant would rate his response. The two groups did not differ in their model fits, neither with regard to their R^2s, $t(39.9) = 1.5,\ p = 0.14$, nor their BIC values, $t < 1$. If this model was compared to model 1 for each group separately, the goodness of fit was not found to be reliably better, neither for the T+ group, $t(23) = 1.4,\ p = 0.19$, nor for the T- group, $t(21) = 1.5,\ p = 0.14$. As noted before, a non-reliable difference in BIC scores does not mean that the fit of the more complex model was worse, only that the new parameter explains enough variance in itself without hurting the overall explanatory value of the entire model.

Finally, a model was fitted which included all four parameters as pre-

dictors, model 3. For the T+ group, only accuracy, $t(23) = 6.0,\ p < 0.001$, and stimulus variance, $t(23) = 3.1,\ p < 0.01$, were reliable predictors, but not RT or stimulus mean, $ts < 1$. These findings replicated those of EXPERIMENT 6. For the T- group, all predictors but stimulus mean, $t < 1$, were reliably different from zero, $ts >= 2.2,\ ps <= 0.04$. The models of the two groups did not differ with regard to their proportion of explained variance, $t < 1$, or goodness of fit, $t(43.6) = 1.3,\ p = 0.21$. If compared to models 2a and 2b, the increment in explained variance was only 'worth' adding the additional parameter when this additional parameter was stimulus variance, $ts = 2.2,\ ps = 0.04$, but not when stimulus mean was added to model 2a, $ts <= 1.7,\ ps >= 0.10$. These results were very similar for the two groups.

Taken together, these findings again support the interpretation that stimulus variance but not mean contributes to confidence over and above the general effect of first-order performance. This effect was very similar for both the T+ and T- groups, therefore going against the hypothesis that tryptophan-depleted participants had a less pronounced effect of evidence reliability on confidence.

SDT model fits. The next set of analyses focused on whether stimulus mean, stimulus variance and ATD had a reliable effect on people's metacognitive ability. For metacognitive sensitivity, *meta-d′* exhibited reliable main effects for both stimulus mean, $F(1,\ 44) = 28.0,\ p < 0.001,\ \eta_p^2 = 0.39$, and stimulus variance, $F(1,\ 44) = 39.3,\ p < 0.001,\ \eta_p^2 = 0.47$, replicating the results of EXPERIMENT 6. Critically, the T+ and T- groups did not significantly differ with regard to their metacognitive sensitivity, $F(1,\ 44) = 1.4,\ p = 0.25,\ \eta_p^2 = 0.03$. No other effects were reliable, $Fs <= 1.8,\ ps >= 0.19$.

However, as previously discussed, first-order performance has to be

taken into account to assess whether differences in first-order processing might have caused this null effect by cancelling out true differences in the ability to distinguish between correct and error trials. There was again a main effect of stimulus mean, $F(1,\ 44) = 155.2,\ p < 0.001,\ \eta_p^2 = 0.78$, and stimulus variance, $F(1,\ 44) = 97.2,\ p < 0.001,\ \eta_p^2 = 0.69$, replicating the findings from Experiment 6. There was now also an interaction between the two factors, $F(1,\ 44) = 20.1,\ p < 0.001,\ \eta_p^2 = 0.31$. I found no difference between the two groups, $F(1,\ 44) = 1.9,\ p = 0.18,\ \eta_p^2 = 0.04$, but a reliable interaction between the group factor and stimulus variance, $F(1,\ 44) = 13.0,\ p < 0.001,\ \eta_p^2 = 0.23$, reflecting the above-reported reduced effect of stimulus variance for the T- group. No other effects were significant, $Fs <= 1.2,\ p >= 0.28$.

First- and second-order sensitivity can then be compared to assess how metacognitively efficient participants were. This analysis is presented in the left panel of Figure 91. Four participants had to be excluded from this analysis because the algorithm could not estimate efficiency parameters due to missing values in some of the data cells. There was first of all a marginally significant effect of stimulus mean, $F(1,\ 40) = 4.0,\ p = 0.05,\ \eta_p^2 = 0.09$, as opposed to the null effect found in Experiment 6. Overall, there was a higher metacognitive efficiency when stimulus mean was low, ($M_{low} = 0.16$) than when it was high ($M_{high} = 0.09$). As in the previous experiment, no effect of stimulus variance was found, $F(1,\ 40) = 1.8,\ p = 0.19,\ \eta_p^2 = 0.04$, and also no interaction between those two factors, $F < 1$. There was no main effect of group, $F < 1$, meaning participants in the T- group were not less metacognitively efficient. Interestingly, though, there was a reliable interaction of the group factor and stimulus variance, $F(1,\ 40) = 5.7,\ p = 0.02,\ \eta_p^2 = 0.12$: Whereas the T+ group showed no effect of stimulus variance on metacognitive efficiency, the T- group was worse at detecting their own errors (independent of how many

errors they actually made) when variability of the stimulus was high. Group and stimulus mean did not interact, $F(1,\ 40) = 1.6,\ p = 0.22,\ \eta_p^2 = 0.04$, but there was a reliable three-way interaction between group, stimulus mean, and stimulus variance, $F(1,\ 40) = 5.3,\ p = 0.03,\ \eta_p^2 = 0.12$. Despite not having perfectly matched medium conditions, two conditions were compared for the two groups separately. There was no reliable difference for the T+ group, $t(22) = 1.3,\ p = 0.20$, but a marginally significant difference for the T- group, $t(18) = 1.9,\ p = 0.08$, with a slightly higher metacognitive efficiency for the *low mean, low variance* condition, $M_{highhigh} = 0.11$ versus $M_{lowlow} = 0.28$.

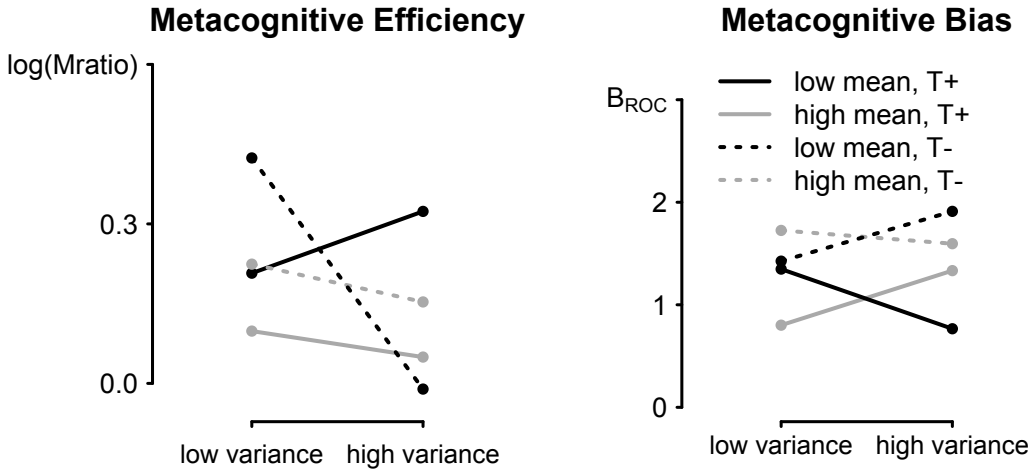

Figure 91: Type-II SDT parameters as a function of condition and treatment group. The left panel presents metacognitive efficiency ($log(M\text{-}ratio)$); the right panel metacognitive bias (B_{ROC}). No error bars are shown because they are highly overlapping and hinder interpretation of the figure. T+: participants who consumed an amino-acid drink with tryptophan; T-: participants who consumed an amino-acid drink without tryptophan.

In conclusion, these findings suggest that overall, tryptophan-depleted participants were not affected in their metacognitive ability. There was, however, a hint of an interaction effect between group and stimulus variance, suggesting that participants in the T- group were less affected by the stimulus variability in their primary choice, but more affected by it in their secondary

judgement.

As already discussed above, participants in the T- group were reliably more confident on both correct and error trials. This finding hints at the fact that there is a difference in metacognitive bias between the two groups, with the tryptophan-depleted participants using the end of the confidence scale more often that classifies responses as *correct*. Also mentioned above, the non-parametric bias parameter B_{ROC} (Kornbrot, 2006) is relatively susceptible to empty data cells. The present experiment had low trial numbers, which unfortunately resulted in 13 participants being excluded from the analysis. The results therefore have to be interpreted with caution. As previously found, there was no effect of stimulus variance on metacognitive bias, $F < 1$. However, the above-found effect hard-easy effect for stimulus mean was also not reliable here, $F < 1$. The two factors showed a marginally significant interaction, though, $F(1,\ 31) = 3.8,\ p = 0.06,\ \eta_p^2 = 0.11$, as reflected in the cross-over patterns that can be seen in the right panel of Figure 91. As expected, participants in the T- group had a higher metacognitive bias, $M_{T-} = 1.65$ versus $M_{T+} = 1.06$. However, this effect was only marginally significant, $F(1,\ 31) = 3.3,\ p = 0.08,\ \eta_p^2 = 0.10$. No other effects were reliable, $Fs <= 2.6,\ ps >= 0.12$. These findings suggest that the difference in confidence between the two groups was caused by a difference in calibration, that is the T- group mapped the confidence scale onto their responses slightly differently, being less likely to classify responses as incorrect. In contrast to the previous study, there was no difference between the two medium conditions, neither for the T+ group nor the T- group, $ts < 1$.

5.2.3 Discussion

The present experiment had two key goals. The first goal was to assess whether the effects reported for Experiment 6 could be replicated with this study. As in the previous experiment, there were reliable effects of signal strength and reliability on response speed, error rates, and efficiency. Due to time constraints, no staircase was applied prior to the main experiment. Instead, the 'best' parameters from the previous experiment were chosen for all participants and difficulty was not further adjusted. Comparisons of these medium-difficulty conditions should therefore be treated with caution. However, post-hoc analyses provided evidence that the conditions were well-matched for at least one of the groups (T-), despite showing a speed-accuracy trade-off as in Experiment 6.

As in the previous experiment, stimulus mean and variance affected both correct- and error-trial confidence, but they did so in opposite ways: A higher stimulus mean made participants more confident on correct trials, but less confident on error trials. The same pattern held for *low variance*, even though this effect was not reliable as a main effect. There was once more a reliable difference in confidence between the two medium conditions with higher confidence in the *low mean, low variance* condition. This effect was reliable for correct trials in both groups. However, interpreting this effect in the present experiment is a little more complicated due to imperfect first-order performance matching. The above-reported difference could therefore – at least for the T+ group – stem from the fact that the T+ group was more efficient in the *low mean, low variance* condition.

The second aim of the present study was to test differences in metacognitive processing for the two groups. Those groups were well-matched with regard to gender, age, intelligence, and impulsivity. Plasma tryptophan

samples confirmed an increase in tryptophan for the group of participants who consumed an amino-acid drink containing tryptophan (T+) but a decrease for participants whose drink did not contain tryptophan (T-), if baseline measures were compared to afternoon samples of blood. However, despite this clear difference in tryptophan, there were no reliable differences in mood rating between the participant groups. However, this does not necessarily mean that the manipulation was unsuccessful. In fact, it has previously been noticed that the effects of ATD on mood are rather unstable (Young & Leyton, 2002). One possible reason for that could be that the mood effects caused by ATD are very weak and that the PANAS scales are not sensitive enough to detect them. However, there were no numerical trends in the hypothesised direction in the PANAS ratings, speaking against a mere lack of statistical power. Instead, one could assume that the ATD effects that mirror depression are not mediated via mood.

Prior to investigating group differences in metacognitive processing, several differences in first-order processing were analysed. First, tryptophan-depleted participants were on average found to respond faster, consistent with prior suggestions that tryptophan depletion leads to more impulsive behaviour (Walderhaug et al., 2002; Worbe et al., 2014). Moreover, tryptophan-depleted people are often better at ignoring irrelevant information (Ahveninen et al., 2002; Schmitt et al., 2000). Consistent with this observation, tryptophan-depleted participants showed a weaker effect of stimulus variance. This effect was present in correct RTs, error rates, and efficiency.

The tryptophan-depleted participants were overall more confident, mirroring Dunning and Story's (1991) finding in patients with subclinical depression. This increased confidence was found for both correct and error trials. This effect was therefore due to an increased metacognitive bias, that is

tryptophan-depleted participants were more likely to use the high-confidence end of the scale independent of their ability to distinguish between correct and error trials. However, there was only a hint of such an effect for the metacognitive bias parameter B_{ROC}. Also, time constraints imposed upon the present study resulted in a problem for the estimation of metacognitive bias parameters. The non-parametric bias measure B_{ROC}, chosen here, could not be calculated for several participants due to missing values in some of the confidence conditions. This finding suggests that even in cases where the parameter could be calculated, this estimate must have been imprecise or unstable. This could have resulted in the failure to replicate the effect of stimulus mean on metacognitive bias.

The causes for the differences in metacognitive processing were further investigated using a model comparison approach. This analysis replicated findings from EXPERIMENT 6, namely that stimulus variance but not stimulus mean had an effect on confidence over and above the effect it had on decision performance. However, the two groups did not differ with regard to how they used the different confidence cues and the present experiment therefore did not provide an explanation as to why metacognitive processing is affected by changes in serotonin levels. Other known confidence cues, such as familiarity or fluency (Schwartz, 1994; Koriat, 1993) should be taken into account as well in future experiments.

Taken together, these findings suggest that low serotonin levels lead to relative overconfidence in participants if compared to a group of participants whose serotonin levels had previously been restored to their normal levels. The tryptophan-depleted participants' performance was also relatively less affected by increases in stimulus variance. Surprisingly, however, this group showed a stronger effect of stimulus variance on their metacognitive efficiency. This find-

ing could mean that those participants are somehow more sensitive to stimulus variability, which leads them to 'protect' themselves more from the detrimental effects low evidence reliability can have on performance. At the same time, however, low evidence reliability makes them worse at distinguishing their own correct and error responses. Future studies could investigate further whether this is a valid explanation of the effects observed here. Those effects could not be explained by differently contributing cues to confidence, at least not for the cues studied in the present experiment. I found no reliable differences in metacognitive efficiency between the two groups.

5.3 Experiment 8: Neurophysiological mechanisms of stimulus mean and variance processing

The last experiment in this thesis investigated the neurophysiological mechanisms of how stimulus mean and variance processing influence cognitive processing. More precisely, I addressed the question of whether the influence of evidence reliability on confidence – that is, less reliable evidence leads to decreases in confidence even for conditions of matched first-order performance – was reflected in the ERN and especially the Pe. These error-related EEG correlates were already studied in Experiments 4 and 5, where they have been found to reflect subjectively-rated decision confidence. The question here therefore becomes whether Pe amplitude reflects an internal source that feeds into confidence judgements.

5.3.1 Methods

Participants were trained and staircased and then completed four blocks on a colour and shape classification task without confidence judgements and then three blocks on the colour discrimination task with confidence judgements. My analysis focused on the latter three blocks. The project was run in collaboration with another student, whose analyses focused on the first four experimental blocks that are not considered further below. In the methods section, I explicitly highlight work conducted together.

5.3.1.1 Participants

We tested 17 participants in total, of whom the first one had to be excluded due to a technical failure with the stimulus presentation. All participants reported to be right-handed and to have intact colour vision. There were 9 female participants. The participants' ages ranged from 19 to 33. Some participants had already taken part in a similar colour experiment, but this was considered not to be a problem given the adaptive nature of the task.

5.3.1.2 Task and procedure

The task used was the same colour judgement task as in previous experiments in the present chapter. I only report the methods where they differed from these earlier studies.

Participants first completed 4 blocks to practise the task, while a staircase algorithm adjusted the level of difficulty. The goal was to find six difficulty conditions for each participant within a roughly similar range, resulting from a factorial combination of two levels of colour mean and three levels of colour variance. For reasons of efficiency and time constraints, the staircase focused

on only two of those conditions. A first 144 trial-long block varied mean colour, while keeping colour variance fixed at 0.2, until an error rate of 25% was obtained. This was repeated for a target error rate of 85%. This resulted in two values of colour mean: *high mean* and *low mean*, for each colour respectively. The remaining four difficulty conditions were automatically created by including variance levels of 0.1 (*low variance*) and 0.3 (*high variance*) in addition to the already used level of 0.2 (*medium variance*). This resulted in a complete, orthogonal design around the already created two conditions of medium difficulty. The same procedure was used to find difficulty conditions for a related task, a shape judgement task, performed on the same stimuli. However, this task was only used by my collaborators in this experiment (see Section 1.9) and is therefore not relevant here.

Responses were made using the *c* and *b* key of a computer keyboard, except for the first 2 participants who used a mouse for this judgement. The key assignment matched the key assignment used later in the main experiment and was therefore counterbalanced over participants. The stimulus was shown for 100 ms, participants' responses were collected up until 1500 ms after stimulus onset. If they failed to respond within this time window, the next trial would automatically begin and the current trial would be labelled as a miss. Participants received auditory feedback after each trial, in form of two tones, each 200 ms long, one of them 600 Hz, the other 1200 Hz. The order of these tones was rising for correct responses and falling for incorrect responses. Between trials, there was an RSI of variable length, on average 1000 ms, with a uniform jitter of 300 ms, ranging from 850 to 1150 ms.

The staircase began to adjust difficulty as soon as the participant had pressed the correct key on 9 consecutive trials. Following a correct trial, difficulty was increased, and following an incorrect trial it was decreased by a

small amount. The increment or decrement by which difficulty was adjusted, changed once over the course of the block. Moreover, these steps were larger for the blocks in which the target accuracy was 75%. After these staircase blocks, a summary figure was presented on screen for inspection by the experimenter. This figure displayed the adjustment of difficulty over trials. The experimenter then chose the final difficulty setting for the entire experiment, adjusting the value suggested by the computer wherever it appeared suspect, for example by being too difficult due to a run of correct trials before the last staircase block finished.

After this first part, participants were prepared for the 32-channel EEG recordings. The methods for these recordings were identical to the methods described in 3.1.1.3, and a description of the procedures will therefore not be repeated here.

Following setup, participants completed another practice block of 20 trials, just of the colour task. We then recorded EEG signals from their scalp while they performing the two parts of the experiment: The first part did not include any confidence judgements and was aimed at testing another research question. This part comprised four blocks, two colour blocks and two blocks of the above-mentioned shape task, which were shuffled in their order. Each block was 144 trials long. They received feedback on all trials.

The second part was then focused on the research question regarding confidence, which was at the centre of this study. Only data from this second half will be reported here. In this part, only colour judgements were used. In the first block of this part, which was 24 trials long, participants familiarised themselves with the 6-point confidence scale already used in previous studies of this thesis. No more feedback tones were played on trials, which instead required confidence judgements. After this block, frequencies for all

confidence categories were displayed on screen so that the experimenter could discuss these with the participant. This was done to reinforce the point that use ought to be made of all confidence categories. There then followed three main experimental blocks. Similar to EXPERIMENT 6, all were comprised of 24 trials with feedback and then 144 trials with confidence judgements. The feedback trials were included to ensure that participants would maintain a stable colour discrimination criterion throughout this part of the experiment. Those trials were discarded for all analyses. Between the feedback and confidence trials, there was a 5-second countdown to help participants prepare for the new upcoming task.

There were 2 (mean colour) x 3 (variance of colour) x 2 (objective colour of the stimulus) x 2 (mean shape) x 3 (variance of shape) x 2 (objective shape of the stimulus), so 144 conditions in total. The block length was therefore set to 144 for this experiment. All shorter practice blocks were a random subset of these conditions. Critically, the shape dimension showed no reliable effects on task performance, even though – unlike in the other experiments described in this chapter – this dimension was task-relevant in some of the previous blocks of the first part in the experiment.

Furthermore, there were too few trials per condition for this study to run SDT analyses. I will therefore not report type-II sensitivity (*meta-d′*) and bias (B_{ROC}) for this study.

5.3.2 Results

5.3.2.1 First-order performance

Basic performance measures and matching of the medium conditions. Figure 92 presents the correct RTs and error rates for the six difficulty

conditions. Participants were faster, $F(1,\ 15) = 11.8,\ p < 0.01,\ \eta_p^2 = 0.44$, and more accurate, $F(1,\ 15) = 24.0,\ p < 0.001,\ \eta_p^2 = 0.61$, when the stimulus mean was high as opposed to when it was low. The main effect of stimulus variance was also replicated for both correct RTs, $F(2,\ 30) = 24.3,\ p < 0.001,\ \eta_p^2 = 0.62$, and error rates, $F(2,\ 30) = 6.2,\ p < 0.01,\ \eta_p^2 = 0.29$: Higher stimulus variance led to increased RTs and higher error rates. These two factors did not interact, for either correct RTs, $F < 1$, nor for error rates, $F(2,\ 30) = 2.3,\ p = 0.12,\ \eta_p^2 = 0.13$, the latter finding not replicating what has been reported above for Experiments 6 and 7.

As in previous experiments, I aimed to contrast two conditions matched for difficulty but different in whether difficulty reflected *low mean* or *high variance*. The *low mean, low variance* condition and *high mean, medium variance* condition suit this purpose, highlighted in all consecutive figures using crosses as symbols. There was no difference between the conditions for correct RTs, $t(15) = 1.3,\ p = 0.23,\ BF_{NULL} = 2.01$. For error rates, there was also no significant difference, $t(15) = 1.5,\ p = 0.15,\ BF_{NULL} = 1.54$. Those conditions also did not differ in efficiency, $t < 1,\ BF_{NULL} = 3.35$. Analyses of drift rate parameters in a diffusion model were consistent with this conclusion, yielding no reliable differences between the two conditions, $t(14) = 1.1,\ p = 0.29,\ BF_{NULL} = 2.27$. This drift rate was fitted to data from not just the confidence part of the experiment, but all blocks, to increase the number of trials and therefore stability of the parameter estimates.

Taken together, the findings for the first-order performance replicated many effects that have already been found for Experiments 6 and 7. Amongst those effects are the main effects of stimulus mean and variance on performance measures such as correct RTs and error rates. Furthermore, two matched conditions were found which will further be analysed in the following

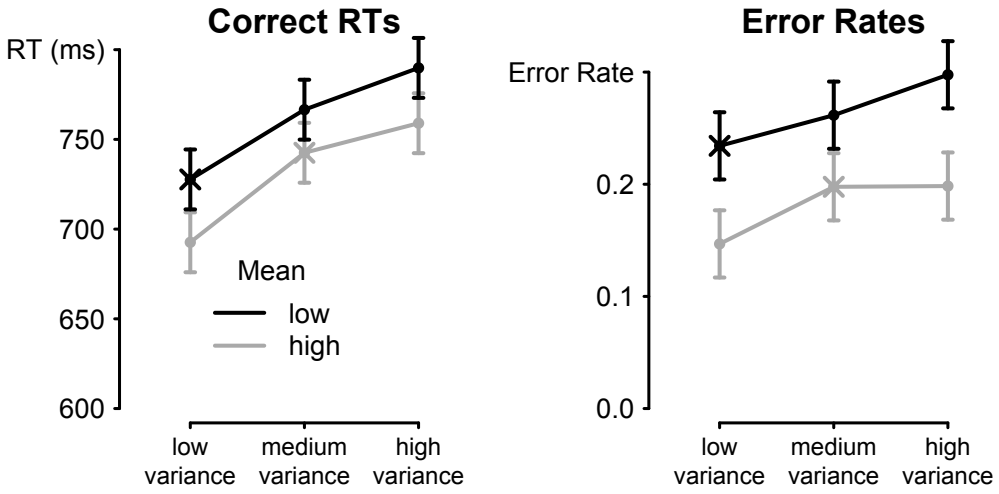

Figure 92: Mean correct-trial response times (RTs) and mean error rates as a function of condition; ms: millisecond. The conditions matched for first-order task performance are highlighted using crosses: the *low mean, low variance* condition and the *high mean, medium variance* condition.

section with regard to their confidence judgements. In conclusion, these data suggest that the colour-judgement paradigm used in this chapter produces consistent and stable effects.

5.3.2.2 Confidence judgements

As for the previously reported experiments, participants displayed a high resolution of confidence, as expressed in a reliable, negative relationship between confidence and error rates, $rs <= -0.94,\ ps < 0.01$, for 14 out of 16 participants. In addition to this, mean, $F(1,\ 15) = 26.7,\ p < 0.001,\ \eta_p^2 = 0.64$, and variance, $F(2,\ 30) = 4.6,\ p = 0.02,\ \eta_p^2 = 0.23$, affected confidence in opposite ways for correct and error trials, replicating the previous findings. Average confidence as a function of accuracy, mean, and variance is presented in Figure 93. All other effects are reported in Appendix C.1.

The two matched conditions of medium difficulty – the *low mean, low*

variance condition and *high mean, medium variance* condition, as highlighted in Figure 93 by cross symbols – were furthermore compared. Consistent with previous findings, participants expressed lower confidence in the *high mean, medium variance* condition, $M_{cor} = 4.74$, $M_{err} = 2.76$, compared to the *low mean, low variance* condition, $M_{cor} = 4.81$, $M_{err} = 3.33$. However, this difference was reliable only for error trials, $t(15) = 4.0,\ p = 0.001,\ BF = 35.81$, and not for correct trials, $t < 1,\ BF_{NULL} = 3.27$.

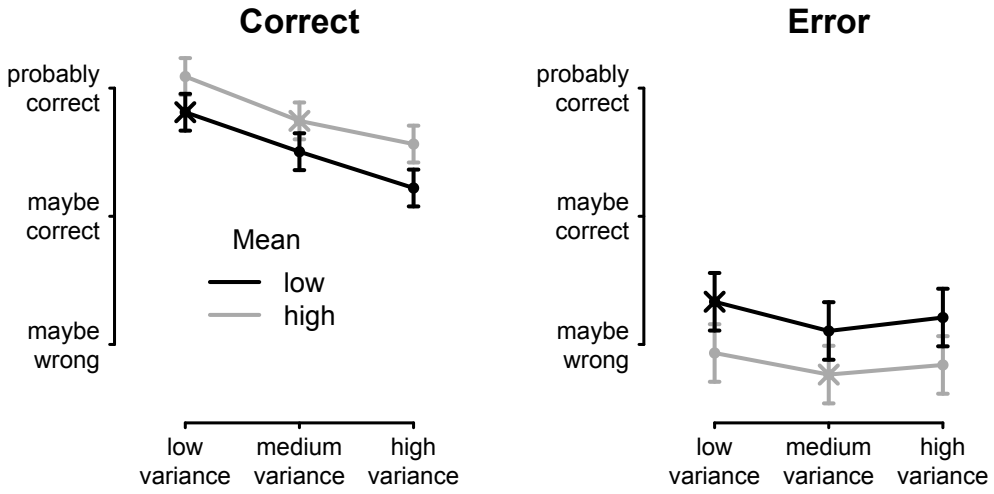

Figure 93: Confidence as a function of objective accuracy and condition. The conditions matched for first-order task performance are highlighted using crosses: the *low mean, low variance* condition and the *high mean, medium variance* condition. Only part of the 6-point confidence scale is used here to improve visibility.

Furthermore, regression model fits replicated findings from Experiments 6 and 7: variance, $t(15) = 2.5,\ p = 0.02$, but not mean, $t(15) = 1.1,\ p = 0.29$, explained confidence variance above and beyond basic task performance. The t-values, R^2s, and BIC values are furthermore presented in Figure 94. The complete analysis of models 0 to 3 is given in Appendix C.2.

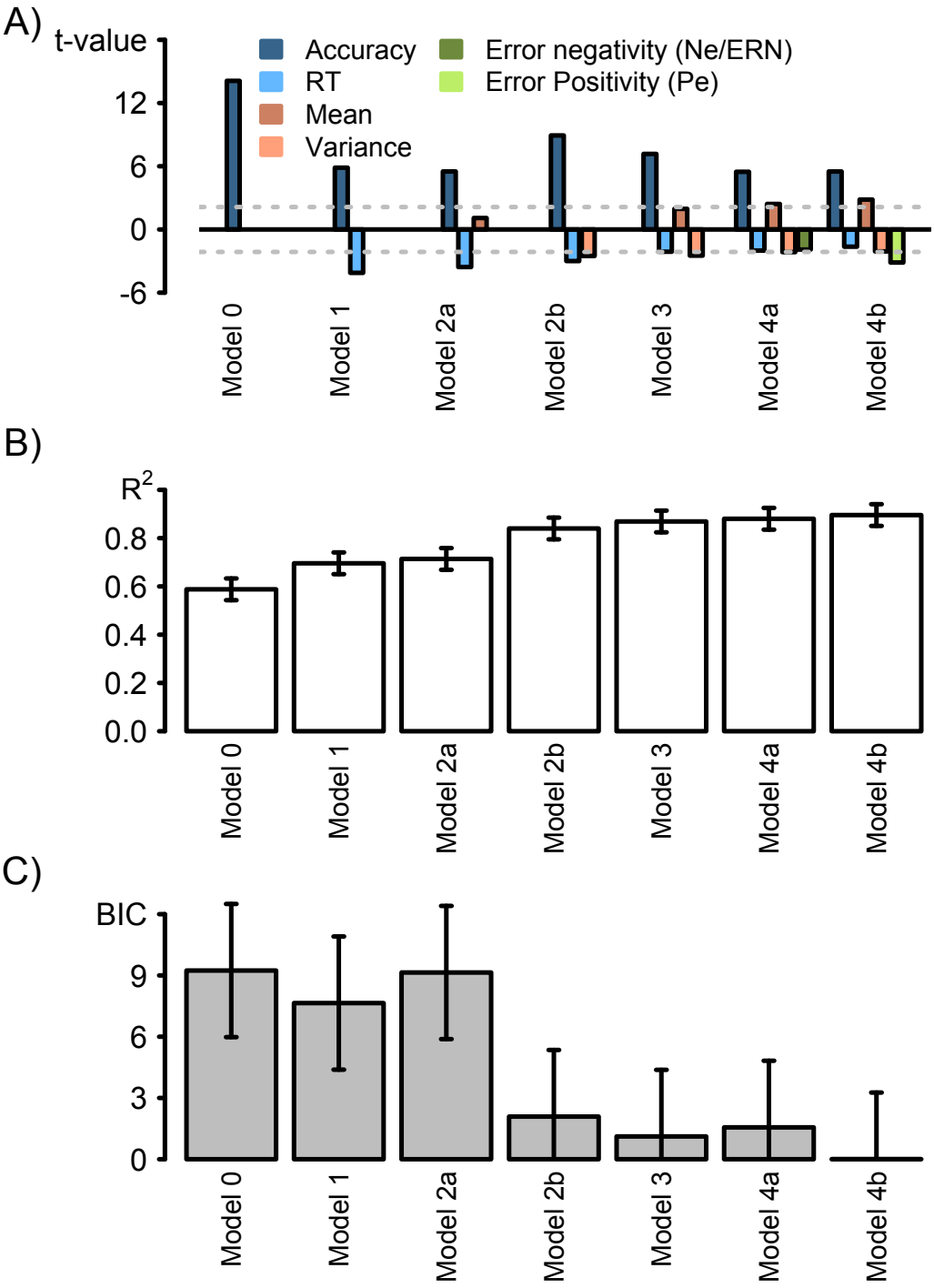

Figure 94: A) Signed t-values for the different models to predict confidence; model parameters above or below the dashed, horizontal lines are significantly different from zero. The horizontal lines highlight the critical t-values. B) Mean R^2 and C) Mean BICs for these models.

5.3.2.3 ERP data

The next section focuses on effects of stimulus mean and variance on neurophysiological correlates of errors and – as previously shown – of confidence. The first analysis asked whether the ERN and the Pe had a reliable effect of objective accuracy. Figure 95 presents the grand-averaged ERP waveforms, time-locked to the colour-task response, at electrode CZ. The time windows of the ERN (-30 to 70 ms) and the Pe (360 to 460 ms) are highlighted. These were picked by visually inspecting the scalp topographies for different time windows. The resulting topographies for those time windows are presented in Figure 96. The ERN (upper panel) had a more central, negative topography. However, the ERN amplitude is small here, and the topography is more posterior and lateralised than what would normally be observed. Presumably, this is a reflection of the difficulty of the task.

Data from all five midline electrodes for both correct and error trials were submitted to a 5 x 2 repeated-measures ANOVA. There was a reliable effect of anteroposterior scalp locations, $F(2.2,\ 33.3) = 24.4,\ p < 0.001,\ \eta_p^2 = 0.62$, expressing that the voltage was lowest at electrode location CZ, $M_{CZ} = 0.49\ \mu V$, compared to all other four locations; $M_{FZ} = -1.44\ \mu V$ $M_{FCZ} = -0.87\ \mu V$, $M_{CPZ} = 1.38\ \mu V$, $M_{PZ} = 1.25\ \mu V$. The ERN was furthermore smaller for error than for correct trials, $M_{cor} = 0.45\ \mu V$ versus $M_{err} = -0.12\ \mu V$, as would have been expected. However, this effect was not reliable, $F(1,\ 15) = 2.6,\ p = 0.12,\ \eta_p^2 = 0.15$. There was also no interaction between location and accuracy, $F < 1$, but the electrode for which data was presented in Figure 95 showed the largest difference between correct and error trials, $M = 0.69\ \mu V$.

The same analysis was conducted for the Pe. There was again a reliable

effect of anteroposterior scalp location, $F(1.8,\ 0.5) = 5.3,\ p = 0.01,\ \eta_p^2 = 0.26$, but now voltages were more positive over posterior scalp locations, $M_{CPZ} = -3.76\ \mu V$, if compared to all other four electrodes; $M_{FZ} = -1.03\ \mu V$, $M_{FCZ} = -1.71\ \mu V$, $M_{CZ} = -2.60\ \mu V$, $M_{PZ} = -3.67\ \mu V$. This is also shown in the typical posterior positive topography shown in the lower panel of Figure 96. The Pe was larger on error trials, $M_{err} = -1.90\ \mu V$, compared to correct trials, $M_{err} = -3.21\ \mu V$, as would have been expected. This difference was indeed reliable, $F(1,\ 15) = 8.8,\ p < 0.01,\ \eta_p^2 = 0.37$. There was furthermore a reliable interaction between scalp location and accuracy, $F(1.7,\ 25.1) = 14.2,\ p < 0.001,\ \eta_p^2 = 0.49$. This interaction effect reflected the fact that the difference between correct and error trials was largest at electrode PZ, $M = -2.37\ \mu V$. Taken together these analyses suggest that the Pe reflected neurophysiological differences in objective accuracy, while the ERN was only a weak index of such an effect.

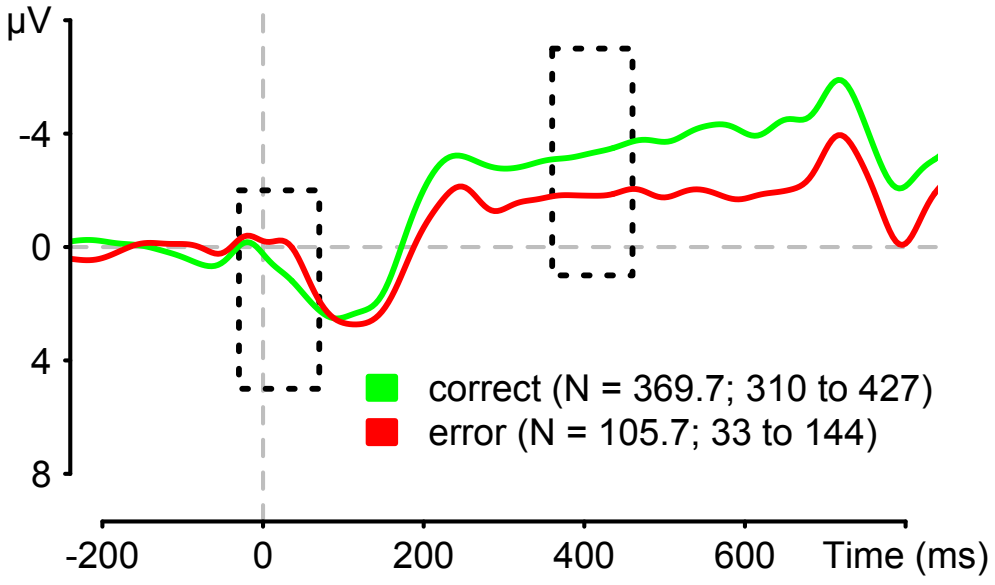

Figure 95: Response-locked ERN (-30 to 70 ms) and Pe (360 to 460 ms) at electrode CZ as a function of objective accuracy. Response was made at 0 ms. The figure legend displays average, minimum, and maximum numbers (N) of trials per participant. ms: millisecond; μV: microvolt.

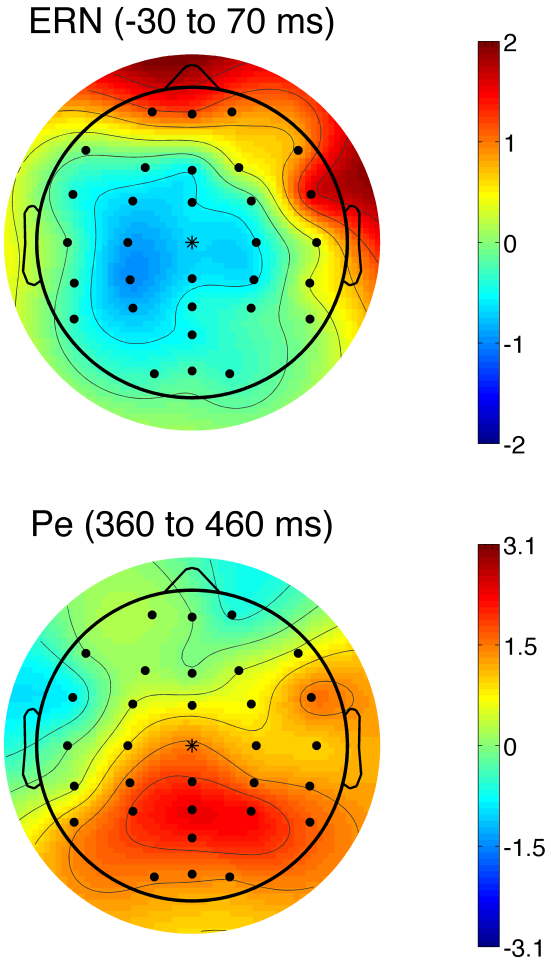

Figure 96: Topography for the response-locked ERN (error-related negativity) and the Pe (error positivity). Both topographies display the difference between correct and error trials. The colour bars to the right of the topography display the corresponding voltages (in microvolts, μV); ms: millisecond.

The next set of analyses focused on whether the ERN and the Pe once more reflected differences in confidence. The upper panel of Figure 97 shows data at electrode CZ with the time window of the ERN highlighted (again, a time window of -30 to 70 ms was chosen). Both correct and error trials were included in this analysis. There was again a reliable effect of anteroposterior scalp location, $F(2.3,\ 33.9) = 25.9,\ p < 0.001,\ \eta_p^2 = 0.63$, as well as a reliable effect of confidence, $F(5,\ 75) = 2.6,\ p = 0.03,\ \eta_p^2 = 0.15$. This effect reflects

that the ERN was the most negative for the *certainly wrong* category, $M = -0.68\ \mu V$ whilst being the most positive for the *certainly correct* category, $M = 0.95\ \mu V$. This was also reflected in a reliable linear trend, $F(1,\ 15) = 9.9,\ p < 0.01,\ \eta_p^2 = 0.77$. The upper panel of Figure 98 furthermore presents the topography of the ERN as a contrast between the highest and the lowest confidence category.

The same analysis was again repeated for the Pe time window (360 to 460 ms). The Pe effect is highlighted in the lower panel of Figure 97, which presents data for the electrode CPZ. As previously shown, there was a reliable effect of location, $F(1.9,\ 29.0) = 5.3,\ p = 0.01,\ \eta_p^2 = 0.26$. Similar to the ERN, the Pe also scaled with confidence, $F(2.9,\ 43.9) = 3.6,\ p = 0.02,\ \eta_p^2 = 0.19$. However, as would have been expected, this effect was in the opposite way, with the largest Pe values for the *certainly wrong* category, $M = -0.68$, and the smallest values for the *certainly correct* category, $M = 0.95$. There was again a reliable linear trend, $F(1,\ 15) = 9.3,\ p < 0.01,\ \eta_p^2 = 0.38$. Location and confidence did not interact, $F(5.5,\ 82.9) = 1.3,\ p = 0.29,\ \eta_p^2 = 0.08$. Furthermore, Figure 98 shows the topography for the Pe, which was a posterior, negative pattern, as expected. Taken together, these findings suggest that both the ERN as well as the Pe scale reflect changes in confidence, replicating findings from Experiments 4 and 5.

The previous analyses provided support for the hypothesis that especially Pe amplitude tracks changes in confidence, replicating findings from Experiment 4. The key question of the present experiment was whether these EEG components also reflect differences in stimulus mean and variance. More precisely, confidence was found to vary both with stimulus mean (higher confidence found at higher levels of stimulus mean) and stimulus variance (higher confidence found at lower levels of stimulus variance). I would expect to find

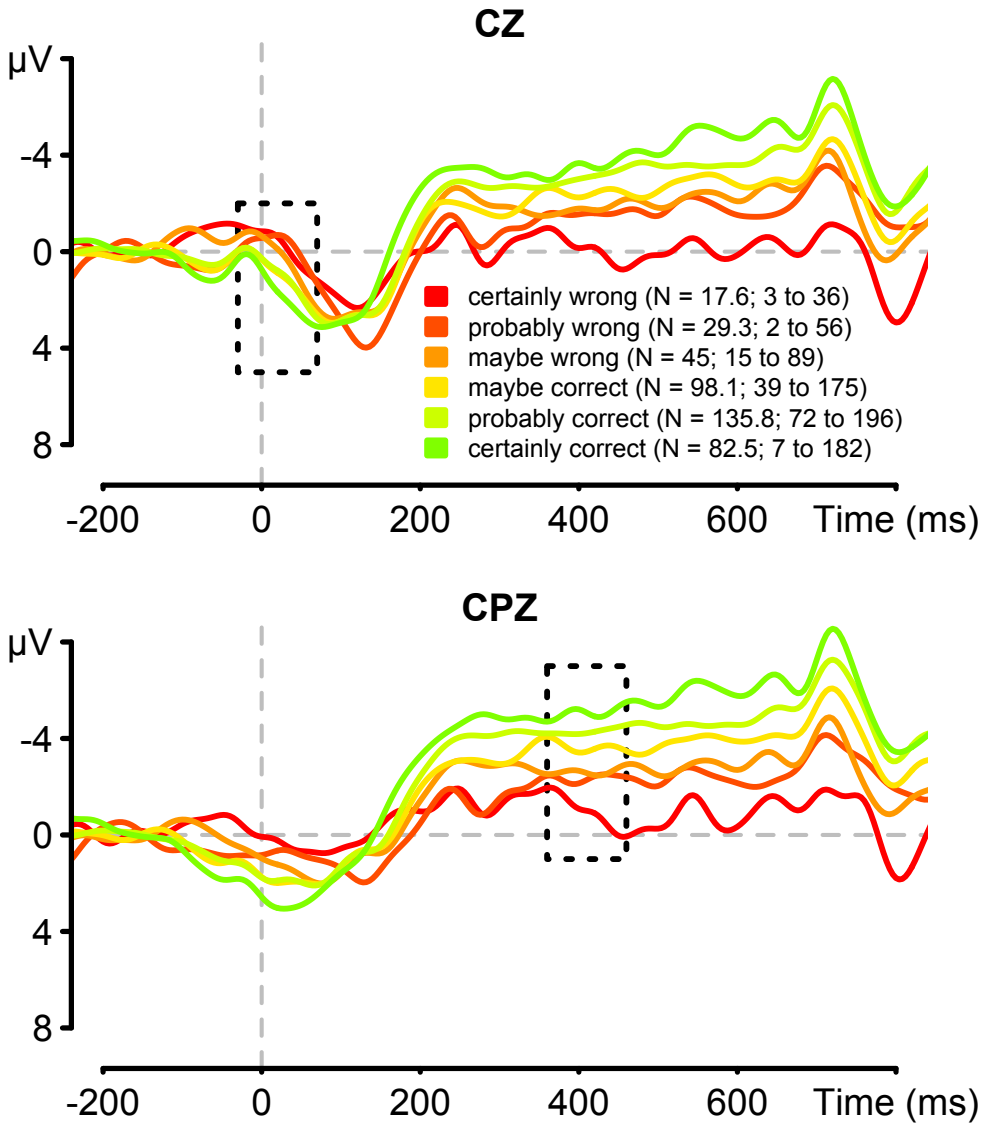

Figure 97: Response-locked ERN (-30 to 70 ms; at electrode CZ; top panel) and Pe (360 to 460 ms; at electrode CPZ; bottom panel) as a function of confidence for correct and error trials. Response was made at 0 ms. The figure legend displays average, minimum, and maximum numbers (N) of trials per participant. ms: millisecond; μV: microvolt.

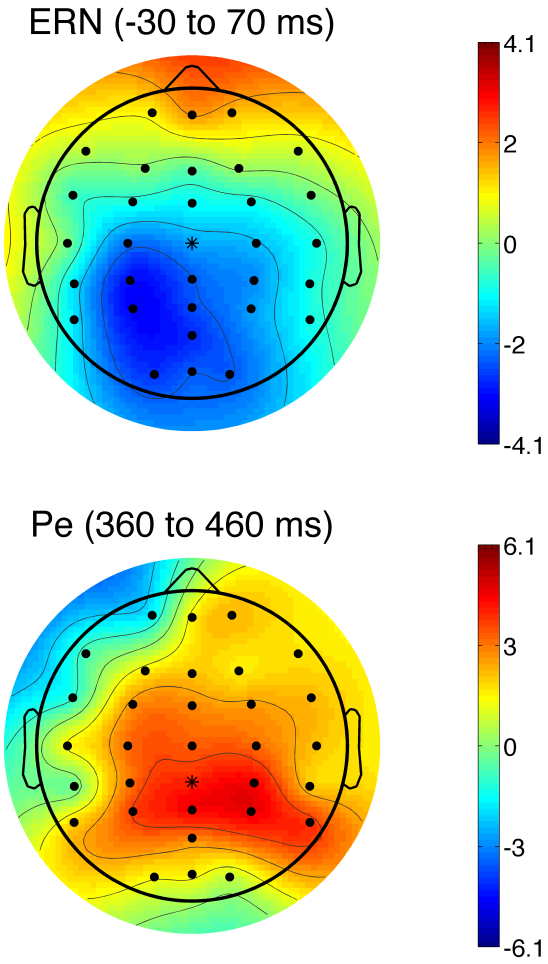

Figure 98: Topography for the response-locked ERN (error-related negativity) and the Pe (error positivity), conditioned on the level of confidence for errors and correct trials. Both topographies display the difference between the most extreme confidence categories: *certainly wrong* and *certainly correct*. The colour bars to the right of the topography display the corresponding voltages (in microvolts, μV); ms: millisecond.

those differences in confidence echoed in the amplitude of the Pe: I would predict a larger amplitude for low stimulus mean and high stimulus variance, respectively. Moreover, I would expect to find that Pe amplitude also reflects the difference in confidence between the two conditions matched for task difficulty, with a larger Pe amplitude for the *high mean, medium variance* condition compared to the *low mean, low variance* condition. This analysis focused on correct trials only, because of the low number of error trials. Furthermore, the analyses were focused on the Pe, rather than the ERN (both highlighted in Figure 99), which did not reveal any reliable interaction effects between electrode location and either mean, $F(2.1,\ 31.3) = 2.6,\ p = 0.08,\ \eta_p^2 = 0.15$, or variance, $F(2.9,\ 43.1) = 1.2,\ p = 0.34,\ \eta_p^2 = 0.07$. The full ERN analysis is given in Appendix C.3.

Figure 99 highlights the Pe in its time window again ranging from 360 to 460 ms (right time window highlighted). The average voltage from this time window was again submitted to a repeated-measures ANOVA with anteroposterior scalp location, stimulus mean, and stimulus variance as factors. This analysis revealed a reliable effect of location, $F(2.0,\ 29.4) = 7.8,\ p < 0.01,\ \eta_p^2 = 0.34$, with the largest activity observed at the most frontal electrode FZ, $M = -1.06$. There were no main effects of either stimulus mean, $F < 1$, or stimulus variance, $F(1.4,\ 21.5) = 2.7,\ p = 0.11,\ \eta_p^2 = 0.15$. There was, however, a reliable interaction between stimulus variance and electrode location, $F(3.2,\ 47.6) = 2.9,\ p = 0.04,\ \eta_p^2 = 0.16$, reflecting that for example the difference between the *high variance* and the *low variance* condition was larger (absolutely) for the more posterior electrodes; $M_{FZ} = -0.2\ \mu V$, $M_{FCZ} = -0.54\ \mu V$, $M_{CZ} = -0.59\ \mu V$, $M_{CPZ} = -1.27\ \mu V$, $M_{PZ} = -1.61\ \mu V$. There were no such interaction for location and stimulus mean, $F < 1$, and also no three-way interaction, $F < 1$. There was, however, a reliable interaction

between stimulus mean and variance, $F(2,\ 30) = 3.4,\ p = 0.04,\ \eta_p^2 = 0.19$, revealing that the difference between the *low mean* and the *high mean* condition was largest for the *medium variance* condition, $M = 0.79\ \mu V$, smallest for the *high variance* condition, $M = -1.05\ \mu V$, and intermediate for the *low variance* condition, $M = -0.54\ \mu V$.

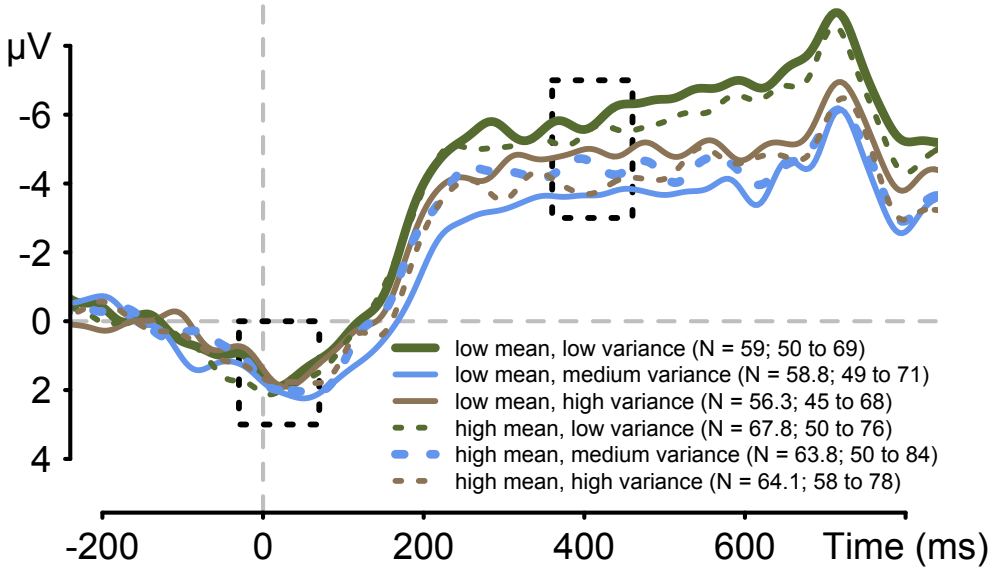

Figure 99: Response-locked ERN (-30 to 70 ms) and Pe (360 to 460 ms) at electrode CPZ as a function of difficulty condition for correct trials only. Response was made at 0 ms. The conditions matched for first-order task performance are highlighted using relatively thicker lines: the *low mean, low variance* condition as a solid, green line and the *high mean, medium variance* condition as a dotted, blue line. The figure legend displays average, minimum, and maximum numbers (N) of trials per participant. ms: millisecond; μV: microvolt.

Given the significant interaction of stimulus variance and location, I conducted another repeated-measures ANOVA with stimulus mean, and stimulus variance as factors for data at electrode CPZ only. There is a reliable effect of stimulus variance, $F(2,\ 30) = 4.7,\ p = 0.02,\ \eta_p^2 = 0.24$, but not of stimulus mean, $F < 1$, and the interaction between the two factors is marginally reliable, $F(2,\ 30) = 2.7,\ p = 0.09,\ \eta_p^2 = 0.15$. Such a main effect of

stimulus variance was also found at PZ. Taken together, Pe amplitude scaled with stimulus variance but not stimulus mean at posterior electrodes. The finding that Pe amplitude tracks differences in stimulus variance but not mean replicates the results of previous experiments and analyses, which have suggested that variance but not mean has an effect on confidence above and beyond its effect on first-order performance.

The second part of the analysis focused on differences in Pe amplitude regarding the matched conditions of medium difficulty. In previous analyses, I have shown that the two medium conditions (*low mean, low variance* and *high mean, medium variance*), which were matched for first-order task performance, show the expected confidence effect: Participants were less confident in the *high mean, medium variance* condition. The key hypothesis of this study was that this effect would also be reflected in the Pe with larger amplitudes for the *high mean, medium variance* condition. Such an effect was indeed found at CPZ and PZ, $ts >= 2.7,\ ps <= 0.02$. The effect was only marginally significant at CZ, $t(15) = 2.1,\ p = 0.05$, and there was no reliable difference at neither FZ, nor FCZ, $ts <= 1.2,\ ps >= 0.10$. Taken together, this suggests that the Pe at the posterior sites reflects the difference in confidence between the two medium conditions matched for primary task performance.

Previous analyses in this thesis provided support for the hypothesis that the Pe, but not the ERN, reflects metacognitive processing. In Experiments 4 and 5, the Pe was even interpreted as a proxy for confidence. The same logic can be applied to the regression analyses reported in Section 5.3.2.2 as well as Appendix C.2: Pe amplitude, but not ERN amplitude should lead to improved prediction of confidence if included in the regression model (Figure 94).

First, model 4a included the ERN as well as the other four predictors

already considered in model 3: accuracy, RT, stimulus mean, and stimulus variance. Those three predictors were all reliable predictors of confidence, $ts >= 2.2,\ ps < 0.05$. The regression weights for RT were only marginally different from zero, $t(15) = 2.0,\ p = 0.06$, as in the previous model with only four predictors. The influence of ERN on confidence was only marginally reliable, $t(15) = 1.9,\ p = 0.054$. The sign of the regression weight for the ERN was negative, which means that the more positive the amplitude of the ERN, the less confident people were. This relationship would have been expected to go in the opposite direction, given than the ERN is a negative component and more negative values were found for less confident judgements. However, this relationship was not as clear in the present experiment, which might have resulted in this finding here. Model 4a explained slightly more variance in the data than model 3, $R^2 = 0.88$, however, this did not justify the additional parameter, as expressed in an increase in BIC scores, $BIC = 1.56$. This difference in BIC values was not reliable, though, $t < 1$. The two models therefore did not differ with regard to their goodness of fit.

Model 4b then added the amplitude of the Pe instead of the ERN. Of all previously analysed predictors, only accuracy and stimulus mean were reliably different from zero, $ts >= 2.8,\ ps <= 0.01$. The influence of stimulus variance on confidence was only marginally significant, $t(15) = 2.1,\ p = 0.05$, whereas the influence of RT was not reliable, $t(15) = 1.6,\ p = 0.12$. Interestingly, Pe amplitude was a reliable predictor of confidence, $t(15) = 3.1,\ p < 0.01$. As can be shown in the upper panel of Figure 94, Pe amplitude was negatively correlated with confidence – the higher the amplitude of the Pe, the less confident people were, as would have been expected. Stimulus variance was a reliable predictor of confidence in model 3, but not in model 4b. This suggests once more that the Pe is sensitive to changes in stimulus variance, leading to a sup-

pression of the stimulus variance predictor in the current model. The model explained absolutely more variance if compared to model 3, $R^2 = 0.90$. The average BIC scores were indeed reduced, $BIC = 0.004$, however, the difference to model 3 was not reliable, $t(15) = 1.1,\ p = 0.29$.

Taken together, this analysis provided additional support for the hypothesis that the Pe is a more reliable correlate of confidence than the ERN. This finding is consistent with previous studies which have established a close link between Pe amplitude and error awareness (Overbeek et al., 2005; Steinhauser & Yeung, 2010; Endrass et al., 2012).

5.3.3 Discussion

The present study focused on neural correlates of the already-established effect of variance on judgements of confidence. In the behavioural data, several key findings from previous experiments were replicated: mean and variance had an effect on response speed, accuracy, as well as on confidence ratings. Highly variable stimuli furthermore resulted in more selective processing of the colour stimuli. Critically, two conditions matched for first-order performance showed the previously-shown difference in confidence: Higher variance (lower evidence reliability) was correlated with lower confidence, statistically reliable at least for error trials.

The main goal of this study was to establish the impact of evidence mean and variance on neurophysiological correlates of confidence – the ERN and Pe. The analyses showed that changes in stimulus variance, but not in stimulus mean, were reflected in neurophysiological recordings: The matched condition with the higher variance showed a larger Pe amplitude. This echoes the earlier proposal that signal reliability reflects the 'native language' of confidence, having an effect on confidence ratings over and above its effect on

first-order task performance. Second, only the Pe but not the ERN was sensitive to these changes in stimulus variance and confidence, supporting again the notion that the error positivity is a good candidate index of error awareness (Overbeek et al., 2005; Steinhauser & Yeung, 2010; Endrass et al., 2012). Furthermore, regression model analyses suggested that the Pe amplitude shares variance with the factor manipulating evidence reliability – colour variance. This was present in a seemingly suppressed effect of stimulus variance on confidence once Pe amplitude was introduced into the model as a new predictor.

5.4 General discussion

The key question addressed in Experiments 6 to 8 was whether different cues affect how confident people judge their decisions. In all three experiments I found that both stimulus mean and variance – both of which I manipulated independently in a colour classification task – affect confidence. Both of these difficulty manipulations affected first-order performance as shown previously by De Gardelle and Summerfield (2011): Participants were faster and more accurate when evidence reliability was high (*low variance* conditions), as well as when evidence strength was high (*high mean* conditions). With regard to second-order effects, participants were more confident when the mean of the stimulus clearly favoured one of the two colours, and when stimulus variability was low. Interestingly, however, stimulus variability affected confidence over and above its effect on first-order performance: For two conditions matched for task difficulty, participants were consistently less confident for trials with high stimulus variability. A detailed analysis of metacognitive bias for Experiment 6 suggested that this effect was caused by the fact that participants' confidence was driven by changes in stimulus variance but not stimulus

mean. In other words, they were overconfident when the stimulus mean was low, a classical hard-easy effect; but there was no such hard-easy effect for stimulus variance.

The main question of whether different cues contribute to decision confidence was addressed by comparing different individual regression models predicting confidence. The results from this approach first of all suggested that accuracy is clearly a very strong predictor of confidence, which does not come as a surprise given that confidence can be regarded as a subjectively estimated probability of being correct in a choice. RT also predicted confidence, with faster RTs being associated with more confident judgements. This is congruent with what the time heuristic would have predicted (Audley, 1960; Moreno-Bote, 2010; Zylberberg et al., 2012; Kiani et al., 2014). The key finding, however, was that stimulus variance also predicted confidence and did so over and above the effect that stimulus variance had on accuracy and RT, given that the regression models only accounted for independent variance proportions. This was not the case for stimulus mean, however. These findings were remarkably stable and replicated for all three studies. I therefore conclude that difficulty due to stimulus variance serve as an important cue to people's confidence.

Taken together, the findings reported in this chapter suggest that confidence judgements are affected by a range of different cues. These cues can be roughly categorised into two different types or models (Koriat, 1993; Schwartz & Metcalfe, 1994). The first class of models assumes that participants' confidence is affected by the same information as the decision itself or some property of it (direct access), for example signal detection theory (SDT; Higham et al., 2009). The second class of models, on the other hand, assumes that participants are inferring their confidence from an external cue (heuristics-based

models), for example from their RTs, judging decisions as high confident if they were made quickly. In the present chapter it is argued that a combination of cues, both direct-access (accuracy and stimulus variance) and heuristics-based cues (RT), affects participants' confidence. This multi-cue model of confidence with different internal sources and signals that contribute to a final confidence judgement is similar to what has previously been suggested in the literature on metamemory (see Nelson et al., 1984; Koriat & Levy-Sadot, 2001, for a similar suggestion regarding feeling-of-knowing judgements; also see p. 71, Dunlosky & Metcalfe, 2009, for a review of these approaches).

The findings reported here suggest that stimulus variance plays a key role in the formation of metacognitive judgements. I have offered the interpretation of this effect that stimulus variance (or evidence reliability) could be regarded as first-order uncertainty, that is information regarding the precision of a stored memory, a percept or another object-level representation. There have indeed been findings supporting the notion that mental representations are stored and accessed as a probability distribution of activations rather than a point estimate (Beck et al., 2008; Fiser et al., 2010; Ma et al., 2006). This means that the brain can 'read out' the uncertainty estimate associated with every representation (Bach & Dolan, 2012), and then easily transform this system-I metacognitive representation into verbal, system-II metacognitive judgements, to use the classifications proposed by Shea et al. (2014). In other words, confidence is more influenced by stimulus variance than stimulus mean because the former is the 'native language' of metacognition, whereas inferring difficulty from signal strength means that this strength first has to be compared to an internal reference point, which could presumably be noisily represented as well.

Experiment 7 addressed the question whether a clinical population,

which shows differences with regard to metacognitive processing, would show a different weighting of some of these cues. Patients diagnosed with clinical depression constitute such a population for which metacognitive judgements differ compared to those of a healthy control group. Acute tryptophan depletion (ATD) was used here to simulate reduced levels of serotonin in healthy adult participants, similar to what has been observed in clinical depression (Heninger et al., 1984). In the literature, underconfidence (Hancock, 1996; Fu et al., 2005; Szu-Ting Fu et al., 2012; Dunlosky & Metcalfe, 2009; Wood et al., 1998), but also overconfidence (Dunning & Story, 1991) has been reported for participants with depression. The findings from Experiment 7 mirror the latter pattern, showing that tryptophan-depleted participants are on average more confident in their responses compared to a control group whose tryptophan levels had been restored to normal prior to the experiment. My findings therefore speak against the "depressive realism hypothesis" (Moore & Fresco, 2012), as well as the "selective processing hypothesis" (Hancock, 1996), which both suggest that participants with depression exhibit better calibration or less overconfidence, respectively.

I would furthermore argue that the experiment reported here overcomes some of the problems and difficulties that can be found for those previous studies on confidence and depression. First, in my study I collected trial-by-trial retrospective confidence judgements rather than a single, overall estimate regarding the percentage correct of all trials completed by the participant, like used by many other studies (Hancock, 1996; Fu et al., 2005; Szu-Ting Fu et al., 2012; Wood et al., 1998). This allowed me to test for condition differences, as conditions occurred intermixed within blocks. Second, studies focusing on confidence in depression have used many different types of decision-making paradigms, like general knowledge tasks (Hancock, 1996), or

prediction tasks for future, personal events (Dunning & Story, 1991). Uncertainty in these types of tasks is often believed to be external (Juslin & Olsson, 1997), making it difficult for the experimenter to control difficulty. Here, I have chosen a perceptual decision-making task for more precise experimental control over these factors. Third, in my experiment, I found group differences with regard to both first- and second-order processing. It was therefore crucial to analyse confidence data using SDT models, to carefully 'disentangle' first- and second-order processing, and to find out whether differences in confidence were driven by differences in metacognitive efficiency or bias. Past studies have not resorted to such models, for example Dunning and Story (1991) found differences in participants' accuracy with regard to judging future events, but did not analyse confidence data with a model that would take into account these differences (Maniscalco & Lau, 2012).

The key question – regarding whether those cues affecting confidence are weighted differently for the tryptophan-depleted participants – was addressed using the same model comparison approach comparing different predictors of confidence. However, no difference was found between the two groups, suggesting that, at least for now, we cannot reject the hypothesis that in both groups these cues contribute to confidence in similar ways. Future research should therefore include other confidence cues to explore the precise mechanisms which caused the difference in metacognitive bias. One such cue would be familiarity of the to-be-judged material. In addition to being a well-characterised influence on metamemory judgements (Reder & Ritter, 1992), familiarity has recently been highlighted as a reliable cue of decision confidence (De Martino et al., 2013).

Several further limitations regarding the ATD study ought to be mentioned; first, this type of design compares two experimental groups without

testing a control group: The tryptophan-depleted participants are compared only with a group whose plasma tryptophan levels were restored prior to the experiment. In the future, untreated control groups should be included as well to be able to directly compare the findings to for example Experiment 6. Another issue comes with the fact that the validity of ATD as a measure to lower serotonin levels is still debated (Van Donkelaar et al., 2011; but see also Crockett et al., 2012; Young, 2013). It will not be possible to resolve this latter issue here and tryptophan depletion should therefore only be considered as a model of some components of depression.

Experiment 8 tested whether changes in the processing of stimulus mean and variance and their influence on confidence are reflected in error-related ERPs. The Pe, which has previously been argued to reflect error awareness (Steinhauser & Yeung, 2012; Nieuwenhuis et al., 2001), also reflected changes in stimulus variability, but not in stimulus mean. Moreover, if only the matched conditions were considered – at least conditions selected post-hoc for matched performance – there was a reliable difference in the Pe, but not the ERN. This difference followed the confidence effect: Participants were more confident in the condition which was difficult due to a *low mean*, which was also reflected in a smaller Pe amplitude, compared to the condition that was difficult due to *high variance*, which was associated with less confident judgements and a larger Pe amplitude.

A third set of analyses linked to the question posed in Experiments 4 and 5, as to whether Pe amplitude serves as a proxy to confidence. Indeed, Pe amplitude was a reliable predictor of confidence if added as predictor to the above-discussed regression models. A model that included Pe as a predictor explained more variance in confidence data, even if the difference in the goodness-of-fit values was not reliable. This was not found for ERN amp-

litude, which was not a reliable predictor of confidence and led to an even worse goodness of fit than a model without ERN as a predictor. Taken together these results replicated that the Pe, but not the ERN, reflected changes in confidence.

Several limitations, which applied to all of these studies should furthermore be addressed here. First of all, the values of stimulus mean and variance were sampled from stimulus distributions within a narrow range, therefore forming discrete, factorial conditions instead of testing the influence of stimulus mean and variance as continuous variables. The main analyses in this chapter were regression analysis, however, which would have been more powerful in detecting effects if continuous variables had been used (Bell, 1992). Future studies should therefore extend this paradigm by using continuously varying independent variables.

Using continuous instead of discrete independent variable could also have affected the above-discussed differences to the recent study by Zylberberg et al. (2014), who found that greater variability in a stimulus lead to higher confidence, the exact opposite to what was found here. The authors suggest that one reason for this finding was that participants are unable to adjust their confidence criteria to the changes in reliability of evidence (stimulus variance). Once could argue that this adjustment was easier for participants in the current experiment, because of the discrete nature of the conditions.

Taken together, the findings from Experiments 6 to 8 reported in this chapter support the hypothesis that confidence is derived from a range of different cues. These cues can rely both on direct, privileged access to the decision process itself, but also on simple heuristics. Future studies could focus on the question whether ways could be found in which participants are influenced in which cues contribute to their confidence judgement. Arguably,

some cues are more valid than others (Gigerenzer et al., 1991). It would be worthwhile to find a method to train participants to 'listen' to the most valid cues for confidence, depending of course on contextual factors which influence the validity of those cues. Such an approach would ultimately result in people becoming more metacognitively accurate, which could in turn mean gains with regard to cognitive control functions. The idea that the cues considered when making a metacognitive judgement links closely to the findings reported by Koriat, Sheffer and Ma'ayan (2002). The authors observed that participants became more underconfident with more practice and explained this effect as a shift in confidence cues, which affected both resolution and calibration.

Chapter 6

General discussion

The research reported in this thesis was focused on metacognition in decision making. I studied methodological issues arising when studying such introspective judgements about the correctness of a choice, as well as links between confidence and error detection – two types of metacognitive judgements that have been studied largely separately over the past decades. Other experiments were focused on the uses that metacognition serves in cognitive control, and the question of how metacognitive judgements are formed. All those questions were explored using behavioural, neurophysiological, and pharmacological methods, together with computational modelling. In the present chapter, findings from all previous chapters will be summarised and theoretical implications will be discussed. Moreover, limitations regarding those findings will be flagged and future avenues to explore will be suggested.

6.1 Summary of research

6.1.1 Methodological issues

The first three experiments reported in this thesis were focused on the question of whether metacognition in decision making is a stable phenomenon or susceptible to influences from fine methodological differences across studies. It was first established that ratings of confidence are indeed not easily disturbed and that measuring them is straightforward and easy. For example, which scale was used did not affect the ratings (Experiment 1), as reflected in there being no reliable difference in metacognitive performance between the two scales (binary error detection versus 6-point, graded confidence scale). It was therefore concluded that the higher resolution obtained with a more fine-grained confidence scale does not come at a cost of additional processing load. This also means participants were just as fast using such a graded scale, compared to judging their performance on a binary scale. Their use of the confidence scales can be described as very intuitive, meaning that hardly any training was required before participants reached impressively high resolution in their metacognitive judgements.

Another reason why I concluded that confidence judgements are stable is that they were not found to be affected by the time point at which participants were instructed to rate their performance (Experiment 2) – whether this judgement took place immediately after the to-be-judged decision or 1.5 seconds later. It should therefore not be expected that a failure to replicate confidence findings stems from such timing specifics. Consistent with the post-decision processing hypothesis, however, it has been observed that for extremely short delays, participants will slow down in their confidence responses,

presumably balancing out any time restrictions imposed by the experimental design. This strategy might have caused the stability of confidence performance observed for a range of different time windows.

Moreover, findings from Experiment 3 led to the conclusion that judging one's confidence in addition to the primary response task does not lead to performance impairments in the primary task. Making metacognitive judgements therefore does not impose substantive additional cognitive burden, suggesting that the cognitive processes underlying the primary task and confidence judgements are highly overlapping. Instead, however, I observed participants to adopt a more accuracy-focused response strategy when they had to judge their decision confidence, meaning they were both slower and more accurate with regard to their primary task performance. Presumably, this was the case because focusing on one's accuracy makes people more aware of their errors and therefore more likely to try to avoid them.

Taken together, the findings reported in Chapter 2 imply that metacognitive judgements are stable, easy to measure, do not impose additional cognitive processing costs, nor do they require excessive training and familiarisation with the confidence scales. A perceptual decision-making paradigm was furthermore introduced in this chapter, which provided a suitable context in which metacognition can be measured with precise control of task difficulty. Participants were found to have impressive metacognitive insight into their own decisions with this paradigm and it was therefore used for the experiments in the majority of the chapters in this thesis.

6.1.2 Confidence and error detection

The key question in Chapter 3 was whether two types of metacognitive judgements – error detection and decision confidence – can be linked, given their

conceptual and methodological similarities. EXPERIMENT 1 from the previous chapter had already shown that participants' metacognitive judgements are not affected by the type of scale (binary or graded) with which they were measured. With this chapter, I followed up on the idea that binary error judgements and confidence are two sides of the same coin, showing that there is a shared neural basis of the two. More specifically, using a cross-classification approach, I showed that subjectively-rated confidence varied with a well-characterised EEG component of error awareness, the Pe. This effect was present even on the single-trial level, as shown using multivariate pattern classification techniques.

There is little compatibility between current theories of confidence and error detection (Yeung & Summerfield, 2012, 2014). However, upon closer inspection, it becomes obvious that the theoretical assumptions proposed by each line of research are by no means mutually exclusive, and can instead be understood as complementing each other and potentially leading to an improved theory of metacognition. For instance, any such integrative theory must be able to explain graded confidence judgements (Vickers & Packer, 1982) – similar to the balance-of-evidence mechanism assumed in many theories of confidence – and should at the same time also explain why participants sometimes state with certainty that their just-given response was incorrect – similar to the mismatch assumptions in many prominent theories of error detection (Falkenstein et al., 1991; Gehring et al., 1993). Those theories often assume that error detection happens after the decision was formed; that is, integrating newly incoming information even after a choice has been formed (Rabbitt et al., 1978; Yeung & Summerfield, 2012, 2014; Baranski & Petrusic, 1998).

Following on from my empirical evidence of links between error detection and confidence judgements, I proposed such a combined model of meta-

cognitive judgements. My single-route model of metacognition in decision making reflects all previously discussed findings: post-decision processing, changes of mind and high-confidence errors, and graded as well as binary metacognitive judgements relying on the same continuous metacognitive signal. At the heart of my race model was a balance-of-evidence mechanism, similar to the previously suggested expanded race model proposed by Van Zandt and Maldonado-Molina (2004): Metacognitive judgements are derived from post-decisional balance of evidence. The model was able to simulate the key data patterns from my previous experiments as well as recent findings presented as a critical challenge to the idea of single-route models of metacognition (Charles et al., 2013), leading me to suggest that it might be a suitable starting point for a formal computational theory of metacognition in decision making aimed at combining theories of error detection and confidence judgements. Taken together, the findings from Experiment 4 and the computational model suggest that it is worthwhile linking two lines of research that have largely been studied separately – error detection and decision confidence – given that there is evidence that both judgements rely on the same internal mechanisms.

6.1.3 Uses of metacognition

The previous chapters were largely focused on how decision making affects metacognition. Chapter 4 instead asked how confidence judgements – once formed – affect future decisions or, as in this case, attention to feedback. Metacognitive evaluations are known to have impressive effects in cognitive control (Fernandez-Duque et al., 2000). For example, studies on error monitoring have shown that participants slow down after they detect an error (Laming, 1979; Dutilh et al., 2012; Notebaert et al., 2009), presumably to prevent mistakes on subsequent trials. Moreover, in the metamemory literature findings have been

reported that suggest that participants use confidence as a cue in guiding the allocation of study time, spending more time on items they are less confident about (Nelson & Leonesio, 1988). Confidence also plays an important role in social interactions (Bahrami et al., 2010; Shea et al., 2014), serving as a cue as to how much weight should be assigned to each person's opinion in a joint decision-making process. While these are all examples of how confidence and other metacognitive signals affect cognitive processing in various domains, the question remains as to whether decision confidence affects decision making. This question was therefore addressed in this chapter.

I hypothesised that participants would pay more attention to feedback when they are in low certainty states, knowing that they guessed the answer. I expected them to ignore feedback whenever they are certain about the outcome of a decision, whether it is positive (sure correct) or negative (sure error). To test this, Experiment 5 used the same multivariate pattern classification techniques as Experiment 4 to infer single-trial confidence without requiring overt metacognitive judgements from the participants.

However, the results did not reflect such an effect of confidence. Instead, there was an effect of feedback valence, as previously found in the literature (Yeung & Sanfey, 2004; Holroyd & Coles, 2002). Presumably, this happened because negative feedback is a salient stimulus that draws attention bottom-up, which however does not necessarily speak against an additional, top-down effect of certainty for which the paradigm presumably lacked the necessary level of power to detect it. Future studies are therefore necessary to test the certainty hypothesis with an improved, more sensitive paradigm and an improved classifier, for which several suggestions have been made. Taken together, how confidence affects attention to feedback remains an open question which I was unable to answer in this thesis. I am looking forward to

following up on this question in the future with the improvements discussed above to the paradigm.

6.1.4 Formation of metacognitive judgements

The experiments in Chapter 5 aimed to shed further light on how metacognitive judgements are formed internally. The hypothesis tested in this chapter was that multiple cues contribute to how confident participants report a choice to be, similar to the hypothesis posed in the context of metamemory judgements (Koriat, 1997; Leonesio & Nelson, 1990). In a recent collaboration, we suggested that a range of such cues exists on system-I level, which is characterised by automatic, fast and effortless processing (Shea et al., 2014), as opposed to system-II processing, which is thought to be slow, effortful, and conscious. According to this view, those system-I cues can be used internally for purposes of cognitive control, or they can be combined to form a system-II metacognitive representation, which can then be expressed on a confidence scale communicated verbally to optimally interact with other agents.

The results of Experiments 6 to 8 indeed suggested that different system-I cues had an influence on (system-II) confidence judgements, such as accuracy of the choice, RT, but also stimulus characteristics. These cues can be both inherent in the decision, as well as external in the form of heuristics, such as RT. Interestingly, the variability of a stimulus had an influence on confidence over and above its effect on first-order performance, which led me to suggest that such stimulus variability can more naturally be translated into confidence. This makes sense if we consider that according to a Bayesian perspective on cognitive processing, every mental representation should be associated with an internal reliability estimate, or spread of the evidence distribution (Ma et al., 2006). Such variability estimates have previously been identified in the brain

in the form of summary statistics (Michael et al., 2015; Alvarez, 2011; Pollard, 1984).

Moreover, findings from Experiment 7 suggested that tryptophan-depleted participants are less likely to be affected in their decisions by stimulus variance, but at the same time it seemed like they were more affected by stimulus variance in their confidence judgements. However, this effect remained very weak and needs to be further tested in future experiments. Overall, tryptophan-depleted participants were found to be more confident, independent of whether or not they were correct.

The extended effect of stimulus variance on confidence was also observed in an EEG study (Experiment 8), which furthermore revealed that the influence of stimulus variability was also reflected in error-related brain activity and that such brain activity also correlated with a participant's confidence. The results showed that the Pe and not the ERN is more likely a correlate of error awareness. Taken together, the results from this chapter highlighted that multiple different cues contribute to metacognitive judgements, also suggesting that stimulus variance plays a prominent role in the formation of metacognitive judgements given that it is already in a format that can be described as the 'native language' of confidence.

6.2 Why study metacognition in decision making?

The findings reported in this thesis highlight several reasons why metacognition in decision making plays an important role in human cognition and why it is therefore a worthwhile object of study. People are capable of reporting metacognitive judgements with impressively high accuracy. Presumably, intro-

spection is precise because it is driven by cues that are directly or indirectly linked to the decision or that drive decision accuracy (Schwartz & Díaz, 2014), such as stimulus variance. Metacognitive performance might be impaired in the case of some clinical groups, however, such as people diagnosed with schizophrenia (David, Bedford, Wiffen & Gilleen, 2012; Charles, 2013), OCD (Tolin et al., 2001; Ben Shachar et al., 2013), or depression (Dunlosky & Metcalfe, 2009). On a related note, it would be worthwhile to gain more understanding of how metacognitive insight develops over the lifespan (Weil et al., 2013; Palmer, David & Fleming, 2014), a question closely linked to issues in research on theory of mind (Meltzoff & Gopnik, 2013; Happé, Winner & Brownell, 1998). To conclude, various people might have impaired metacognition that contributes to the difficulties they face in everyday life, so understanding the nature of those metacognitive deficits might be helpful to them. This is therefore the first reason why I believe metacognition is a worthwhile object of study.

A second reason why metacognition in decision making is worthwhile studying is its already discussed crucial role in social interactions (Shea et al., 2014; Bahrami et al., 2010) and cognitive control (Yeung & Summerfield, 2012, 2014; Fernandez-Duque et al., 2000). In this thesis, I have assumed that metacognitive judgements are omnipresent, which means every mental process or representation is accompanied by system-I cues such as stimulus variance or evidence reliability (Beck et al., 2008; Fiser et al., 2010; Ma et al., 2006) that can be utilised to optimise behaviour if needed. Often, people have to learn without external feedback. In such situations, metacognition can serve as an internal feedback system, therefore enabling successful learning. Future research should therefore ultimately be focused on how these interactions and instances of cognitive control can be improved through the study of their underlying metacognitive processes, potentially even with the

goal to develop training programmes that helps people from the affected risk populations to develop more precise metacognitive judgements. Arguably, this could be achieved using a mindfulness training approach, which is known to improve interoceptive attention (Farb, Segal & Anderson, 2013) and metacognitive awareness (Teasdale et al., 2002; Jankowski & Holas, 2014). Such mindfulness interventions have been shown to reduce ruminative thought in clinical depression (Ramel, Goldin, Carmona & McQuaid, 2004) and to decrease mind wandering in the general population (Mrazek, Franklin, Phillips, Baird & Schooler, 2013).

A third reason in favour of studying metacognition is that it constitutes an interesting tool for studying other cognitive processes, such as decision making. In the case of value-based decision making, for example, a recent study by De Martino et al. (2013) has suggested that confidence reflects the difference in value between the chosen and an unchosen choice alternative, similar to a balance-of-evidence mechanism. Other cases in which confidence sheds light on underlying decision processes is post-decision processing as well as choice reversals (Van Zandt & Maldonado-Molina, 2004). Moreover, Zylberberg et al. (2012) have argued that if confidence depends on a balance-of-evidence mechanism – as suggested by their model comparison approach – then underlying the first-order decision must also be a race model, rather than integrated accumulation of balance of evidence as assumed by diffusion-model approaches. In other words, the nature of confidence in decisions constrains possible theories of how those decisions are made. Metacognition can therefore be regarded as a window into cognitive control and decision making and not just as noise in evidence accumulation. Therefore, accounts of confidence ought to be incorporated into theories of decision making to provide explanations for this dependent variable in addition to the usually studied accuracy and RT.

6.3 Future directions

In this thesis, decision confidence was studied only in the context of perceptual decision-making paradigms. Limiting my experiments to only this type of task was a deliberate decision, as discussed in Chapter 1: Perceptual choices are largely driven by external stimulus properties and therefore provide precise experimental control over the noisy evidence on which decisions are based (Gold & Shadlen, 2007). Real-world decision making, on the other hand, can be far more complex, but within this complexity lie several challenging, but useful extensions of existing work, two of which I would like to discuss here.

First, many real-world decisions are based on internal value representations: The decision maker is faced with various objects, each associated with a subjective, internal value. The decision maker's choice should reflect underlying preferences. Confidence in such neuroeconomical choices has – to the best of my knowledge – so far only been studied by De Martino et al. (2013), who suggested that confidence is a function of a difference in value – much like classical balance-of-evidence models (Vickers & Packer, 1982). One potential challenge for the study of such value-based decision confidence is the fact that value is difficult to measure objectively and often based on a history of life-long exposure to the items in question. However, even if exposure is largely controlled for in laboratory experiments – for example, through multi-armed bandit tasks in which participants learn the value of different lotteries over time – there exists a second challenge for the study of value-based confidence: Representations of value can be noisy and can develop over time. For instance, the more familiar people are with a lottery, and the less variable the lottery's outcomes have been over time, the less uncertainty should be associated with the value of this lottery. Importantly, such uncertainty can exist on a purely

first-order basis (Bach & Dolan, 2012), meaning that it is inherent in how value representations are stored (Ma et al., 2006), and is therefore qualitatively different from metacognitive, and therefore re-evaluative confidence, as studied in this thesis (Shea et al., 2014). However, as suggested in Chapter 5, increases in stimulus uncertainty lead to decreases in confidence. This evidence was found in the context of a classification task regarding the colour of the stimulus. This is also shown in the left panel of Figure 100: The distribution represents an internal estimate of the relevant stimulus dimension, in this case colour value. This estimate is associated with a certain degree of uncertainty, shown here as the spread of the distribution. This first-order uncertainty has an influence on confidence. It might be worthwhile to extend this approach using a comparative decision task, which would thus require participants to integrate first-order uncertainty from different sources of evidence. This extension is shown in the right panel of Figure 100: Two objects or stimuli are compared on an internal task-relevant dimension, such as value. The estimated value of each of these items is represented with a different precision – expressed in the spread of the respective distribution. The question is whether and how these uncertainties are integrated regarding their effect on decision confidence. Such an approach could shed light on how first-order uncertainty is 'translated' into second-order uncertainty, or confidence, therefore linking the lines of research on uncertainty and metacognition.

A second way in which real-world decisions are more complex than the perceptual decisions studied in this thesis is that they are usually hierarchical (Botvinick, 2008): They can be regarded as being organised to serve different subgoals, whilst ultimately aimed towards reaching certain higher-order goals. For instance, when humans or animals forage for food, they face the constant lower-level decisions of whether they should stay in a given situation or try

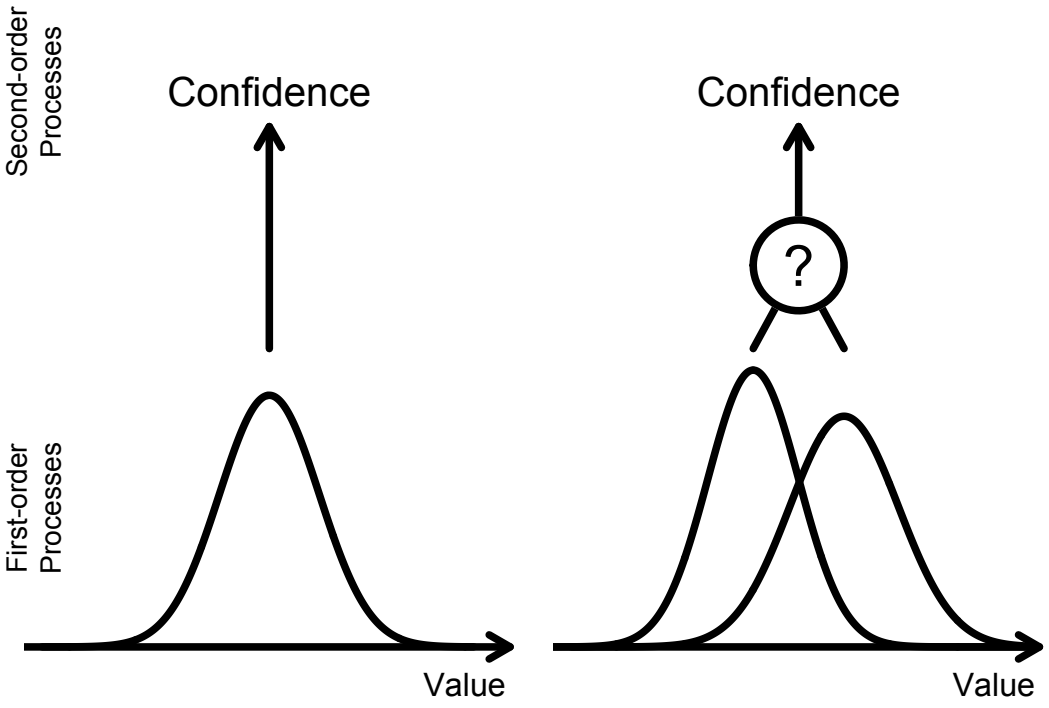

Figure 100: Influence of first-order uncertainty, reflected in the spreads of the distributions, on decision confidence. The left panel shows a classification task, similar to the one reported in Chapter 5. The right panel shows a comparison of two items on a task-relevant dimension, such as value.

something new at the risk of worsening the current reward level (Kolling, Behrens, Mars & Rushworth, 2012) – all while aiming to optimise food intake or survival as a higher-level goal. This tradeoff is known as an exploration-exploitation dilemma. Information seeking in the general sense of uncertainty reduction has been suggested as a mechanism to manage this tradeoff: People have been observed to choose to explore especially when the to-be-explored choice option would result in reducing the uncertainty (Frank et al., 2009; Badre et al., 2012). These findings have usually been interpreted as evidence of how first-order uncertainty affects decision making. The question arises as to whether decision confidence also serves as such a cue to the arbitration between exploration and exploitation. For instance, one could expect that participants are more likely to explore after a run of low-confidence trials, which might be interpreted as a signal that there is insufficient evidence regarding the value difference between the choice option and therefore exploration is necessary. This would constitute another example of how metacognition is used to enhance cognitive control, as studied in Chapter 4.

6.4 Conclusion

The research discussed in this thesis has investigated metacognition in decision making. Across all the experiments reported here it was shown that people have accurate insight into the accuracy of their perceptual decisions. In this context, experiments have shown that metacognitive judgements are stable, straightforward to measure, and do not lead to additional processing costs. Moreover, a range of different cues on the basis of which confidence judgements are formed have been identified. Critically, the findings presented here indicate that amongst those cues, stimulus variance is a very promin-

ent one, influencing confidence more strongly and more directly compared to signal strength. Furthermore, error detection and decision confidence – two metacognitive functions that have commonly been treated separately – have been linked here, suggesting that theoretical insights into each process could potentially be transferred to and constrain theorising about the other. Finally, how confidence affects feedback processing remains and open question and will have to be investigated into in the future together with the question how precisely metacognitive judgements are used internally to enable cognitive control and guide decision making in more complex real-world scenarios.

References

Ahveninen, J., Kähkönen, S., Pennanen, S., Liesivuori, J., Ilmoniemi, R. J. & Jääskeläinen, I. P. (2002). Tryptophan depletion effects on EEG and MEG responses suggest serotonergic modulation of auditory involuntary attention in humans. *NeuroImage*, *16*(4), 1052–1061. doi: 10.1006/nimg.2002.1142

Alain, C., McNeely, H. E., He, Y., Christensen, B. K. & West, R. (2002). Neurophysiological evidence of error-monitoring deficits in patients with schizophrenia. *Cerebral Cortex*, *12*(8), 840–846. doi: 10.1093/cercor/12.8.840

Alexander, W. H. & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. doi: 10.1038/nn.2921

Allport, D. A., Styles, E. A. & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance 15: Conscious and nonconscious information processing. Attention and performance series.* (pp. 421–452). Cambridge, MA: The MIT Press.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. doi: 10.1016/j.tics.2011.01.003

Asberg, M., Thoren, P., Traskman, L., Bertilsson, L. & Ringberger, V. (2003). "Serotonin depression" – a biochemical subgroup within the

affective disorders? *Science*, *191*(4226), 478–480. doi: 10.1126/science.1246632

Audley, R. J. (1960). A stochastic model for individual choice behavior. *Psychological Review*, *67*(1), 1–15. doi: 10.1037/h0046438

Bach, D. R. & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586. doi: 10.1038/nrn3289

Başar-Eroglu, C., Başar, E., Demiralp, T. & Schürmann, M. (1992). P300-response: Possible psychophysiological correlates in delta and theta frequency channels. A review. *International Journal of Psychophysiology*, *13*(2), 161–179. doi: 10.1016/0167-8760(92)90055-G

Badre, D., Doll, B. B., Long, N. M. & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*, *73*(3), 595–607. doi: 10.1016/j.neuron.2011.12.025

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G. & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1350–1365. doi: 10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G. & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085. doi: 10.1126/science.1185718

Baranski, J. V. & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. doi: 10.2307/1423410

Baranski, J. V. & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 929–945. doi: 10.1037/0096-1523.24.3.929

Baranski, J. V. & Petrusic, W. M. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, *55*(3), 195–206. doi: 10.1037/h0087366

Barrett, A. B., Dienes, Z. & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. doi: 10.1037/a0033268

Barrouillet, P., Bernardin, S. & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. doi: 10.1037/0096-3445.133.1.83

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., ... Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142–1152. doi: 10.1016/j.neuron.2008.09.021

Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. doi: 10.1038/nn1954

Bell, R. W. (1992). Continuously distributed random variables in factorial designs. *Experimental Aging Research*, *18*(1-2), 47–50. doi: 10.1080/03610739208253910

Ben Shachar, A., Lazarov, A., Goldsmith, M., Moran, R. & Dar, R. (2013). Exploring metacognitive components of confidence and control in individuals with obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(2), 255–261. doi: 10.1016/j.jbtep.2012.11.007

Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. doi: 10.1037/0033-295X.113.4.700

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U. & Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, *33*(1), 10–16. doi: 10.1016/j.tins.2009.09.002

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208. doi: 10.1016/j.tics.2008.02.009

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. doi: 10.1037/0033-295X.108.3.624

Botvinick, M. M., Cohen, J. D. & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. doi: 10.1016/j.tics.2004.10.003

Brainard, D. H. (1997). The Psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. doi: 10.1163/156856897X00357

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brown, G. D. A., Neath, I. & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. doi: 10.1037/0033-295X.114.3.539

Brunia, C. (1988). Movement and stimulus preceding negativity. *Biological Psychology*, *26*(1-3), 165–178. doi: 10.1016/0301-0511(88)90018-X

Brunia, C., Hackley, S., Van Boxtel, G., Kotani, Y. & Ohgami, Y. (2011). Waiting to perceive: Reward or punishment? *Clinical Neurophysiology*, *122*(5), 858–868. doi: 10.1016/j.clinph.2010.12.039

Bruyer, R. & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica*, *51*(1), 5–13. doi: 10.5334/pb-51-1-5

Budescu, D. V. & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. Busemeyer, R. Hastie & D. L. Medin (Eds.), *Psychology of learning and motivation* (pp. 275–318). Academic Press.

Buratti, S. & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive Processing*, *13*(3), 243–253. doi: 10.1007/s10339-012-0440-5

Burgess, N. & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language*, *55*(4), 627–652. doi: 10.1016/j.jml.2006.08.005

Busey, T. A., Tunnicliff, J., Loftus, G. R. & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*(1), 26–48. doi: 10.3758/BF03210724

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. doi: 10.3102/00346543065003245

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language*, *23*(1), 58–89. doi: 10.1111/j.1468-0017.2007.00329.x

Castel, A. D., McCabe, D. P. & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review*, *14*(1), 107–111. doi: 10.3758/BF03194036

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X. & Frank, M. J. (2012). Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cerebral Cortex*, *22*(11), 2575–2586. doi: 10.1093/cercor/bhr332

Cavanagh, J. F. & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. doi: 10.1016/j.tics.2014.04.012

Cerný, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, *45*(1), 41–51. doi: 10.1007/BF00940812

Charles, L. (2013). *Neural mechanisms of conscious and non-conscions meta-decisions* (Doctoral dissertation). L'Université Pierre et Marie Curie.

Charles, L., King, J.-R. & Dehaene, S. (2014). Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. *Journal of Neuroscience*, *34*(4), 1158–1170. doi: 10.1523/JNEUROSCI.2465-13.2014

Charles, L., Van Opstal, F., Marti, S. & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94. doi: 10.1016/j.neuroimage.2013.01.054

Chua, E. F., Pergolizzi, D. & Weintraub, R. R. (2014). The cognitive neuroscience of metamemory monitoring: Understanding metamemory processes, subjective levels expressed, and metacognitive accuracy. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 267–291). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-45190-4_12

Chwilla, D. J. & Brunia, C. H. (1991). Event-related potentials to different feedback stimuli. *Psychophysiology*, *28*(2), 123–132. doi: 10.1111/j.1469-8986.1991.tb00400.x

cognition. (n.d.). Retrieved from `http://www.oxforddictionaries.com/definition/english/cognition`

Cohen, M. X. & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *The Journal of Neuroscience*, *27*(2), 371–378. doi: 10.1523/JNEUROSCI.4421-06.2007

Coles, M. G., Scheffers, M. K. & Holroyd, C. B. (2001). Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. *Biological Psychology*, *56*(3), 173–189. doi: 10.1016/S0301-0511(01)00076-X

Compumedics Neuroscan. (2003). *SCAN.* Charlotte, NC, USA.

Compumedics Neuroscan. (2007). *SCAN.* Charlotte, NC, USA.

Crockett, M. J., Clark, L., Roiser, J. P., Robinson, O. J., Cools, R., Chase, H. W., ... Robbins, T. W. (2012). Converging evidence for central 5-HT effects in acute tryptophan depletion. *Molecular Psychiatry*, *17*(2), 121–123. doi: 10.1038/mp.2011.106

Damen, E. & Brunia, C. (1987). Changes in heart rate and slow brain potentials related to motor preparation and stimulus anticipation in a time estimation task. *Psychophysiology*, *24*(6), 700–713. doi: 10.1111/j.1469-8986.1987.tb00353.x

Danielmeier, C. & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology*, *2*(September), 233. doi: 10.3389/fpsyg.2011.00233

Davelaar, E. J. (2009). Conflict-monitoring and (meta)cognitive control. In J. Mayor, N. Ruh & K. Plunkett (Eds.), *Connectionist models of behaviour and cognition ii. proceedings of the 11th neural computation and psychology workshop* (pp. 91–102). Singapore: WorldScientific.

David, A. S., Bedford, N., Wiffen, B. & Gilleen, J. (2012). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1379–1390. doi: 10.1098/rstb.2012.0002

Davies, P. L., Segalowitz, S. J., Dywan, J. & Pailing, P. E. (2001). Error-negativity and positivity as they relate to other ERP indices of attentional control and stimulus processing. *Biological Psychology*, *56*(3), 191–206.

De Gardelle, V. & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, *25*(6), 1286–1288. doi: 10.1177/0956797614528956

De Gardelle, V. & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(32), 13341–13346. doi: 10.1073/pnas.1104517108

De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. doi: 10.1038/nn.3279

De Quadros, A. (2012). *The Cambridge companion to choral music*. Cambridge, England: Cambridge University Press.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E. & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132*, 2531–1540. doi: 10.1093/brain/awp111

Delorme, A. & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Dienes, Z. (2008). Subjective measures of unconscious knowledge. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in brain research* (Vol. 168, pp. 49–64). doi: 10.1016/S0079-6123(07)68005-4

Ding, L. & Gold, J. I. (2012). Neural correlates of perceptual decision making before, during, and after decision commitment in monkey frontal eye field. *Cerebral Cortex*, *22*(5), 1052–1067. doi: 10.1093/cercor/bhr178

Dunlosky, J. & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: SAGE Publications.

Dunlosky, J., Serra, M. & Baker, J. (2007). Metamemory. In F. Durso, R. Nickerson, S. Dumais, S. Lewandowsky & T. Perfect (Eds.), *Handbook of applied cognition* (2nd ed., pp. 137– 160). Wiley, NJ: Hoboken.

Dunning, D. & Story, A. L. (1991). Depression, realism, and the overconfidence effect: Are the sadder wiser when predicting future actions and events? *Journal of Personality and Social Psychology*, *61*(4), 521–532. doi: 10.1037/0022-3514.61.4.521

Dutilh, G., Forstmann, B. U., Vandekerckhove, J. & Wagenmakers, E.-J. (2013). A diffusion model account of age differences in posterror slowing. *Psychology and Aging*, *28*(1), 64–76. doi: 10.1037/a0029875

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M. & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception & Psychophysics*, *74*(2), 454–465. doi: 10.3758/s13414-011-0243-2

Edelson, M. G., Dudai, Y., Dolan, R. J. & Sharot, T. (2014). Brain substrates of recovery from misleading influence. *The Journal of Neuroscience*, *34*(23), 7744–7753. doi: 10.1523/JNEUROSCI.4720-13.2014

Egner, T. & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), 1784–1790. doi: 10.1038/nn1594

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. *Electroencephalography and Clinical Neurophysiology*, *99*(3), 225–234. doi: 10.1016/S0921-884X(96)95711-2

Endrass, T., Franke, C. & Kathmann, N. (2005). Error awareness in a saccade countermanding task. *Journal of Psychophysiology*, *19*(4), 275–280. doi: 10.1027/0269-8803.19.4.275

Endrass, T., Klawohn, J., Preuss, J. & Kathmann, N. (2012). Temporospatial dissociation of Pe subcomponents for perceived and unperceived errors. *Frontiers in Human Neuroscience*, *6*(June), 178. doi: 10.3389/fnhum.2012.00178

Evans, S. & Azzopardi, P. (2007). Evaluation of a 'bias-free' measure of awareness. *Spatial Vision*, *20*(1-2), 61–77. doi: 10.1163/156856807779369742

Falkenstein, M., Hohnsbein, J., Hoormann, J. & Blanke, L. (1990). Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In C. Brunia, A. Gaillard & A. Kok (Eds.), *Psychophysiological brain research* (pp. 192–195). Tilburg, the Netherlands: Tilburg University Press.

Falkenstein, M., Hohnsbein, J., Hoormann, J. & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. doi: 10.1016/0013-4694(91)90062-9

Farb, N. A. S., Segal, Z. V. & Anderson, A. K. (2013). Mindfulness meditation training alters cortical representations of interoceptive attention. *Social Cognitive and Affective Neuroscience*, *8*(1), 15–26. doi: 10.1093/scan/nss066

Fernandez-Duque, D., Baird, J. A. & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition*, *9*, 288–307. doi: 10.1006/ccog.2000.0447

Fetsch, C. R., Kiani, R., Newsome, W. T. & Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron*, *83*(4), 797–804. doi: 10.1016/j.neuron.2014.07.011

First, M. B., Spitzer, R. L., Gibbon, M. & Williams, J. B. W. (2002). *Structured clinical interview for DSM-IV-TR axis I disorders: Patient edition (research version).* New York, NY: Biometrics Research.

Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130. doi: 10.1016/j.tics.2010.01.003

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231–236). Hillsdale, NJ: Erlbaum.

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

Fleming, S. M. & Dolan, R. J. (2014). The neural basis of metacognitive ability. In *The cognitive neuroscience of metacognition* (pp. 245–265). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-45190-4_11

Fleming, S. M., Dolan, R. J. & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1280–1286. doi: 10.1098/rstb.2012.0021

Fleming, S. M. & Frith, C. D. (Eds.). (2014). *The cognitive neuroscience of metacognition*. Berlin, Heidelberg: Springer. doi: 10.1007/978-3 -642-45190-4

Fleming, S. M., Huijgen, J. & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience*, *32*(18), 6117–6125. doi: 10.1523/JNEUROSCI.6489 -11.2012

Fleming, S. M. & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(July), 443. doi: 10.3389/fnhum .2014.00443

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. doi: 10.1126/science.1191883

Foote, A. L. & Crystal, J. D. (2007). Metacognition in the rat. *Current Biology*, *17*(6), 551–555. doi: 10.1016/j.cub.2007.01.061

Foti, D. & Hajcak, G. (2012). Genetic variation in dopamine moderates neural response during reward anticipation and delivery: Evidence from event-related potentials. *Psychophysiology*, *49*(5), 617–626. doi: 10.1111/j.1469-8986.2011.01343.x

Frank, M., Doll, B., Oas-Terpstra, J. & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062. doi: 10.1038/nn.2342

Fu, T., Koutstaal, W., Fu, C. H. Y., Poon, L. & Cleare, A. J. (2005). Depression, confidence, and decision: Evidence against depressive realism. *Journal of Psychopathology and Behavioral Assessment*, *27*(4), 243–252. doi: 10.1007/s10862-005-2404-x

Fuentemilla, L., Cucurell, D., Marco-Pallarés, J., Guitart-Masip, M., Morís, J. & Rodríguez-Fornells, A. (2013). Electrophysiological correlates of anticipating improbable but desired events. *NeuroImage*, *78*, 135–144. doi: 10.1016/j.neuroimage.2013.03.062

Fullerton, G. S. & Cattell, J. M. (1892). *One the perception of small differences* (Vol. 3) (No. 2). Philadelphia: Publications of the University of Pennsylvania. Retrieved from `http://babel.hathitrust.org/cgi/pt?id=mdp.39015028619891;view=1up;seq=1`

Galvin, S. J., Podd, J. V., Drga, V. & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and

incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. doi: 10.3758/BF03196546

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E. & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385–390. doi: 10.1111/j.1467-9280.1993.tb00586.x

Gherman, S. & Philiastides, M. G. (2014). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage*, *106C*, 134–143. doi: 10.1016/j.neuroimage.2014.11.036

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

Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Goodman, L. A. & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732–764. doi: 10.1080/01621459.1954.10501231

Graziano, M., Parra, L. C. & Sigman, M. (2010). Neurophysiology of perceived confidence. *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, *2010*(282), 2818–2821. doi: 10.1109/IEMBS.2010.5626567

Graziano, M. & Sigman, M. (2009). The spatial and temporal construction of confidence in the visual scene. *PloS ONE*, *4*(3), e4909. doi: 10.1371/journal.pone.0004909

Greenhouse, S. W. & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. doi: 10.1007/BF02289823

Griffing, H. (1895). *On sensations from pressure and impact* (Doctoral dissertation). Columbia University.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, *21*(1), 27–58. doi: 10.1214/aoms/1177729885

Grützmann, R., Endrass, T., Klawohn, J. & Kathmann, N. (2014). Response accuracy rating modulates ERN and Pe amplitudes. *Biological Psychology*, *96*, 1–7. doi: 10.1016/j.biopsycho.2013.10.007

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(9), 5359–5362. doi: 10.1073/pnas.071600998

Hancock, J. A. (1996). "Depressive realism" assessed via confidence in decision-making. *Cognitive Neuropsychiatry*, *1*(3), 213–220. doi: 10.1080/135468096396514

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E. & Shadlen, M. N. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience*, *31*(17), 6339–6352. doi: 10.1523/JNEUROSCI.5613-10.2011

Happé, F. G. E., Winner, E. & Brownell, H. (1998). The getting of wisdom: Theory of mind in old age. *Developmental Psychology*, *34*(2), 358–362. doi: 10.1037/0012-1649.34.2.358

Harrell Jr, F. E. & Dupont, C. (2014). Hmisc: Harrell Miscellaneous [Computer software manual]. Retrieved from `http://cran.r-project.org/package=Hmisc`

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*(4), 208–216. doi: 10.1037/h0022263

Hartwig, M. K. & Dunlosky, J. (2014). The contribution of judgment scale to the unskilled-and-unaware phenomenon: How evaluating others can exaggerate over- (and under-) confidence. *Memory & Cognition*, *42*(1), 164–173. doi: 10.3758/s13421-013-0351-4

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S. & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, *84*, 159–168. doi: 10.1016/j.neuroimage.2013.08.028

Hays, M. J., Kornell, N. & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review*, *17*(6), 797–801. doi: 10.3758/PBR.17.6.797

Heitz, R. P. & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, *76*(3), 616–628. doi: 10.1016/j.neuron.2012.08.030

Heninger, G. R., Charney, D. S. & Sternberg, D. E. (1984). Serotonergic function in depression. Prolactin response to intravenous tryptophan in depressed patients and healthy subjects. *Archives of General Psychiatry*, *41*(4), 398–402. doi: 10.1001/archpsyc.1984.01790150088012

Henmon, V. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, 196–206. Retrieved from `http://psycnet.apa.org/journals/rev/18/3/186/`

Hertzog, C., Dunlosky, J. & Sinclair, S. M. (2010). Episodic feeling-of-knowing resolution derives from the quality of original encoding. *Memory & Cognition*, *38*(6), 771–784. doi: 10.3758/MC.38.6.771

Hester, R., Foxe, J. J., Molholm, S., Shpaner, M. & Garavan, H. (2005). Neural mechanisms involved in error processing: A comparison of errors made with and without awareness. *NeuroImage*, *27*, 602–608. doi: 10.1016/j.neuroimage.2005.04.035

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26. doi: 10.1080/17470215208416600

Higham, P. A., Perfect, T. J. & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 57–80. doi: 10.1037/a0013865

Hillyard, S. A., Squires, K. C., Bauer, J. W. & Lindsay, P. H. (1971). Evoked potential correlates of auditory signal detection. *Science*, *172*(3990), 1357–1360. doi: 10.1126/science.172.3990.1357

Holroyd, C. B. & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. doi: 10.1037//0033-295X.109.4.679

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H. & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(5), 497–498. doi: 10.1038/nn1238

IBM. (2013). *IBM SPSS Statistics for Mac OS.* Armonk, NY.

Irwin, F. W., Smith, W. A. S. & Mayfield, J. F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology*, *51*(4), 261–268. doi: 10.1037/h0041911

Jack, A. I. (2013). Introspection: The tipping point. *Consciousness and Cognition*, *22*(2), 670–671. doi: 10.1016/j.concog.2013.03.005

James, W. (1890). *The principles of psychology.* New York: Henry Holt & Co.

Jankowski, T. & Holas, P. (2014). Metacognitive model of mindfulness. *Consciousness and Cognition*, *28*(1), 64–80. doi: 10.1016/j.concog.2014.06.005

Jentzsch, I. & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *Quarterly Journal of Experimental Psychology*, *62*(2), 209–218. doi: 10.1080/17470210802240655

Jersild, A. T. (1927). Mental set and shift. *Archives of Psychology*, *14*(89), 1–81.

Johnson, R. (1986). A triarchic model of P300 amplitude. *Psychophysiology*, *23*(4), 367–384. doi: 10.1111/j.1469-8986.1986.tb00649.x

Juslin, P. & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*(2), 344–366. doi: 10.1037/0033-295X.104.2.344

Juslin, P., Winman, A. & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*(2), 384–396. doi: 10.1037//0033-295X.107.2.384

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics*, *49*(3), 227–229. doi: 10.3758/BF03214307

Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Kepecs, A. & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. doi: 10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. doi: 10.1038/nature07200

Kerkhof, G. A. (1982). Event-related potentials and auditory signal detection: Their diurnal variation for morning-type and evening-type subjects. *Psychophysiology*, *19*(1), 94–103. doi: 10.1111/j.1469-8986.1982.tb02607.x

Kiani, R., Corthell, L. & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342. doi: 10.1016/j.neuron.2014.12.015

Kiani, R. & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. doi: 10.1126/science.1169405

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. doi: 10.1126/science.220.4598.671

Kleiner, M., Brainard, D. & Pelli, D. (2007). What's new in Psychtoolbox-3? In *Perception 36 ECVP Abstract Supplement.*

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. doi: 10.1037/0033-2909.119.2.254

Kolling, N., Behrens, T. E. J., Mars, R. B. & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, *336*(6077), 95–98. doi: 10.1126/science.1216930

Komsta, L. (2011). outliers: Tests for outliers [Computer software manual]. Retrieved from `http://cran.r-project.org/package=outliers`

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639. doi: 10.1037/0033-295X.100.4.609

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. doi: 10.1037//0096-3445.126.4.349

Koriat, A. (2008). When confidence in a choice is independent of which choice is made. *Psychonomic Bulletin & Review*, *15*(5), 997–1001. doi: 10.3758/PBR.15.5.997

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, *140*(1), 117–139. doi: 10.1037/a0022171

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. doi: 10.1037/a0025648

Koriat, A. & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 34–53. doi: 10.1037//0278-7393.27.1.34

Koriat, A., Sheffer, L. & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*(2), 147–162. doi: 10.1037//0096-3445.131.2.147

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: Model-based and distribution-free measures and evaluation. *Perception & Psychophysics*, *68*(3), 393–414. doi: 10.3758/BF03193685

Kornell, N. (2009). Metacognition in humans and animals. *Current Directions in Psychological Science*, *18*(1), 11–15. doi: 10.1111/j.1467-8721.2009.01597.x

Kotani, Y., Kishida, S., Hiraku, S., Suda, K., Ishii, M. & Aihara, Y. (2003). Effects of information and reward on stimulus-preceding negativity prior to feedback stimuli. *Psychophysiology*, *40*(5), 818–826. doi: 10.1111/1469-8986.00082

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Kulhavy, R. W. & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308. doi: 10.1007/BF01320096

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, *43*(3), 199–224. doi: 10.1016/0001-6918(79)90026-X

Langsrud, O. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, *13*(2), 163–167. doi: 10.1023/A:1023260610025

Lau, H. & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. doi: 10.1016/j.tics.2011.05.009

Lawrence, M. A. (2013). ez: Easy analysis and visualization of factorial experiments. [Computer software manual]. Retrieved from `http://cran.r-project.org/package=ez`

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 686–708. doi: 10.1037/a0026478

Leonesio, R. J. & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 464–470. doi: 10.1037/0278-7393.16.3.464

Li, Q., Hill, Z. & He, B. J. (2014). Spatiotemporal dissociation of brain activity underlying subjective awareness, objective performance and confidence. *The Journal of Neuroscience*, *34*(12), 4382–4395. doi: 10.1523/JNEUROSCI.1820-13.2014

Loftus, G. R. & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. doi: 10.3758/BF03210951

Logan, G. D. & Crump, M. J. C. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, *330*(6004), 683–686. doi: 10.1126/science.1190483

Luck, S. J. & Hillyard, S. A. (1994a). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology*, *31*(3), 291–308. doi: 10.1111/j.1469-8986.1994.tb02218.x

Luck, S. J. & Hillyard, S. A. (1994b). Spatial filtering during visual search: Evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1000–1014. doi: 10.1037/0096-1523.20.5.1000

Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. doi: 10.1038/nn1790

Macdonald, J. S. P., Mathan, S. & Yeung, N. (2011). Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus EEG alpha oscillations. *Frontiers in Psychology*, *2*(May), 82. doi: 10.3389/fpsyg.2011.00082

Malenka, R. C., Angel, R. W., Hampton, B. & Berger, P. A. (1982). Impaired central error-correcting behavior in schizophrenia. *Archives of General Psychiatry*, *39*(1), 101–107. doi: 10.1001/archpsyc.1982.04290010073013

Maniscalco, B. & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. doi: 10.1016/j.concog.2011.09.021

Masson, M. E. J. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203–220. doi: 10.1037/h0087426

MathWorks. (2012). *MATLAB and Statistics Toolbox.* Natick, Massachusetts.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*, 204–209. doi: 10.1214/aoms/1177731915

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory,*

and Cognition, *22*(6), 1423–1442. doi: 10.1037//0278-7393.22.6.1423

Meltzoff, A. N. & Gopnik, A. (2013). Learning about the mind from evidence: Children's development of intuitive theories of perception and personality. In S. Baron-Cohen, H. Tager-Flausber & M. Lombardo (Eds.), *Understanding other minds* (3rd ed., pp. 19–34). Oxford, England: Oxford University Press.

Merkle, E. C. & Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*(3), 391–408. doi: 10.1037/0096-3445.135.3.391

meta-. (n.d.). Retrieved from `http://www.oxforddictionaries.com/definition/english/meta-`

Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 623–634. doi: 10.1037/0278-7393.12.4.623

Metcalfe, J. & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, *15*(3), 238–246. doi: 10.3758/BF03197722

Michael, E., De Gardelle, V., Nevado-Holgado, A. & Summerfield, C. (2015). Unreliable evidence: 2 sources of uncertainty during perceptual choice. *Cerebral Cortex*, *25*(4), 937–947. doi: 10.1093/cercor/bht287

Michael, E., De Gardelle, V. & Summerfield, C. (2014). Priming by the variability of visual information. *Proceedings of the National Academy*

of Sciences of the United States of America, *111*(21), 7873–7878. doi: 10.1073/pnas.1308674111

Middlebrooks, P. G. & Sommer, M. A. (2011). Metacognition in monkeys during an oculomotor task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 325–337. doi: 10.1037/a0021611

Miltner, W. H. R., Braun, C. H. & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. doi: 10.1162/jocn.1997.9.6.788

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. doi: 10.1016/S1364-6613(03)00028-7

Montgomery, D. A. & Sorkin, R. D. (1996). Observer sensitivity to element reliability in a multielement visual display. *Human Factors*, *38*(3), 484–494. doi: 10.1518/001872096778702024

Moore, M. T. & Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, *32*(6), 496–509. doi: 10.1016/j.cpr.2012.05.004

Moran, R., Teodorescu, A. R. & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. doi: 10.1016/j.cogpsych.2015.01.002

Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural Computation*, *22*(7), 1786–1811. doi: 10.1162/neco.2010.12-08-930

Morey, R. D. & Rouder, J. N. (2014). *BayesFactor: Computation of Bayes factors for common designs.* Retrieved from `http://cran.r-project.org/package=BayesFactor`

Morís, J., Luque, D. & Rodríguez-Fornells, A. (2013). Learning-induced modulations of the stimulus-preceding negativity. *Psychophysiology*, *50*(9), 931–939. doi: 10.1111/psyp.12073

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–784). Taylor & Francis.

Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B. & Schooler, J. W. (2013). Mindfulness training improves working memory capacity and GRE performance while reducing mind wandering. *Psychological Science*, *24*(5), 776–781. doi: 10.1177/0956797612459659

Murphy, P. R., Robertson, I. H., Allen, D., Hester, R. & O'Connell, R. G. (2012). An electrophysiological signal that precisely tracks the emergence of error awareness. *Frontiers in Human Neuroscience*, *6*(March), 1–16. doi: 10.3389/fnhum.2012.00065

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. doi: 10.1037/0033-2909.95.1.109

Nelson, T. O. & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The

"delayed-JOL effect". *Psychological Science*, *2*(4), 267–270. doi: 10 .1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., Gerler, D. & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, *113*(2), 282–300. doi: 10.1037//0096-3445.113.2.282

Nelson, T. O. & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 676–686. doi: 10.1037/0278-7393.14.4.676

Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 125–173). San Diego, CA: Academic Press.

Nelson, T. O. & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (xiii ed., pp. 1–25). Cambridge, MA, US: The MIT Press.

Nieuwenhuis, S., Holroyd, C. B., Mol, N. & Coles, M. G. H. (2004). Reinforcement-related brain potentials from medial frontal cortex: Origins and functional significance. *Neuroscience and Biobehavioral Reviews*, *28*(4), 441–448. doi: 10.1016/j.neubiorev.2004.05.003

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P. & Kok, A. (2001). Error-related brain potentials are differentially related to

awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, *38*(5), 752–760. doi: 10.1111/1469-8986.3850752

Norman, D. A. & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation* (pp. 1–17). New York: Plenum. doi: 10.1007/978-1-4757-0629-1

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W. & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition*, *111*(2), 275–279. doi: 10.1016/j.cognition.2009.02.002

Oliveira, F. T. P., McDonald, J. J. & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: Expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, *19*(12), 1994–2004. doi: 10.1162/jocn.2007.19.12.1994

Overbeek, T. J., Nieuwenhuis, S. & Ridderinkhof, K. R. (2005). Dissociable components of error processing. *Journal of Psychophysiology*, *19*(4), 319–329. doi: 10.1027/0269-8803.19.4.319

Overgaard, M., Koivisto, M., Sørensen, T. A., Vangkilde, S. & Revonsuo, A. (2006). The electrophysiology of introspection. *Consciousness and Cognition*, *15*, 662–672. doi: 10.1016/j.concog.2006.05.002

Overgaard, M. & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1287–1296. doi: 10.1098/rstb.2011.0425

Overgaard, M. & Sandberg, K. (2014). Kinds of access: Different methods for report reveal different kinds of metacognitive access. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 67–85). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-45190-4_4

Pailing, P. E. & Segalowitz, S. J. (2004). The effects of uncertainty in error monitoring on associated ERPs. *Brain and Cognition*, *56*(2), 215–233. doi: 10.1016/j.bandc.2004.06.005

Palmer, E. C., David, A. S. & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*(1), 151–160. doi: 10.1016/j.concog.2014.06.007

Pannu, J. K., Kaszniak, A. W. & Rapcsak, S. Z. (2005). Metamemory for faces following frontal lobe damage. *Journal of the International Neuropsychological Society*, *11*, 668–676. doi: 10.1017/S1355617705050873

Parra, L. C., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A. & Sajda, P. (2002). Linear spatial integration for single-trial detection in encephalography. *NeuroImage*, *17*, 223–230. doi: 10.1006/nimg.2002.1212

Parra, L. C., Spence, C. D., Gerson, A. D. & Sajda, P. (2005). Recipes for the linear analysis of EEG. *NeuroImage*, *28*(2), 326–341. doi: 10.1016/j.neuroimage.2005.05.032

Patton, J. H., Stanford, M. S. & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psy-*

chology, *51*(6), 768–774. doi: 10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Peirce, C. S. (1877). Illustrations of the logic of science: The probability of induction. *The Popular Science Monthly*, *12*, 705–718.

Peirce, C. S. & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*, 73–83.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. doi: 10.1163/156856897X00366

Persaud, N., McLeod, P. & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257–261. doi: 10.1038/nn1840

Petitmengin, C., Remillieux, A., Cahour, B. & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Consciousness and Cognition*, *22*(2), 654–669. doi: 10.1016/j.concog.2013.02.004

Petrusic, W. M. & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review*, *10*(1), 177–183. doi: 10.3758/BF03196482

Pfabigan, D. M., Alexopoulos, J., Bauer, H. & Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. *Psychophysiology*, *48*(5), 656–664. doi: 10.1111/j.1469-8986.2010.01136.x

Philiastides, M. G., Ratcliff, R. & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *The Journal of Neuroscience*, *26*(35), 8965–8975. doi: 10.1523/JNEUROSCI.1655-06.2006

Pleskac, T. J. & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi: 10.1037/a0019737

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychology*, *3*(1), 5–18. doi: 10.1007/BF02686528

R Core Team. (2013). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.r-project.org/`

Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264–272. doi: 10.1037/h0022853

Rabbitt, P. M. (1968). Three kinds of error-signalling responses in a serial choice task. *The Quarterly Journal of Experimental Psychology*, *20*(2), 179–188. doi: 10.1080/14640746808400146

Rabbitt, P. M. (2002). Consciousness is slower than you think. *The Quarterly Journal of Experimental Psychology*, *55*(4), 1081–1092. doi: 10.1080/02724980244000080

Rabbitt, P. M., Cumming, G. & Vyas, S. (1978). Some errors of perceptual analysis in visual search can be detected and corrected. *The Quarterly Journal of Experimental Psychology*, *30*(2), 319–332. doi: 10.1080/14640747808400679

Rabbitt, P. M. & Rodgers, B. (1977). What does a man do after he makes an error? An analysis of response programming. *Quarterly Journal of Experimental Psychology*, *29*(4), 727–743. doi: 10.1080/14640747708400645

Rabbitt, P. M. & Vyas, S. (1981). Processing a display even after you make a response to it. How perceptual errors can be corrected. *The Quarterly Journal of Experimental Psychology Section A*, *33*(3), 223–239. doi: 10.1080/14640748108400790

Rahnev, D. A., Maniscalco, B., Luber, B., Lau, H. & Lisanby, S. H. (2012). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of Neurophysiology*, *107*(6), 1556–1563. doi: 10.1152/jn.00985.2011

Ramel, W., Goldin, P. R., Carmona, P. E. & McQuaid, J. R. (2004). The effects of mindfulness meditation on cognitive processes and affect in patients with past depression. *Cognitive Therapy and Research*, *28*(4), 433–455. doi: 10.1023/B:COTR.0000045557.15923.96

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446–461. doi: 10.1037/0033-2909.86.3.446

Ratcliff, R., McKoon, G. & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763–785. doi: 10.1037/0278-7393.20.4.763

Ratcliff, R., Philiastides, M. G. & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6539–6544. doi: 10.1073/pnas.0812589106

Ratcliff, R. & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. doi: 10.1037/0033-295X.111.2.333

Ratcliff, R., Van Zandt, T. & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300. doi: 10.1037/0033-295X.106.2.261

Raven, J. (1996). *Standard progressive matrices sets A, B, C, D & E.* Oxford, England: Oxford Psychologists Press.

Reder, L. M. & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 435–451. doi: 10.1037/0278-7393.18.3.435

Reingold, E. M. & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*(6), 563–575. doi: 10.3758/BF03207490

Resulaj, A., Kiani, R., Wolpert, D. M. & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266. doi: 10.1038/nature08275

Richardson, J. & Erlebacher, A. (1958). Associative connection between paired verbal items. *Journal of Experimental Psychology*, *56*(1), 62–69. doi: 10.1037/h0049034

Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. doi: 10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R. & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–175. doi: 10.1080/17588921003632529

Ruchkin, D., Sutton, S., Mahaffey, D. & Glaser, J. (1986). Terminal CNV in the absence of motor response. *Electroencephalography and Clinical Neurophysiology*, *63*(5), 445–463. doi: 10.1016/0013-4694(86)90127-6

Sallet, J., Camille, N. & Procyk, E. (2013). Modulation of feedback-related negativity during trial-and-error exploration and encoding of behavioral shifts. *Frontiers in Neuroscience*, *7*(November), 209. doi: 10.3389/fnins.2013.00209

Sandberg, K., Timmermans, B., Overgaard, M. & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. doi: 10.1016/j.concog.2009.12.013

Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H. & Kuboki, T. (2005). Effects of value and reward magnitude on feedback negativity and P300. *NeuroReport*, *16*(4), 407–411. doi: 10.1097/00001756-200503150-00020

Scheffers, M. K. & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 141–151. doi: 10.1037/0096-1523.26.1.141

Schmitt, J. A. J., Jorissen, B. L., Sobczak, S., Van Boxtel, M. P. J., Hogervorst, E., Deutz, N. E. P. & Riedel, W. J. (2000). Tryptophan depletion impairs memory consolidation but improves focussed attention in healthy young volunteers. *Journal of Psychopharmacology*, *14*(1), 21–29. doi: 10.1177/026988110001400102

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, *1*(3), 357–375. doi: 10.3758/BF03213977

Schwartz, B. L. & Díaz, F. (2014). Quantifying human metacognition for the neurosciences. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 9–23). Springer. doi: 10.1007/978-3-642-45190-4_2

Schwartz, B. L. & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (xiii ed., pp. 93–113). The MIT Press: Cambridge, MA, US.

Selimbeyoglu, A., Keskin-Ergen, Y. & Demiralp, T. (2012). What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision. *Clinical Neurophysiology*, *123*(6), 1158–1167. doi: 10.1016/j.clinph.2011.10.037

Selmeczy, D. & Dobbins, I. G. (2013). Metacognitive awareness and adaptive recognition biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 678–690. doi: 10.1037/a0029469

Semlitsch, H. V., Anderer, P., Schuster, P. & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology*, *23*(6), 695–703. doi: 10.1111/j.1469-8986.1986.tb00696.x

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M. & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–321. doi: 10.1016/j.tics.2008.04.008

Shea, N. (2012). Reward prediction error signals are meta-representational. *Noûs*, *48*(2), 314–341. doi: 10.1111/j.1468-0068.2012.00863.x

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C. & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193. doi: 10.1016/j.tics.2014.01.006

Smith, J. D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, *13*(9), 389–396. doi: 10.1016/j.tics.2009.06.009

Smith, J. D., Beran, M. J., Couchman, J. J. & Coutinho, M. V. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review*, *15*(4), 679–691. doi: 10.3758/PBR.15.4.679

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R. & Erb, L. (1995). The uncertain response in the bottlenosed dolphin (Tursiops truncatus). *Journal of Experimental Psychology: General*, *124*(4), 391–408. doi: 10.1037/0096-3445.124.4.391

Smith, J. D., Shields, W. E., Schull, J. & Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition*, *62*(1), 75–97. doi: 10.1016/S0010-0277(96)00726-3

Smith, J. D., Shields, W. E. & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, *26*(3), 317–373. doi: 10.1017/S0140525X03000086

Smith, P. L. & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. doi: 10.1016/j.tins.2004.01.006

Sniezek, J. A. & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. doi: 10.1006/obhd.2000.2926

Spitta, G. (2011). *Lasst Kinder zeigen, was sie können: Neu aufgelegte Beiträge zu einer demokratischen Deutschdidaktik* (S. Nickel, Ed.). Norderstedt, Germany: Books on Demand GmbH.

Squires, K. C., Hillyard, S. A. & Lindsay, P. H. (1973). Cortical potentials evoked by confirming and disconfirming feedback following an auditory discrimination. *Perception & Psychophysics*, *13*(1), 25–31. doi: 10.3758/BF03207230

Squires, K. C., Squires, N. K. & Hillyard, S. A. (1975a). Decision-related cortical potentials during an auditory signal detection task with cued observation intervals. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 268–279. doi: 10.1037/0096-1523.1.3.268

Squires, K. C., Squires, N. K. & Hillyard, S. A. (1975b). Vertex evoked potentials in a rating-scale detection task: Relation to signal probability. *Behavioral Biology*, *13*(1), 21–34. doi: 10.1016/S0091-6773(75)90748-8

Steinhauser, M. & Yeung, N. (2010). Decision processes in human performance monitoring. *The Journal of Neuroscience*, *30*(46), 15643–15653. doi: 10.1523/JNEUROSCI.1899-10.2010

Steinhauser, M. & Yeung, N. (2012). Error awareness as evidence accumulation: Effects of speed-accuracy trade-off on error signaling. *Frontiers in Human Neuroscience*, *6*(August), 240. doi: 10.3389/fnhum.2012.00240

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. doi: 10.1037/h0054651

Sugrue, L. P., Corrado, G. S. & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*(5), 363–375. doi: 10.1038/nrn1666

Sutton, S., Braren, M., Zubin, J. & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188. doi: 10.1126/science.150.3700.1187

Sutton, S., Ruchkin, D. S., Munson, R., Kietzman, M. L. & Hammer, M. (1982). Event-related potentials in a two-interval forced-choice detection task. *Perception & Psychophysics*, *32*(4), 360–374. doi: 10.3758/BF03206242

Sutton, S., Tueting, P., Hammer, M. & Hakerem, G. (1978). Evoked potentials and feedback. In D. Otto (Ed.), *Multidisciplinary perspectives in event-related potential research* (pp. 184–188). Washington, DC: United States Government Printing Office.

Szu-Ting Fu, T., Koutstaal, W., Poon, L. & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(2), 699–704. doi: 10.1016/j.jbtep.2011.09.014

Talmi, D., Atkinson, R. & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction

errors. *The Journal of Neuroscience*, *33*(19), 8264–8269. doi: 10 .1523/JNEUROSCI.5695-12.2013

Teasdale, J. D., Moore, R. G., Hayhurst, H., Pope, M., Williams, S. & Segal, Z. V. (2002). Metacognitive awareness and prevention of relapse in depression: Empirical evidence. *Journal of Consulting and Clinical Psychology*, *70*(2), 275–287. doi: 10.1037/0022-006X.70.2 .275

Tenney, E. R., MacCoun, R. J., Spellman, B. A. & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46–50. doi: 10.1111/j.1467-9280.2007.01847.x

Terrace, H. S. & Son, L. K. (2009). Comparative metacognition. *Current Opinion in Neurobiology*, *19*(1), 67–74. doi: 10.1016/j.conb.2009.06 .004

Timmers, C. & Veldkamp, B. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education*, *56*(3), 923–930. doi: 10.1016/j.compedu .2010.11.007

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P. & Foa, E. B. (2001). Memory and memory confidence in obsessive-compulsive disorder. *Behaviour Research and Therapy*, *39*(8), 913–927. doi: 10.1016/S0005-7967(00)00064-4

Tunney, R. J. (2012). Sources of confidence judgments in implicit cognition. *Psychonomic Bulletin & Review*, *12*(2), 367–373. doi: 10.3758/BF03196386

Tunney, R. J. & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition*, *31*(7), 1060–1071. doi: 10.3758/BF03196127

Ullsperger, M. & Von Cramon, D. Y. (2006). How does error correction differ from error signaling? An event-related potential study. *Brain Research*, *1105*(1), 102–109. doi: 10.1016/j.brainres.2006.01.007

Van Boxtel, G. & Böcker, K. (2004). Cortical measures of anticipation. *Journal of Psychophysiology*, *18*(2-3), 61–76. doi: 10.1027/0269-8803.18.23.61

Van Donkelaar, E. L., Blokland, A., Ferrington, L., Kelly, P. A. T., Steinbusch, H. W. M. & Prickaerts, J. (2011). Mechanism of acute tryptophan depletion: Is it only serotonin? *Molecular Psychiatry*, *16*(7), 695–713. doi: 10.1038/mp.2011.9

Van Veen, V. & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*, *14*(4), 593–602. doi: 10.1162/08989290260045837

Van Zandt, T. & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1147–1166. doi: 10.1037/0278-7393.30.6.1147

Vandekerckhove, J. & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*(1), 61–72. doi: 10.3758/BRM.40.1.61

Vesonder, G. T. & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language*, *24*(3), 363–376. doi: 10.1016/0749-596X(85)90034-8

Vickers, D. & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*(2), 179–197. doi: 10.1016/0001-6918(82)90006-3

Walderhaug, E., Lunde, H., Nordvik, J. E., Landrø, N. I., Refsum, H. & Magnusson, A. (2002). Lowering of serotonin by rapid tryptophan depletion increases impulsiveness in normal individuals. *Psychopharmacology*, *164*(4), 385–391. doi: 10.1007/s00213-002-1238-4

Wallsten, T. S. & Budescu, D. V. (2009). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review*, *10*(1), 43–62. doi: 10.1017/S0269888900007256

Wallsten, T. S., Budescu, D. V. & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*(2), 176–190. doi: 10.1287/mnsc.39.2.176

Walsh, M. M. & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(47), 19048–19053. doi: 10.1073/pnas.1117189108

Walsh, M. M. & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural ad-

aptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, *36*(8), 1870–1884. doi: 10.1016/j.neubiorev.2012.05.008

Watson, D., Clark, L. A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., ... Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271. doi: 10.1016/j.concog.2013.01.004

Wells, G. L., Ferguson, T. J. & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*(6), 688–696. doi: 10.1037//0021-9010.66.6.688

Wells, G. L., Lindsay, R. C. & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, *64*(4), 440–448. doi: 10.1037//0021-9010.64.4.440

Wessel, J. R. (2012). Error awareness and the error-related negativity: Evaluating the first decade of evidence. *Frontiers in Human Neuroscience*, *6*(April), 88. doi: 10.3389/fnhum.2012.00088

Wessel, J. R., Danielmeier, C. & Ullsperger, M. (2011). Error awareness revisited: Accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of Cognitive Neuroscience*, *23*(10), 3021–3036. doi: 10.1162/jocn.2011.21635

Wierzchon, M., Paulewicz, B., Asanowicz, D., Timmermans, B. & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, *27*, 109–120. doi: 10.1016/j.concog.2014.04.009

Wilkinson, R. T. & Seales, D. M. (1978). EEG event-related potentials and signal detection. *Biological Psychology*, *7*(1-2), 13–28. doi: 10.1016/0301-0511(78)90039-X

Wilson, T. D. & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*(2), 181–192. doi: 10.1037/0022-3514.60.2.181

Wood, J., Moffoot, A. & O'Carroll, R. (1998). 'Depressive realism' revisited: Depressed Patients are realistic when they are wrong but are unrealistic when they are right. *Cognitive Neuropsychiatry*, *3*(2), 119–126. doi: 10.1080/135468098396198

Worbe, Y., Savulich, G., Voon, V., Fernandez-Egea, E. & Robbins, T. W. (2014). Serotonin depletion induces 'waiting impulsivity' on the human four-choice serial reaction time task: Cross-species translational significance. *Neuropsychopharmacology*, *39*(6), 1519–1526. doi: 10.1038/npp.2013.351

Yeung, N. (2013). Conflict monitoring and cognitive control. In K. N. Ochsner & S. Kosslyn (Eds.), *The Oxford handbook of cognitive neuroscience: Volume 2: The cutting edges* (pp. 275–299). Oxford,

United Kingdom: Oxford University Press. doi: 10.1093/oxfordhb/9780199988709.013.0018

Yeung, N., Botvinick, M. M. & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959. doi: 10.1037/0033-295X.111.4.939

Yeung, N., Holroyd, C. B. & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*(5), 535–544. doi: 10.1093/cercor/bhh153

Yeung, N. & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *The Journal of Neuroscience*, *24*(28), 6258–6264. doi: 10.1523/JNEUROSCI.4537-03.2004

Yeung, N. & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. doi: 10.1098/rstb.2011.0416

Yeung, N. & Summerfield, C. (2014). Shared mechanisms for confidence judgements and error detection in human decision making. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 147–167). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-45190-4_7

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., ... Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in

short-term recognition memory performance. *Neuroscience Research*, *68*(3), 199–206. doi: 10.1016/j.neures.2010.07.2041

Young, S. N. (2013). Acute tryptophan depletion in humans: A review of theoretical, practical and ethical aspects. *Journal of Psychiatry & Neuroscience*, *38*(5), 294–305. doi: 10.1503/jpn.120209

Young, S. N. & Leyton, M. (2002). The role of serotonin in human mood and social interaction: Insight from altered tryptophan levels. *Pharmacology, Biochemistry and Behavior*, *71*(4), 857–865. doi: 10.1016/S0091-3057(01)00670-0

Young, S. N., Smith, S. E., Pihl, R. O. & Ervin, F. R. (1985). Tryptophan depletion causes a rapid lowering of mood in normal males. *Psychopharmacology*, *87*(2), 173–177. doi: 10.1007/BF00431803

Yu, A. J. & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–692. doi: 10.1016/j.neuron.2005.04.026

Zellermayer, M., Salomon, G., Globerson, T. & Givon, H. (1991). Enhancing writing-related metacognitions through a computerized writing partner. *American Educational Research Journal*, *28*(2), 373–391. doi: 10.3102/00028312028002373

Zimmer, A. C. (1983). Verbal vs. numerical processing of subjective probabilities. In R. Scholz (Ed.), *Decision making under uncertainty* (First ed., pp. 159–182). Elsevier Science.

Zylberberg, A., Barttfeld, P. & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*(September), 79. doi: 10.3389/fnint.2012.00079

Zylberberg, A., Roelfsema, P. R. & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition*, *27*, 246–253. doi: 10.1016/j.concog.2014.05.012

Appendix A

Experiment 5: Does decision confidence predict attention to feedback?

A.1 Staircase procedure

The difficulty for each participant was set using a staircase algorithm, inspired by the up-down staircase method developed by Kaernbach (1991) and used in Steinhauser and Yeung (2010). In the first four blocks of the experiment, which were 64 trials long each, the staircase adjusted the difficulty while participants familiarised themselves with the dot task. No confidence judgements were required in this block and no feedback was given. In this block, participants were asked to respond as accurately as possible and difficulty was adjusted so that an error rate of about 15% was obtained. The first trial had a difficulty of 2000 – that means there were 1048 dots on one side and 3048 on the other side. This very easy level of difficulty was reduced by steps of 200 over the first six trials until it reached a difference of 800 dots. This was done to ease

participants into the task. The program then changed the difficulty after every response, making the task more difficult after correct responses and easier after errors. The steps by which difficulty was adjusted were relatively large in the beginning, and then reduced twice until adjustments were made in only small steps. In the beginning, the adjustments were 44 dots (to make trials easier) and 8 (to make trials more difficult). As soon as the last 20 trials showed an error rate of between 15 and 20%, these in- and decrements were adjusted to 22 dots (to make trials easier) and 4 (to make trials more difficult). Twenty trials after this adjustment, the mean error rate of the previous 20 trials were tracked again. If it reached an error rate within 15 and 20%, the in- and decrements were adjusted once more to 10 dots (to make trials easier) and 2 (to make trials more difficult). Repeatedly decreasing the step-size ensured convergence of the staircase at 15 to 20% errors.

Again, replicating the procedure from Steinhauser and Yeung (2010), blocks 5 and 6 were similar to the previous blocks, only that now participants were asked to respond as quickly as possible. These blocks were 32 trials long each. Difficulty was not adapted during these blocks, but the experimenter monitored the participants' error rates, encouraging them to go faster so that the overall error rate was doubled to about 30%, now including both slow data-limitation errors, as well as fast-guess errors (Scheffers & Coles, 2000).

In blocks 7 to 12 (each containing 32 trials), participants were asked to make binary error judgements after each dot decision, using their index fingers. Only binary error signalling information was required for the staircase, I therefore did not use a graded scale as in the previous experiments. There was again no immediate feedback given after their responses. However, how often the *error* or *correct* category was chosen was presented on screen after the first of these blocks, giving the experimenter the opportunity to discuss with par-

ticipants how well their error signalling matched the actual task performance and encouraging them to pay close attention to this rating.

Difficulty was again adjusted in this part to ensure that participants would detect about half of their errors. For that reason, the task was made more difficult if a correct response was given (decreasing the difference in dots by 2 dots), but only after undetected errors, the task was made easier, increasing the difference in dots by 8 dots. In contrast to Steinhauser and Yeung (2010), no adjustments in difficulty were made after detected errors. For all following blocks, the level of difficulty did not change.

Throughout the entire study, not just the staircase part, participants received feedback after each block, both with regard to their average correct RT, and their average error rate. During blocks 7 to 12, however, this feedback was accompanied by a sentence that aimed at pushing participants' performance further towards a more balanced equal number of detected and undetected errors: If the ratio of the number of detected and undetected errors was smaller than 0.8, participants were reminded to "Please keep in mind to react as QUICKLY as possible!". If this message was not shown, however, and they committed more than 40% errors in the previous block, they saw the message "Please try to be more accurate!" displayed along their performance feedback.

Figure 101 shows the block-wise development of error rates (bars) and difficulty levels (connected dots). Blocks in which difficulty was adjusted are highlighted in blue. There was a substantial increase in difficulty (i.e., decrease in dot difference) from block 1 to block 2, due to the fact that the first trials in block 1 were designed to be particularly easy to ease participants into the task. First, the task was made more and more difficult over the course of the first four staircase blocks. An average error rate of 15% was aimed for in these blocks, and participants indeed committed 13.2% errors in the last of these

blocks. Blocks 5 and 6 introduced speed stress with the aim of increasing error rates to 30%. Indeed, error rates had increased to 32.8% by the end of block 6. The second part of the staircase aimed for a 2:3 to 1:1 ratio of detected and undetected errors, that means the shaded dark red and red proportions of the bars should be similar. Overall, the task was adjusted towards a higher level of difficulty in these blocks. There were still more detected errors ($M = 17.5\%$) than undetected errors ($M = 11.6\%$) by the end of block 12. Difficulty was not adjusted thereafter. The last three blocks of the experiment introduced the confidence scale. The proportion of detected errors had increased to ($M = 20.4\%$) by the end of the experiment. However, roughly two thirds of all errors were still undetected ($M = 9.4\%$), ensuring enough trials for the planned analyses. It can therefore be concluded that the staircase worked as desired.

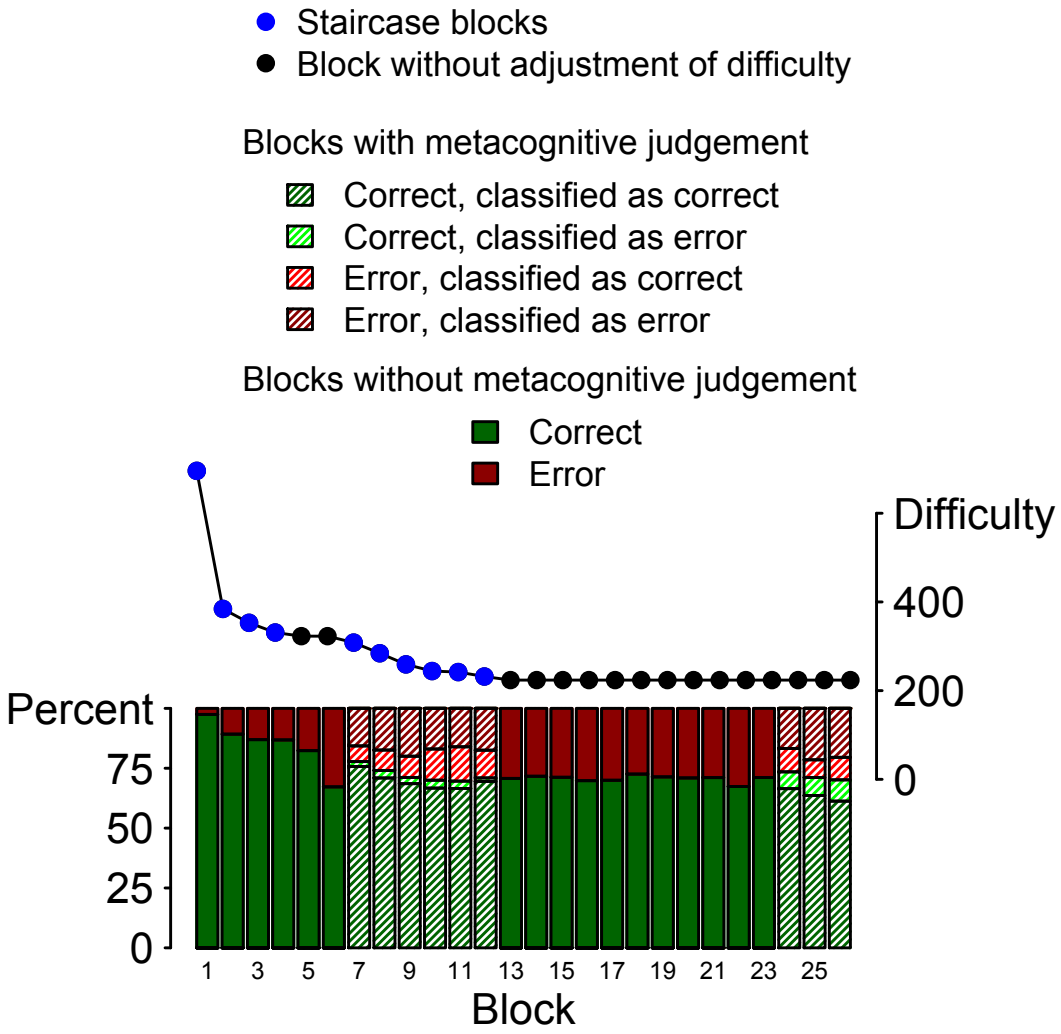

Figure 101: Development of performance (bar plots) and difficulty (filled circles) over blocks. The filled, blue circles represent blocks in which difficulty was adjusted via a staircase procedure. The scale of difficulty corresponds to the difference in dots out of a square of 4096 possible locations. Blocks in which participants were required to rate their confidence or detect their errors are highlighted in diagonally shaded bars with darker colours representing correctly classified proportions of trials.

A.2 SPN as a function of classifier bin: Results from the full four-way ANOVA model

Table 3: SPN full ANOVA. Bin: classifier quartile; Lat: lateral scalp location (left, centre, and right); Anteropost: anteroposterior scalp location (frontal, fronto-central, central, posterio-central, and posterior); Time: time window (-800 to -600 ms, -600 to -400 ms, -400 to -200 ms, -200 to 0 ms.

Factor names	**Statistical effect**
Bin	$F < 1$
Lat	$F(2,\ 32) = 13.6,\ p < 0.001,\ \eta_p^2 = 0.46$
Anteropost	$F(1.7,\ 26.4) = 16.8,\ p < 0.001,\ \eta_p^2 = 0.51$
Time	$F(1.7,\ 26.9) = 31.8,\ p < 0.001,\ \eta_p^2 = 0.67$
Bin x Lat	$F < 1$
Bin x Anteropost	$F(2.8,\ 44.3) = 3.0,\ p = 0.04,\ \eta_p^2 = 0.16$
Lat x Anteropost	$F(8,\ 128) = 2.7,\ p < 0.01,\ \eta_p^2 = 0.14$
Bin x Time	$F(3.6,\ 57.7) = 3.0,\ p = 0.03,\ \eta_p^2 = 0.16$
Lat x Time	$F(3.1,\ 49.3) = 1.5,\ p = 0.23,\ \eta_p^2 = 0.08$
Anteropost x Time	$F(1.7,\ 27.8) = 23.0,\ p < 0.001,\ \eta_p^2 = 0.59$
Bin x Lat x Anteropost	$F < 1$
Bin x Lat x Time	$F(18,\ 288) = 1.5,\ p = 0.10,\ \eta_p^2 = 0.08$
Bin x Anteropost x Time	$F(36,\ 576) = 1.6,\ p = 0.01,\ \eta_p^2 = 0.09$
Lat x Anteropost x Time	$F(24,\ 384) = 7.8,\ p < 0.001,\ \eta_p^2 = 0.33$
Bin x Lat x Anteropost x Time	$F(72,\ 1152) = 1.0,\ p = 0.38,\ \eta_p^2 = 0.06$

A.3 SPN as a function of feedback valence: Results from the full four-way ANOVA model

Table 4: SPN full ANOVA. Bin: classifier quartile; Lat: lateral scalp location (left, centre, and right); Anteropost: anteroposterior scalp location (frontal, fronto-central, central, posterio-central, and posterior); Time: time window (-800 to -600 ms, -600 to -400 ms, -400 to -200 ms, -200 to 0 ms.

Factor names	**Statistical effect**
Err	$F(1,\ 16) = 2.1,\ p = 0.16,\ \eta_p^2 = 0.12$
Lat	$F(2,\ 32) = 12.1,\ p < 0.001,\ \eta_p^2 = 0.43$
Anteropost	$F(1.7,\ 26.6) = 15.1,\ p < 0.001,\ \eta_p^2 = 0.49$
Time	$F(1.8,\ 28.1) = 27.8,\ p < 0.001,\ \eta_p^2 = 0.64$
Err x Lat	$F < 1$
Err x Anteropost	$F(1.5,\ 24.6) = 2.1,\ p = 0.15,\ \eta_p^2 = 0.12$
Lat x Anteropost	$F(3.7,\ 58.9) = 2.8,\ p = 0.04,\ \eta_p^2 = 0.15$
Err x Time	$F(3,\ 48) = 19.3,\ p < 0.001,\ \eta_p^2 = 0.55$
Lat x Time	$F(3.0,\ 48.4) = 1.7,\ p = 0.18,\ \eta_p^2 = 0.10$
Anteropost x Time	$F(1.6,\ 26.1) = 19.3,\ p < 0.001,\ \eta_p^2 = 0.55$
Err x Lat x Anteropost	$F(1.5,\ 24.7) = 2.2,\ p = 0.14,\ \eta_p^2 = 0.12$
Err x Lat x Time	$F(3.1,\ 49.6) = 3.5,\ p = 0.02,\ \eta_p^2 = 0.18$
Err x Anteropost x Time	$F(2.5,\ 40.3) = 4.7,\ p < 0.01,\ \eta_p^2 = 0.23$
Lat x Anteropost x Time	$F(24,\ 384) = 6.1,\ p < 0.001,\ \eta_p^2 = 0.28$
Err x Lat x Anteropost x Time	$F < 1$

A.4 N2pc as a function of classifier bin: Results from the full five-way ANOVA model

Table 5: N2pc full ANOVA. Bin: classifier quartiles; Stim: side on which the stimulus was shown; Resp: response hand; Elec: electrode pair (P7/P8, P3/P4, and O1/O2); Time: time window (100 to 200 ms, 140 to 240 ms, 180 to 280 ms, 220 to 320 ms.

Factor names	**Statistical effect**
Bin	$F(1.7,\ 26.7) = 3.5,\ p = 0.05,\ \eta_p^2 = 0.18$
Continued on next page	

Table 5 – continued from previous page

Factor names	Statistical effect
Stim	$F(1,\ 16) = 2.7,\ p = 0.12,\ \eta_p^2 = 0.14$
Resp	$F < 1$
Time	$F(1.8,\ 29.5) = 4.2,\ p = 0.03,\ \eta_p^2 = 0.21$
Elec	$F(2,\ 32) = 17.5,\ p < 0.001,\ \eta_p^2 = 0.52$
Bin x Stim	$F(1.8,\ 29.3) = 1.0,\ p = 0.39,\ \eta_p^2 = 0.06$
Bin x Resp	$F < 1$
Stim x Resp	$F < 1$
Bin x Time	$F(4.7,\ 74.4) = 2.0,\ p = 0.09,\ \eta_p^2 = 0.11$
Stim x Time	$F(1.6,\ 25.7) = 11.5,\ p = 0.001,\ \eta_p^2 = 0.42$
Resp x Time	$F(1.7,\ 27.6) = 4.2,\ p = 0.03,\ \eta_p^2 = 0.21$
Bin x Elec	$F(2.2,\ 34.8) = 1.6,\ p = 0.22,\ \eta_p^2 = 0.09$
Stim x Elec	$F(2,\ 32) = 2.7,\ p = 0.08,\ \eta_p^2 = 0.14$
Resp x Elec	$F(2,\ 32) = 2.3,\ p = 0.12,\ \eta_p^2 = 0.13$
Time x Elec	$F(2.7,\ 42.8) = 6.3,\ p < 0.01,\ \eta_p^2 = 0.28$
Bin x Stim x Resp	$F < 1$
Bin x Stim x Time	$F < 1$
Bin x Resp x Time	$F < 1$
Stim x Resp x Time	$F < 1$
Bin x Stim x Elec	$F < 1$
Bin x Resp x Elec	$F < 1$
Stim x Resp x Elec	$F < 1$
Bin x Time x Elec	$F < 1$
Stim x Time x Elec	$F(2.9,\ 45.7) = 4.3,\ p = 0.01,\ \eta_p^2 = 0.21$
Resp x Time x Elec	$F < 1$
Bin x Stim x Resp x Time	$F(2.7,\ 43.0) = 1.3,\ p = 0.29,\ \eta_p^2 = 0.07$
Bin x Stim x Resp x Elec	$F(1.7,\ 26.6) = 1.4,\ p = 0.25,\ \eta_p^2 = 0.08$
Bin x Stim x Time x Elec	$F < 1$
Bin x Resp x Time x Elec	$F < 1$
Stim x Resp x Time x Elec	$F < 1$
Bin x Stim x Resp x Time x Elec	$F(18,\ 288) = 1.3,\ p = 0.16,\ \eta_p^2 = 0.08$

A.5 N2pc as a function of feedback valence: Results from the full five-way ANOVA model

Table 6: N2pc full ANOVA. Err: objective accuracy or feedback valence; Stim: side on which the stimulus was shown; Resp: response hand; Elec: electrode pair (P7/P8, P3/P4, and O1/O2); Time: time window (100 to 200 ms, 140 to 240 ms, 180 to 280 ms, 220 to 320 ms.

Factor names	**Statistical effect**
Err	$F(1,\ 16) = 2.4,\ p = 0.14,\ \eta_p^2 = 0.13$
Stim	$F(1,\ 16) = 4.7,\ p = 0.04,\ \eta_p^2 = 0.23$
Resp	$F < 1$
Time	$F(1.8,\ 29.5) = 4.6,\ p = 0.02,\ \eta_p^2 = 0.22$
Elec	$F(2,\ 32) = 13.8,\ p < 0.001,\ \eta_p^2 = 0.46$
Err x Stim	$F < 1$
Err x Resp	$F < 1$
Stim x Resp	$F(1,\ 16) = 1.1,\ p = 0.30,\ \eta_p^2 = 0.07$
Err x Time	$F(1.7,\ 27.2) = 2.4,\ p = 0.11,\ \eta_p^2 = 0.13$
Stim x Time	$F(1.6,\ 26.1) = 10.2,\ p = 0.001,\ \eta_p^2 = 0.39$
Resp x Time	$F(1.8,\ 29.5) = 3.6,\ p = 0.04,\ \eta_p^2 = 0.18$
Err x Elec	$F < 1$
Stim x Elec	$F(2,\ 32) = 1.6,\ p = 0.23,\ \eta_p^2 = 0.09$
Resp x Elec	$F(2,\ 32) = 1.7,\ p = 0.20,\ \eta_p^2 = 0.10$
Time x Elec	$F(2.5,\ 39.5) = 5.6,\ p < 0.01,\ \eta_p^2 = 0.26$
Err x Stim x Resp	$F(1,\ 16) = 6.3,\ p = 0.02,\ \eta_p^2 = 0.28$
Err x Stim x Time	$F < 1$
Err x Resp x Time	$F < 1$
Stim x Resp x Time	$F < 1$
Err x Stim x Elec	$F < 1$
Err x Resp x Elec	$F < 1$
Stim x Resp x Elec	$F < 1$
Err x Time x Elec	$F(3.0,\ 48.7) = 1.5,\ p = 0.23,\ \eta_p^2 = 0.09$
Stim x Time x Elec	$F(3.0,\ 48.3) = 4.2,\ p = 0.01,\ \eta_p^2 = 0.21$
Resp x Time x Elec	$F < 1$
Err x Stim x Resp x Time	$F < 1$
Err x Stim x Resp x Elec	$F(2,\ 32) = 2.0,\ p = 0.15,\ \eta_p^2 = 0.11$
Err x Stim x Time x Elec	$F < 1$
Err x Resp x Time x Elec	$F < 1$
Stim x Resp x Time x Elec	$F(2.9,\ 46.9) = 1.2,\ p = 0.31,\ \eta_p^2 = 0.07$
Err x Stim x Resp x Time x Elec	$F(2.6,\ 41.5) = 1.1,\ p = 0.35,\ \eta_p^2 = 0.07$

A.6 P3 results from the full four-way ANOVA model

Table 7: P3 full ANOVA. Cert: classifier certainty (high and low); Conf: classifier confidence (high and low); Elec: electrode (FZ, FCZ, CZ, CPZ, and PZ); Err: accuracy (error and correct); Time: time window (100 to 200 ms, 140 to 240 ms, 180 to 280 ms, 220 to 320 ms.

Factor names	**Statistical effect**
Bin	$F(3,\ 48) = 1.8,\ p = 0.15,\ \eta_p^2 = 0.10$
Elec	$F(1.8,\ 29.3) = 15.1,\ p < 0.001,\ \eta_p^2 = 0.48$
Err	$F < 1$
Time	$F(1.5,\ 24.0) = 27.9,\ p < 0.001,\ \eta_p^2 = 0.64$
Bin x Elec	$F(5.3,\ 85.0) = 1.1,\ p = 0.37,\ \eta_p^2 = 0.06$
Bin x Err	$F(3,\ 48) = 2.1,\ p = 0.11,\ \eta_p^2 = 0.12$
Elec x Err	$F(1.6,\ 25.9) = 2.8,\ p = 0.09,\ \eta_p^2 = 0.15$
Bin x Time	$F < 1$
Elec x Time	$F(1.7,\ 27.5) = 6.0,\ p < 0.01,\ \eta_p^2 = 0.27$
Err x Time	$F(1.8,\ 28.3) = 4.2,\ p = 0.03,\ \eta_p^2 = 0.21$
Bin x Elec x Err	$F(4.5,\ 71.2) = 1.1,\ p = 0.35,\ \eta_p^2 = 0.07$
Bin x Elec x Time	$F(36,\ 576) = 2.1,\ p < 0.001,\ \eta_p^2 = 0.12$
Bin x Err x Time	$F(3.0,\ 48.1) = 1.9,\ p = 0.15,\ \eta_p^2 = 0.10$
Elec x Err x Time	$F < 1$
Bin x Elec x Err x Time	$F(36,\ 576) = 3.2,\ p < 0.001,\ \eta_p^2 = 0.17$

Appendix B

Experiment 6 – Signal reliability affects metacognitive judgements

B.1 First-order effects for the two median-split groups

For the median-split group with the small difference in correct RTs, there was again a main effect of stimulus mean on correct RTs, $F(1,\ 9) = 38.9,\ p < 0.001,\ \eta_p^2 = 0.81$, and error rates, $F(1,\ 9) = 110.3,\ p < 0.001,\ \eta_p^2 = 0.92$. The main effect of stimulus variance was replicated for both correct RTs, $F(1,\ 9) = 54.6,\ p < 0.001,\ \eta_p^2 = 0.86$, and error rates, $F(1,\ 9) = 40.6,\ p < 0.001,\ \eta_p^2 = 0.82$. The two factors showed an interaction only for error rates, $F(1,\ 9) = 6.5,\ p = 0.03,\ \eta_p^2 = 0.42$, and not for correct RTs, $F < 1$.

The other median-split group showed a reliable difference in correct RTs between the two medium conditions, $t(9) = 5.7,\ p < 0.001$, but no difference in error rates, $t < 1,\ BF_{NULL} = 2.77$. This group also showed a main effect of stimulus mean on both correct RTs, $F(1,\ 9) = 41.2,\ p < 0.001,\ \eta_p^2 = 0.82$, and error rates, $F(1,\ 9) = 87.2,\ p < 0.001,\ \eta_p^2 = 0.91$, as well as a reliable

effect of stimulus variance on correct RTs, $F(1,\ 9) = 47.5,\ p < 0.001,\ \eta_p^2 = 0.84$, and error rates, $F(1,\ 9) = 59.2,\ p < 0.001,\ \eta_p^2 = 0.87$. Once more, a reliable interaction was found only for error rates, $F(1,\ 9) = 5.7,\ p = 0.04,\ \eta_p^2 = 0.39$, and not for correct RTs, $F < 1$.

B.2 Confidence effects for the two median-split groups

For the sake of completeness, confidence analysis for the two median-split groups will be reported here. These data are presented in Table 8. For the median-split group with the small difference for correct RTs, there was a reliable main effect of stimulus mean, $F(1,\ 9) = 49.8,\ p < 0.001,\ \eta_p^2 = 0.85$. Participants were more confident in the *high mean* condition on correct trials, and less confident in the *high mean* condition on error trials. This pattern was also reflected in a reliable interaction between stimulus mean and accuracy, $F(1,\ 9) = 74.9,\ p < 0.001,\ \eta_p^2 = 0.89$. There was also a reliable effect of accuracy, $F(1,\ 9) = 83.4,\ p < 0.001,\ \eta_p^2 = 0.90$, with higher confidence on correct, $M = 5.0$, compared to error trials, $M = 3.3$. There was, however, no reliable effect of stimulus variance, $F < 1$. Stimulus mean and variance showed a reliable interaction, $F(1,\ 9) = 8.0,\ p = 0.02,\ \eta_p^2 = 0.47$, as well as stimulus variance and accuracy, $F(1,\ 9) = 18.4,\ p < 0.01,\ \eta_p^2 = 0.67$. There was also a marginally significant three-way interaction, $F(1,\ 9) = 3.9,\ p = 0.08,\ \eta_p^2 = 0.30$.

For the median-split group with the larger difference in correct RTs, there was no reliable effect of stimulus mean, $F(1,\ 9) = 2.7,\ p = 0.14,\ \eta_p^2 = 0.23$, but instead of stimulus variance, $F(1,\ 9) = 6.4,\ p = 0.03,\ \eta_p^2 = 0.42$. This effect was again in opposite directions for correct and error trials, with higher confidence for *low variance* trials compared to *high variance* trials for

corrects and the reverse pattern for error trials. This effect was also reflected in a significant interaction between stimulus variance and the objective accuracy of the trial, $F(1,\ 9) = 15.1,\ p < 0.01,\ \eta_p^2 = 0.63$. This group also showed an effect of objective accuracy with higher confidence for correct, $M = 4.89$, compared to error trials, $M = 3.38$, $F(1,\ 9) = 69.6,\ p < 0.001,\ \eta_p^2 = 0.89$. Stimulus mean and accuracy showed a reliable interaction, $F(1,\ 9) = 33.1,\ p < 0.001,\ \eta_p^2 = 0.79$, in the same direction as for the other median-split group. No other interactions were significant, $Fs < 1$. For this median-split group, there was a reliable difference in confidence between medium conditions only for correct, $t(9) = 3.2,\ p = 0.01$, but not for error trials, $t(9) = 1.4,\ p = 0.20,\ BF_{NULL} = 1.53$.

Table 8: Median split groups

Groups	Small RT difference				Large RT difference			
Mean	high		low		high		low	
Variance	low	high	low	high	low	high	low	high
Correct RTs (ms)	590	652	647	704	608	708	652	753
Error Rate	0.05	0.11	0.15	0.24	0.06	0.13	0.14	0.26
Confidence (cor)	5.4	5.0	5.1	4.6	5.4	4.7	5.1	4.4
Confidence (err)	2.5	3.1	3.6	3.8	3.0	3.3	3.5	3.7

Appendix C

Experiment 8 – Neurophysiological mechanisms of stimulus mean and variance processing

C.1 Average confidence

The good metacognitive resolution of participants was also reflected in the reliable difference between average correct- and error-trial confidence, as plotted in the two panels of Figure 93, $F(1,\ 15) = 72.3,\ p < 0.001,\ \eta_p^2 = 0.83$: Participants were more confident on correct than on error trials. As can be seen in the figure, the grey line lies on top of the black line in the left panel, but the order of lines is reversed in the right panel. In other words, participants were more confident for the easier, *high mean* condition on correct trials and less confident on error trials. This was expressed in a reliable interaction between stimulus mean and accuracy, $F(1,\ 15) = 26.7,\ p < 0.001,\ \eta_p^2 = 0.64$. If tested separately, there was indeed a reliable main effect for both errors, $F(1,\ 15) = 28.5,\ p < 0.001,\ \eta_p^2 = 0.66$, and correct trials, $F(1,\ 15) = 17.5,\ p <$

$0.001,\ \eta_p^2 = 0.54$. There was no reliable overall main effects of stimulus mean, $F(1,\ 15) = 1.3,\ p = 0.27,\ \eta_p^2 = 0.08$. There was also a reliable interaction between stimulus variance and accuracy, $F(2,\ 30) = 4.6,\ p = 0.02,\ \eta_p^2 = 0.23$, but only for correct trials was there a reliable main effect of stimulus variance with higher confidence for the easier, *low variance* conditions, $F(2,\ 30) = 16.5,\ p < 0.001,\ \eta_p^2 = 0.52$. The opposite pattern was not found for the error trials, $F(2,\ 30) = 1.4,\ p = 0.27,\ \eta_p^2 = 0.08$. Overall, there was a reliable main effect of stimulus variance, $F(2,\ 30) = 10.2,\ p < 0.001,\ \eta_p^2 = 0.40$. Finally, stimulus mean and variance did not interact, $F < 1$, neither for the correct, $F(2,\ 30) = 1.1,\ p = 0.36,\ \eta_p^2 = 0.07$, nor the error trials, $F < 1$, if analysed separately. This is also expressed in a null effect for the three-way interaction between stimulus mean, stimulus variance, and accuracy, $F < 1$.

C.2 Comparing the influence of stimulus mean and variance on confidence

Findings from Experiments 6 and 7 suggested that stimulus variance – and not stimulus mean – explains confidence over and above its effect on first-order performance. The question arises as to whether this also applied here. Similar individual regression models were therefore fitted to the data with the results shown in Figure 94. The first model, the normative model, or null model is shown in the left-most bars. In this model, only accuracy predicts confidence. The regression weights for this predictor were reliably different from zero, $t(15) = 14.1,\ p < 0.001$, as can also be seen in the upper panel of Figure 94, which displays the (signed) t-value for this test. This value is positive, meaning that accuracy and confidence are positively correlated, as would have been expected: The more accurate a participant is, the more confident he will

be. This model already explained a substantial proportion of variance in the data, $R^2 = 0.59$. The R^2 values are shown in the medium panels of Figure 94. The BIC of this model was 9.24, as shown in the bottom panel of Figure 94.

This null model can then be compared to a slightly more complex model, model 1. This model adds RT as an additional predictor. Both predictors were reliably different from zero, $ts >= 4.1,\ ps < 0.001$. The absolute t-values shown in Figure 94 show that RT had a negative t-value, meaning that the higher RT was, the less confident participants judged their responses, as would have been expected according to the time heuristic. On average, this model explained slightly more variance with an R^2 of 0.70. The question is though, whether this increase in explained variance was 'worth' adding another predictor. The BIC was lower, $BIC = 7.65$, suggesting a better model fit. However, the difference in BIC values was not reliable over participants, $t(15) = 1.7,\ p = 0.12$. This suggests that adding RT as an additional predictor does not hurt the goodness of fit, but it also does not reliably increase it.

Model 2a adds stimulus mean as a third predictor. As in the previous model, both accuracy and RT were reliable predictors of confidence, $ts >= 3.6,\ ps < 0.01$. Stimulus mean, however, did not predict confidence significantly, $t(15) = 1.1,\ p = 0.29$. The explained variance was almost identical to model 1, $R^2 = 0.71$, and the mean BIC was even increased, $BIC = 9.14$, suggesting that this model's goodness of fit was actually worse compared to model 1. This difference in the fits of the two models was indeed reliable, $t(15) = 3.3,\ p < 0.01$, suggesting that the model with only accuracy and RT ought to be preferred over the more complex model, which also includes stimulus mean as a predictor. Stimulus mean does not seem to explain confidence over and above its effect on accuracy and RT.

A similar model was then tested for accuracy, RT, and stimulus vari-

ance, model 2b. For this model, all three predictors were reliably different from zero, $ts >= 2.5,\ ps <= 0.02$. RT was again negatively correlated with confidence, as was stimulus variance: The more variable a stimulus was, the less confident people judged their responses, echoing findings from the previous two experiments. Compared to model 1, this model explained numerically more variance, $R^2 = 0.84$. Whether this increase in explained variance justified the additional parameter was tested by comparing the BIC values of the two models across participants. The BIC of model 2b was indeed reduced, $BIC = 2.09$, and this reduction was reliable, $t(15) = 2.7,\ p = 0.02$. This suggests that stimulus variance is a meaningful parameter when explaining confidence, which affects confidence judgements over and above its influence on accuracy and RT.

Finally, model 3 was the most complex model with all four variables predicting confidence. Only accuracy, $t(15) = 7.2,\ p < 0.001$, and stimulus variance, $t(15) = 2.5,\ p = 0.03$, were reliable predictors of confidence. RT, however, was only marginally significant, $t(15) = 2.1,\ p = 0.05$, and so was stimulus mean, $t(15) = 2.0,\ p = 0.07$. The reason why RT was no longer a reliable predictor of confidence was presumably that RT shared a substantial amount of variance with both stimulus mean and variance, which was reduced when both variables were included into the model and their shared variance components therefore excluded. Overall, this model explained the most variance, $R^2 = 0.87$, and also had the best goodness of fit, $BIC = 1.12$. When compared to both models with only three predictors, only the step from model 2a to model 3 was significantly better, $t(15) = 2.8,\ p = 0.01$. Adding stimulus mean to the regression model with accuracy, RT, and stimulus variance (model 2b) did not reliably increase the goodness of fit, though, $t < 1$.

Taken together, these findings are highly consistent with what was

reported for the previous two studies: People's confidence was not just by first-order performance such as accuracy and RT, but also by the variance of the stimulus and this effect was present over and above the effect stimulus variance had on accuracy and RT. This was not found for stimulus mean.

C.3 ERN amplitude as a function of difficulty

For an analysis of the ERN (presented in Figure 99) in relation to difficulty condition, scalp activity averaged over the time window of the ERN (-30 to 70 ms as in previous analyses for this data set) were submitted to a repeated-measures ANOVA with anteroposterior scalp location, stimulus mean, and variance as factors. As previously reported, there was a reliable effect of location, $F(2.0,\ 29.5) = 16.8,\ p < 0.001,\ \eta_p^2 = 0.53$, reflecting that the voltages were strongest (i.e., most negative) at the most frontal electrode FZ, $M = -1.17$. There were, however, no main effects of neither stimulus mean nor variance, $Fs < 1$. The interaction between location and stimulus mean was found to be marginally significant, $F(2.1,\ 31.3) = 2.6,\ p = 0.08,\ \eta_p^2 = 0.15$. This reflected the fact that the difference between the low and the *high mean* condition was largest for the more frontal electrodes and decreased monotonically towards the rear of the head; $M_{FZ} = 0.17\ \mu V$, $M_{FCZ} = 0.14\ \mu V$, $M_{CZ} = -0.06\ \mu V$, $M_{CPZ} = -0.15\ \mu V$, $M_{PZ} = -0.32\ \mu V$. The interaction between the location and stimulus variance was not reliable, though, $F(2.9,\ 43.1) = 1.2,\ p = 0.34,\ \eta_p^2 = 0.07$. Stimulus mean and variance also did not show an interaction, $F < 1$, but there was a reliable three-way interaction, $F(3.2,\ 48.1) = 5.0,\ p < 0.01,\ \eta_p^2 = 0.25$. When comparing the two medium conditions, which were matched for first-order task performance, none of the electrode locations showed a reliable difference, $ts <= 1.5,\ ps >= 0.15$.