A metacognitive journey in monitoring:

Modelling students' multiple-choice calibration in a non-traditional task environment

Blanka Sara Palfi

Supervisors: Daina Crafa, Joshua Charles Skewes

A thesis presented for the degree of
Cognitive Science Master's

Linguistics, Cognitive Science and Semiotics
School of Communication and Culture
Aarhus University
Denmark
01.12.2021

Summary

The overarching theme of the present thesis is metacognition, inspired by the natural bedding of the project, my position as an instructor of Study Techniques at the Cognitive Science programme at Aarhus University. My approach to instructing the course has been to engage students in discovering their own best ways to learning, thinkink about their patterns of thinking. Stemming from the field of metacognition, monitoring became the focus of this thesis and can be defined as, simply put, the process of overseeing cognitive activities. From a philosophical perspective, researchers have long debated the paradoxical underpinnings of meta-level processes, also relevant in metacognitive monitoring, the homunculus problem: how can the conscious cognitive system oversee itself in action while, at the same time, also being a part of it? This thesis hypothesises that metacognitive monitoring is not a parallel, but instead a sequential process elicited by external prompting.

With regards to monitoring, this thesis presents a local, online calibration study investigated in relation to a multiple-choice task environment. Monitoring is then assessed as confidence judgments after each question without feedback on performance. The multiple-choice task environment, however, does not conform to previously used 'traditional' settings. In traditional metamemory tasks, the information acquisition phase and the testing phase of the experiment is controlled by the researchers. Qualities of the trials in the testing phase can therefore objectively be assessed and classified. In traditional multiple-choice tasks in educational research, although not controlling the information acquisition phase as such, the experimenters set the questions under study and, thus, are able to assess and classify them. In this project, however, the questions for the multiple-choice test were made by the students themselves, creating a rather decontrolled task environment from a research perspective. This characteristic is both a strength and a weakness of this thesis. It is a weakness as questions may have been overlapping and I, as a researcher, had no control over the saliency of the trials. On the other hand, this characteristic eliminates experimenter biases and offers additional support for the hard-easy effects of question difficulty.

The chosen method of investigation is Bayesian cognitive modelling of behavioural data. Computational models are built to create algorithmic hypotheses describing observed behaviour in the task. This is done following probabilistic procedures to make inferences about latent variables associated with specific psychological interpretations. Overall, four models were built: Two incorporating null-hypotheses, and two operationalising hypotheses based on behavioural patterns found in multiple-choice tasks in educational settings and calibration studies. The first of these hypotheses is related to the philosophical research question by directly linking performance and calibration accuracy, as the cognitive event generating the answer is suggested to be repeated for evaluation to produce a confidence judgment. The second hypothesis

set out to replicate the hard-easy effect of question difficulty on an individual basis. Trials, consisting of a multiple-choice question and an associated confidence rating, are modelled as unique events for each participant, a novel element of this project in comparison to previous research due to the non-traditional task environment. The procedure of cognitive modelling involved a process of model evaluation through parameter and model recovery studies, and a descriptive adequacy study, posterior-predictive checks, where models were fit on the real, collected data to evaluate which model describes behaviour the best. The model that accounted for effects of question difficulty outperformed all three alternative models suggesting some universality of the hard-easy effect regardless of experimenter-bias. Findings also suggest validity of the first hypothesis, a repetition account of monitoring. However, many limitations of this study are discussed in detail, such as the uncontrolled task environment, alternative approaches and models, and conceptual issues.

Keywords: metacognitive monitoring, online calibration, bayesian cognitive modelling, signal detection theory

Contents

Summary **1**

Introduction **5**

1. Metacognition and the context and motivation of the thesis 5

1.1 Relevance of metacognition 5

1.2 Philosophical questions and cognitive constructs 7

2. Metacognitive monitoring 9

3. Monitoring in multiple-choice tasks 11

3.1 Metamemory, monitoring, and calibration 11

3.2 Behavioural patterns found in monitoring 13

4. Measurements of online calibration 16

Research questions **21**

Methods **23**

1. Data collection and task environment 23

2. Cognitive modelling 24

3. modelling behaviour in the task and signal detection theory 26

3.1 Simple signal detection theory on 'Yes/No' tasks 26

3.2 Equal versus Unequal variance SDT 30

3.3 modelling the 4-point rating scale 32

3.4 The competing models investigating the research questions 34

Results **41**

1. Parameter recovery study 41

2. Model recovery study 48

3. Descriptive adequacy (posterior-predictive checks) 49

Discussion **52**

1. Interpretation of the results and contributions 52

2. Limitations and alternative approaches 56

    2.1 Method of observation: task environment 56

    2.2 Method of investigation: modelling 60

    2.3 Conceptual issues and alternative potentials 62

3. Concluding remarks 64

Acknowledgments **65**

References **66**

Appendices **71**

Introduction

1. Metacognition and the context and motivation of the thesis

The overarching theme of this thesis is metacognition. In this first part of the introduction, the concept of metacognition and its philosophical background, its relevance in research, and proposed general constructs of metacognition will be examined. This will be done in the natural bedding of this study, its context, a course called Study Techniques which was offered for first semester students at the Cognitive Science Bachelor's programme at Aarhus University (AU). Metacognition as a topic of choice was motivated by my position as an instructor of the course between 2018 and 2020. Cognitive Science at AU is an international programme welcoming increasingly more international students over the years since its foundation in 2015. The addition of Study Techniques to the first semester was motivated by student welfare initiatives at the programme. *Metacognition* is most commonly referred to as the knowledge about and regulation of cognitive processes in learning (Flawell, 1979; Brown, 1978 as cited in (Veenman, Hout-Wolters, & Afflerbach, 2006, p. 3)). Studying cognition naturally made me study my own cognition which gave me a sense of control, agency over my learning processes resulting in more efficient strategies that fit me. As an instructor of Study Techniques, over the years, my approach shifted focusing more and more on metacognition and self-discovery.

1.1 Relevance of metacognition.

Starting university has never been an easy transition. There are many new guidelines, rules and unique traditions which can be confusing, especially for foreigners. Over the three years of instructing Study Techniques my approach to the course became more open-ended, exploratory and focused on the journeys of students discovering their own best practices in learning. Studying cognitive science and researching cognition conditions us in this field to think about thinking all the time. The object of study in cognitive science is also the subject of study. We study people and their mental processes and we need to think about what they could possibly think about, systematically, going through all possible scenarios. When I was going through the semesters of my bachelor's, I learned a great deal about the mind in general, different minds, and my own mind. I studied a variety of processes, skills and associated disorders. I remember feeling frustrated that if I would have known all this earlier, I would have been more successful at learning with less effort. When I take the time to explore and reflect on my own practices, habits, strengths, and weaknesses with the knowledge I have about cognition, I feel empowered. With some effort and engagement I managed to be able to clearly identify my weaknesses and develop strategies relying on my strengths to overcome them. This has been my first-hand experience of

the saying from Sir Francis Bacon, "knowledge itself is power" (Bacon, 1597). Non-pathological populations, while commonly referred to as traditionally developed in research, can be very cognitively different from one another as it has been recognised in association with the emergence of the neurodiversity paradigm (Armstrong, 2012). Many may be on the spectrums of even multiple cognitive impairments. These might be hidden because of naturally developed coping strategies and surrounding cultural norms. Discovering weaknesses and associated coping strategies can be empowering because with awareness comes conscious decision making and control (Armstrong, 2012). Next time I face a challenge I know is difficult for me I am aware of my toolbox, how to alter the task for myself to be able to solve it. My personal experience of taking responsibility for my own learning and cognition motivated me to apply for the position as an instructor of Study Techniques and to devote my time to emphasise the importance of metacognitive skills.

Metacognition has been a focus in education research due to its powerful relationship with learning processes. It has been found that metacognitive skills are the most powerful predictors of learning (Wang, Haertel, and Walberg, 1990, as cited in Veenman, Hout-Wolters, and Afflerbach (2006, p. 3)). Additionally, even though metacognition is strongly tied to cognitive skills, it seems that metacognitive skills are not the same as intellectual abilities; metacognitive skills alone can explain about 17% variance in learning, and can thus even compensate for limited cognitive abilities (Sternberg, 1990; Veenman, Wilhelm & Beishuizen, 2004; Veenman & Spaans, 2005, as cited in Veenman et al. (2006, p. 6)). There is evidence concerning the generality of metacognitive skills showing that they seem to be transferable across domains (Veenman, Elshout & Meijer, 1997; Veenman & Verheij, 2003; Veenman et al., 2004, as cited in Veenman et al. (2006, p. 7)). Through the engagement of metacognition in specific learning scenarios students may develop 'rules of thumbs' in their ways of thinking that they can utilise in different domains and new scenarios. However, findings are ambiguous around the generality of metacognitive skills (Veenman et al., 2006, p. 3).

From a developmental perspective, any conceptual understanding of self starts with the notion of the *theory of mind* (Meltzoff, 1990, 2007a; Schneider & Löffler, 2016). There is a stage in infant development when a baby starts to recognise that others around them are similar to them and that they have minds and mental states. Behavioural indicators of this phenomenon are observable signature patterns such as following the gaze of a parent or imitating others in the environment (Meltzoff, 1990, 2007a; Schneider & Löffler, 2016). Theory of mind, the implicit understanding of mental states and minds, is the most primitive prerequisite of any higher-order process of reflection. It is most likely that seeds of actual metacognitive skills are already developing during early years of schooling and it seems that intellectual abilities offer a head-start at this process (Veenman et al., 2006, p. 8). Veenman and Spaans (2005, as cited in Veenman et al. (2006, p. 8)) found that, initially, metacognitive

skills are domain-specific and then later become transferable across subjects. Furthermore, Veenman argues that metacognitive skills become more sophisticated over time as formal educational contexts call for them more extensively (Veenman et al., 2006, p. 8). Weil et al. (2013) investigated metacognitive monitoring skills, measured as judgments of confidence in accuracy, in a visual perceptual task with participants between the ages of 11 and 41 years. They found that these metacognitive abilities improve significantly in adolescence and show a prolonged developmental trajectory. Adolescence has been associated with an increased emergence of self-concept and self-awareness, thus this period can offer a great opportunity for acquiring high level metacognitive competencies (Blakemore & Choudhury, 2006; Schneider, 2008; Schneider & Löffler, 2016; Weil et al., 2013). The first year of starting a university degree is an amazing opportunity to reflect on and explore one's own ways of learning. While students in the first semester at the Cognitive Science Bachelor's are diverse in their ages, the majority are around 20 years of age (refer to Methods). The journey of self-discovery is personal, but students can be prompted to engage in it for their own benefit and to discuss and learn from each other. This has been the overarching aim of Study Techniques that I was lucky to instruct. As I started researching metacognition for the course, it naturally also became the topic and bedding of my thesis. My journey of taking control of my learning made me fascinated by the concept of agency in metacognition and its relevance for education. The upcoming section dives deeper into the nature of metacoginitive processes.

1.2 Philosophical questions and cognitive constructs.

The notion of agency is central to both metacognition and *self-concept* (Hacker, Dunlosky, & Graesser, 2009). Successful students take responsibility and control over their learning processes. They observe, reflect, evaluate and change. Such activities allow one to become an agent of one's own cognition (Hacker et al., 2009). Taking charge of one's self-image and learning, realising goals, needs, and current states through observation of oneself, and implementing necessary changes is empowering. Self-concept may be summarised as the description of oneself in relation to one's norms and values or how one is ought to be (prescriptions), and to expectations (future projections). These may change over time through confirmations and disconfirmation by oneself or others (Hattie, 1992, p. 37). Crucial characteristics of self-concept include the actuality of a flow of conscious experiences, reflections and memory of such experiences, flaws in making conclusions about oneself in light of observations, and its changing nature over time with the accumulation of more experiences (Hattie, 1992, pp. 12-16). In short, self-concepts are ever-changing beliefs of oneself. In comparison to self-concept, metacognition is most commonly referred to as the 'knowledge about and regulation of one's cognitive activities in learning processes' (Veenman et al., 2006, p. 3). Bringing the two together, one can have self-concepts (beliefs) associated with their learning self that originate from metacognitive

activities. Multiple selves have been proposed since William James' investigation of self-concept that highlights the shaping, evolving agent and their beliefs. However, behind both notions of self-concept and metacognition lies the *homunculus problem* (Veenman et al., 2006, p. 5). Namely, who and what is this entity overlooking all cognitive activities and changes of self? The philosophical problem is that if there is a mind in charge of the mind then that mind also needs a 'smaller' mind in charge and so on, like a never-ending Russian doll. How can something govern the cognitive system having the same capabilities as its object, the cognitive system itself, while also being a part of the cognitive process taking place? The concept of metacognition inherently proposes a higher-order agent, which is problematic. This puzzle became the philosophical research question of the present thesis.

The aforementioned philosophical paradox leads up to a core idea in the present study. The central aim of computational cognitive model-building, the method applied in this thesis, is to operationalise latent processes in a mathematical way that incorporates conceptual hypotheses. In the chosen task environment, where the main process was information recall and reflection on it, the implicit hypothesis of the models indirectly proposes the idea that metacognition is a sequential process and involves the repetition of the cognitive event it monitors. To illustrate it with an example, consider asking someone what they had for dinner. This person now recalls their meal from their memory of the event. Now asking them whether they are sure about their answer they will again recall the same memory as they did the first time (repeating the cognitive activity) while searching for errors whether they made a mistake. Thus, the hypothesis is that metacognition in this specific context is repeating the same cognitive activity for a second time with an aim of finding errors in the process. It is not a higher-order agent automatically and in parallel overlooking the cognitive system, but it is a linear second process prompted post-task, and something, e.g. a question ('are you sure?'), is needed to elicit the metacognitive response. The related research question is; whether it is possible to build models directly linking knowledge and the monitoring of that, simulating a repetition of the cognitive event, and if so, to what degree does it seem meaningful. While the idea sounds a reasonable proposition, I have to admit I was not planning on taking a stance on the philosophical issues around the nature of metacognition.

In the following, cognitive frameworks and constructs are examined in an attempt to place the skill under investigation in this thesis. Nelson (1996) proposes an account of the relationship and information flow between the general cognitive system and metacognition (as cited in Veenman et al. (2006, p. 4)). Based on his paradigm, cognitive activity takes place at a so-called 'object-level' while a 'meta-level' regulates the object-level. Information flows bi-directionally in this conceptual framework. On one hand, there is a bottom-up process in which the state of the object-level is monitored and registered for the meta-level. On the other hand, the meta-level controls the

object-level through executive processes. This theory is in alignment with the commonly used definition of metacognition that also implies an awareness or the ability to observe and thus gain knowledge of cognition, and a conscious regulation of cognitive activities (Flavell, 1979; Brown, 1978 as cited in Veenman et al. (2006, p. 3)). The most widely accepted components of metacognition are knowledge and beliefs about cognition, which resonates with self-concepts around cognition and learning; monitoring of cognition, that is, the ability to consciously observe cognitive processes; and regulating cognition, Nelson's top-down flow that implies executive processes on top of cognition that guides it (Hacker et al., 2009; Veenman et al., 2006, p. 4). An earlier work from Schraw and Dennison (1994, p. 460) distinguishes only between knowledge of cognition and regulation of cognition for the purpose of developing their metacognitive awareness scale. Under knowledge of cognition they specify three subcomponents; declarative knowledge (about self and strategies), procedural knowledge (how to apply strategies), and conditional knowledge (when and why to use strategies). Similarly, five subprocesses of regulation of cognition are distinguished; planning, information management strategies, comprehension monitoring, debugging strategies, and evaluation. They, however, could not validate the subcomponents of the two main categories as part of their scale. This may be since such subprocesses can be intertwined to a great degree. Additionally, any subjective self-report is prone to flaws. Wordings of the items naturally create bias since researchers came up with them and not the students by themselves individually. While the ultimate abstract goal in educational settings is to enhance students' awareness of metacognition, and being able to assess such comprehension of oneself is important for the associated research, the present thesis takes on a less ambitious goal.

Initially, multiple aspects of metacognition in the context of Study Techniques was planned to be investigated, but due to time and resource limits of a master thesis, the scope has been narrowed down. The focus of this research is to study metacognitive monitoring, specifically. However, the definition of monitoring is a little fuzzy as the use of the concept changes slightly from paper to paper, even from the same authors. Comprehension monitoring, for example, best described as the 'working self' regulating preexisting knowledge for new information to enter memory (Conway, 2005 as cited in Hacker et al. (2009, p. 2)), is not the monitoring skill investigated here. It resonates more with Schraw and Dennison (1994) the evaluation subcomponent of regulatory processes, but the following section is dedicated entirely to specify monitoring, the specific process under study.

2. Metacognitive monitoring

In an attempt to summarise and clarify the place and relevance of monitoring skills, a hierarchy of the concepts, introduced in the previous section, is proposed. First,

from the bottom-up, there is cognition. Metacognition always draws on cognition. Cognition gives the context and it thus naturally follows that without domain knowledge, in which competence needs to be evaluated, metacognition is difficult to elicit if even possible (Veenman et al., 2006, p.5). Then, monitoring knowledge is the next prerequisite for any higher order metacognitive processes (Tobias & Everson, 2009, p. 109). The ability to access the state of current knowledge generated from prior learning experiences is essential to accordingly act and regulate cognitive activities (Tobias & Everson, 2009). Monitoring in this context, and the stance that this thesis adopts, is the accuracy with which one can discriminate between what they know and what they have yet to know, thus needing to learn or revise. Higher order metacognitive skills, such as strategy selection, evaluation of learning, and planning, may also be monitored (observed) and aid the forming of knowledge and beliefs about oneself, i.e. self-concepts around learning. On the top of the hierarchy stands the awareness of that knowledge and beliefs of self alongside with the awareness of all the other components further down acquired through experiences and the monitoring of those. From this line of thought it follows that monitoring is indeed crucial and the first such process starts at monitoring prior knowledge, distinguishing between what one knows and what one does not, accessing memory of domain-specific knowledge and evaluating success at recalling information.

Monitoring is considered a regulatory process of metacognitive skills (Nietfeld, Cao, and Osborne,2005, p. 9; Tobias and Everson, 2009, p. 107). As proposed by Nelson (1996), it is through monitoring that the meta-level gains information about the object-level, making it essential to driving control processes (Chua, Pergolizzi, & Weintraub, 2014; Hacker et al., 2009). Monitoring occurs at multiple stages of metacognitive processes and is an important signature of self-awareness. Studying monitoring skills is not only prevalent in education. Metacognitive impairments and flaws in monitoring can signal a lack of self-awareness important in the study of many neurological and psychiatric disorders, e.g. dementia or schizophrenia (Bertrand, Landeira-Fernandez, & Mograbi, 2016; Cosentino, 2014; David, Bedford, Wiffen, & Gilleen, 2014). Interestingly, Bertrand et al. (2016) found that patients better recognised cognitive deficits from a third person perspective which supports the idea that switching perspectives in monitoring processes may be indeed an important valid aspect. Since meta-level processes are tied to the object-level, or cognition itself, monitoring cannot be studied in isolation. There is always a task requiring a certain kind of cognitive activity which constitutes the object of monitoring. Kelley and Jacoby (1996) studied biases in judgments of task difficulty of anagrams manipulating the prior experience of participants with similar problem-solving tasks. There is extensive monitoring research on signal detection, in which awareness of visual recognition is investigated; reading comprehension; and metamemory, where conscious control of memory processes are studied (Tobias & Everson, 2009, p. 110). These studies highlight the diversity of monitoring research and thus the difficulty

of defining the construct. As Tobias and Everson (2009, p. 107) states, in the research of metacognition "the method of observation defines the construct". While the specific context in which monitoring occurs is crucial and defines its nature, there is explanatory value in research from different areas and tasks. A common characteristic, however, is that the task of monitoring always involves some kind of judgment under uncertainty. It is usually measured in comparison to objective estimates of performance and participants' subjective judgment of that (Tobias & Everson, 2009, p. 109).

Research and literature around metamemory was found to be the most appropriate conceptual basis for this study. On the other hand, signal detection theory and research methods were most useful in model building. In the following section, the former is utilised to create a conceptual framework for the phenomenon under investigation. In the methods section, the signal detection paradigm is then explained in more detail, drawing on the concepts from metamemory. The third pillar of this thesis organizes findings from educational research within monitoring in multiple-choice tasks similar to this study's context and task environment, providing a basis for the two main hypotheses that the models operationalise.

3. Monitoring in multiple-choice tasks

3.1 Metamemory, monitoring, and calibration.

Research of metacognitive monitoring in education focuses mainly on judgment of performance (Isaacson & Fujita, 2006; Kelley & Jacoby, 1996; Nietfeld et al., 2005; Schraw & Dennison, 1994; Tobias & Everson, 2009). Studied monitoring processes may be in relation to problem solving activities, reading comprehension, or multiple-choice tasks of learned knowledge. Monitoring seems to be so important for academic success that it was shown to be able to predict dropping out of college better than anxiety measures (Tobias & Everson, 2009, p. 110). Thus, studying these processes and when they are suboptimal is crucial to understand students' practices and how to target them. With regards to reading comprehension, suboptimal monitoring behaviour gives rise to an *'illusion of knowledge'*. Sometimes students falsely judge the level of difficulty of the material learned and spend insufficient time and effort on reading the material (Pressley & Ghatala, 1990). Research highlights the importance of consistent information retrieval practices as successful strategies to (self-)test one's acquired knowledge objectively, however, it has not been reported as a frequent strategy among students, running the risk of an *'illusion of competence'* (Karpicke, Butler, & III, 2009). The task environment of the present study is a multiple-choice test. Pedagogically, it was implemented as part of the course to practice for one of the students' exams, which was based on their knowledge acquired throughout the semester, poten-

tially reducing illusions of (in)competence among students. The course, Introduction to Cognitive Science, is one of the foundational theoretical courses of the degree where students need to learn the vocabulary of the field, core concepts, and phenomena around cognition. Their first multiple-choice test provided data for the investigation. The test included questions from the course material with three alternative answers to choose from. After each question they were asked to rate their confidence level on a scale based on whether they think they answered correctly without receiving any feedback (refer to Methods).

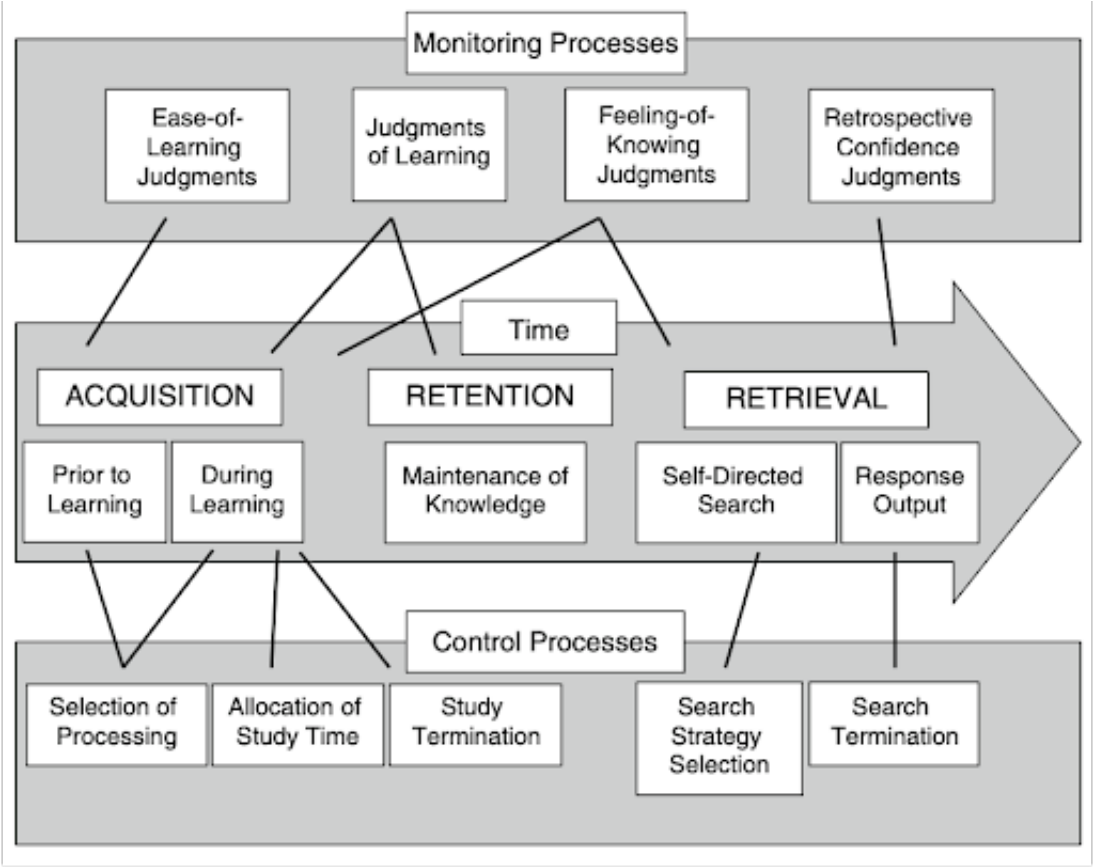

Figure 1

Figure 1. Theoretical framework of monitoring and control processes from metamemory. This figure shows the components of information acquisition, retention and retrieval in the middle, the cognitive activity that metacognitive processes are instantiated upon. On the top, metacognitive monitoring processes as observable behaviours (judgments) are shown, linked to the relevant cognitive process that they draw on. Similarly, on the bottom, regulatory control processes are depicted. From: Chua, Pergolizzi, and Weintraub (2014, p. 269)

The goal of the cognitive task in a multiple-choice test is to correctly retrieve relevant

information, assuming that information has been acquired prior to the test. Afterwards, when prompted to rate their confidence, students repeat the retrieval process in an attempt to find errors and/or points of uncertainty. To better understand these processes, research on and theoretical frameworks of metamemory was proven useful. In 1, a theoretical framework of monitoring and control processes in metamemory can be seen (Chua et al., 2014, p. 269). In the middle are the processes on the object-level, the cognitive activities that metacognitive processes draw on. On the top, monitoring processes are listed for all stages of memory processes, and on the bottom, control processes are shown. Monitoring processes are usually evaluated as the participants' confidence judgments against actual performance in the task. These measures are also referred to as *calibration*, i.e., how well the participant's perception is calibrated to their actual level of performance. In comprehension calibration, participants are asked right after information acquisition (reading, learning) to predict their ability to answer an upcoming question about the material (Nietfeld et al., 2005, p. 10). As calibration of comprehension takes place right after learning, that is, information acquisition, the related monitoring processes are thus judgments of learning. In comparison, calibration of performance is a retrospective confidence judgment about a question already answered by the participant. Another important distinction in the ways monitoring can be observed is whether participants are prompted to make judgments globally or locally (Chua et al., 2014; Schraw, 2009b). Global judgments of learning, also referred to as *offline monitoring*, are overall estimates made prior to or after taking an entire test consisting of many questions. These could be score predictions by the participant, i.e., how many out of the total number of questions that they think they will be/were able to answer correctly. On the other hand, local confidence judgments, or *online monitoring*, constitute confidence ratings after each and every item or question (Nietfeld et al., 2005; Schraw, 2009b). Within online calibration, Schraw (2009b) differentiates between confidence judgments and probability estimates of accuracy, but this is uncommon and most researchers use confidence ratings to evaluate monitoring (Fleming & Lau, 2014; Nietfeld et al., 2005). This study focuses on online monitoring observed through confidence ratings. Students indicated their confidence in their response, trial-by-trial after answering each multiple-choice question, on a four-point rating scale. Again, the initial plan was to study both online and offline monitoring, but ambitions had to be narrowed down. Nonetheless, some findings from research on global calibration (offline monitoring) are introduced below, as a comparison to clarify and better understand the online, local processes and how they are different.

3.2 Behavioural patterns found in monitoring.

Two key research papers are used to describe the behavioural patterns found in monitoring of multiple-choice performance in educational settings: Isaacson and Fujita (2006) and Nietfeld et al. (2005). First, none of them found evidence of improve-

ment of monitoring skills over time in multiple-choice contexts. They both argue that the mere subjection of students to these tests without any additional intervention is therefore insufficient in enhancing monitoring skills. Accordingly, the present investigation does not attempt to find improvement over time and focuses only on modelling the baseline, first-test monitoring skills. Isaacson and Fujita (2006) did not measure local calibration, only global, but they set up the test to allow students to choose their questions to a certain degree, based on varying levels of difficulty, thus studying control processes of strategy selection. Globally, they collected data on pre-test judgments (self-efficacy, pride goal, effort) and postdictions (judgments of scores after the test). Nietfeld et al. (2005) assessed local calibration and global postdictions. Isaacson and Fujita (2006) found that postdictions and performance were correlated, i.e., higher performance was associated with better monitoring skills. Higher achievers were more accurate in their calibration overall which was most striking on difficult questions. They were also better at judging whether the test was more or less difficult than the previous one, and their choice of questions were more deliberate based on their subjective estimation of skills to maximise their points. Students with good monitoring skills thus align their perception of question difficulty with the objective level of difficulty and act accordingly. Based on their choices of questions they also seem to be better at knowing what they do and what they do not know. The findings of Nietfeld et al. (2005) suggest that overall global postdictions tend to be more accurate than local ones across students of all performance levels. This also supports the idea that better performing students are more aware of what they know and what they do not. Performance was the best predictor of monitoring accuracy (calibration), better than overall academic ability (GPA or grades). Their results also highlight the importance of question difficulty which was the strongest predictor of calibration and bias (over- or underconfidence). Students were better calibrated on easy items than on difficult ones, and while they tended to be rather underconfident on easy items they became increasingly overconfident on more difficult ones.

The findings of Nietfeld et al. (2005) suggest two different processes of global versus local monitoring. It appears that the process behind global judgments is more analytic in nature. Especially with predictive judgments of performance, one would evaluate one's current stage in learning the material based on certain cues. Brunswik's lens model conceptualises probability judgments under uncertainty that can offer a broad account on the processes involved in global judgments and introduce the reader to the concept of cues (Hammon & Stewart, 2001, as cited in Newell, Lagnado, and Shanks (2015, pp. 27-28)). This model suggests that the object of the process is a criterion in the external world that needs to be matched by a judgment. This matching is made through the use of cues translated from the external world to the judge's mind. The abstract broad question of global judgments is 'how likely am I to succeed?'. To illustrate, when asked to predict how well they will score on a test, this individual might think of the time and effort they invested in studying prior to

testing or previous experiences with multiple-choice tests in similar contexts. When asked after taking the test, they might attempt to quantify their positive feelings toward recall in comparison to more uncertain events during the test, or overall how difficult they found the questions. These are all indirect cues related to the object of judgment helping to make an informed decision. The difficulty with postdicting scores lies in forgetting item-by-item experiences over time during the test. An 'ideal' judgment process would require the participant to keep all questions, their answers and momentary judgments about their answers in mind until the end of the test

In comparison, local monitoring is a more direct judgment by assessing monitoring right away. As proposed before, after answering the question, the participant can just repeat the thinking process for reaching a solution one more time to estimate its accuracy and make a judgment. To dive deeper into the processes behind local monitoring, *feeling-of-knowing (FOK)* theories were found to be useful (Koriat, 1993, 1994, 1995). As can be seen on Figure 1, FOK judgments are linked to both information acquisition and retrieval. FOK is an inference-based mechanism in which cues of information elicit memory and response (Koriat, 1994). A FOK theory, the *trace-access model* proposes a process of following cues even when they are misleading until the correct cue is found (Koriat, 1994). When the search terminates from failing to find an answer or a valid cue before a response is reached, according to the *accessibility account*, FOK is calculated online based on the partial information accessed during retrieval. Each cue contributes to a positive FOK until proven wrong or irrelevant (Koriat, 1994). FOK in retrieval processes draw on memory from the acquisition stages (Figure 1). Accessing memory storage is based on cues that lead to the piece of information in question. For instance, when trying to remember, one tries to recall the situation in which the material was studied, pictures in the book, feelings, or anything from the surroundings. Such cues are followed until falsified, i.e., proven irrelevant as in giving access to the answer of a different question or not giving anything. FOK theories also explain canceling-out strategies very well, as for each alternative answer the process becomes an attempt to falsify them until only one answer is left. The confidence judgment related to FOK becomes the 'sum' of feelings attached to each stage of the retrieval process.

Both Nietfeld et al. (2005) and Isaacson and Fujita (2006) found that levels of question difficulty influenced both the accuracy of monitoring and overall bias of confidence ratings. The relationship between calibration accuracy and question difficulty gives further support to the main pattern that performance and monitoring of performance are linked. Performance is not only affected by the participants' knowledge about the material but also the specific item's difficulty, thus question difficulty also influences calibration accuracy. Then underconfidence on easy items and overconfidence on more difficult ones are the biases describing the tendencies in responding. The findings that performance correlates with monitoring accuracy and question difficulty plays a sig-

nificant role constitute the behavioural patterns that are used to frame the hypotheses and research questions for modelling. Additional theoretical and conceptual accounts were given in this section, in an attempt to provide some explanations for and insight into the processes under investigation. In the following section, measures of online calibration and more precise accounts of the abovementioned patterns are introduced to provide a conceptual background of the main ideas expressed in the models.

4. Measurements of online calibration

The models built as part of this study are based on the found relationship between performance and calibration accuracy in online monitoring (Isaacson & Fujita, 2006; Nietfeld et al., 2005). As introduced before, this link is reasonable to make as metacognition draws on cognition itself; thus, task performance generated by the cognitive activity required to solve it provides the basis upon which monitoring is applied and produces calibration accuracy. Following these lines of thought, calibration should fluctuate in concert with the level of knowledge about the material and the skills required in doing the task. The first research question of this study, whether this relationship between knowledge and calibration can be modelled, and if so, to what degree will that model be coherent and descriptive of behaviour in comparison to alternative competing models.

Schraw (2009b) introduces two main families of measures of calibration accuracy: absolute and relative accuracy. *Absolute accuracy* can be measured through an index of discrepancy between confidence and performance, the Hamann coefficient of discrepancy between correct and incorrect judgments, and the bias index of overall degree of under- or overconfidence (ibid.). *Relative accuracy* may be estimated through correlation coefficients between judgments and corresponding performance scores, gamma coefficients of the dependence of confidence and performance, and the discrimination index indicating the ability in classifying correct versus incorrect responses (ibid.). All these measures describe calibration accuracy. In the next paragraph, the ideas of absolute and relative accuracy will be introduced through the Hamann and the gamma coefficient (Schraw, 1995).

In the simplest calibration task there is no confidence scale with multiple ratings. Instead, participants indicate their judgment on a binary scale whether they think they answered correctly or not. The outcome of the first cognitive task upon which they reflect is also binary as it may be correct or incorrect. Their trial-by-trial performance is then 0 or 1 for the first task, and 0 or 1 for the second task of monitoring judgments indicating their calibration accuracy. This produces a 2x2 table. There are four possible outcomes: hits (H) when they were correct and indicated so; *misses (M)* when they were correct but responded 'no'; *correct rejections (CR)* when they

were incorrect and indicated so; and *false alarms (FA)* when they were incorrect but responded 'yes'. *Agreement accuracy* is defined as the degree to which FOK judgments (yes/no) match correct/incorrect performance (Schraw, 1995). In conditional probability, this can be expressed as the probability of hits equals the conditional probability of success given a positive FOK (yes) (ibid.). Association between performance and confidence is then the probability of concordant trials, where judgment was correct about outcome (hits and correct rejections), minus the probability of disconcordant trials (misses and false alarms), divided by all possible pairs of outcomes (ibid.). The Hamann coefficient is a measure of absolute accuracy, or how well FOK matches reality. This score can take a value on a scale between -1 and 1. The Hamann coefficient is estimated additively as (ibid.):

$$\frac{(H + CR) - (M + FA)}{H + M + CR + FA}$$

The Hamann coefficient is the product of four mutually exclusive conditional probabilities and is independent of performance on the first cognitive task. In comparison, gamma, a relative accuracy measure, is the bivariate association between judgments and performance expressed as the difference between the joint probability of concordant and disconcordant trials (ibid.). It takes on the joint distribution of FOK judgments and performance. More specifically, gamma is calculated multiplicatively as:

$$\frac{(HxCR) - (MxFA)}{(HxCR) + (MxFA}$$

Gamma is a subjective internal variable conveying the relationship between confidence judgments and performance. However, if any of the four cases are 0 then it is undefined. Gamma can show whether associations between judgments are coherent. To illustrate the difference between the two measures with an example, if someone is asked to judge the weight of three bags one, they may be rather correct by getting close enough estimates (absolute accuracy). However, if a heavier bag is judged lighter than an actually lighter bag then judgments are not coherent internally (relative accuracy) (ibid.). Judgments can be off but still coherent. Both measures are important to assess calibration accuracy (Nietfeld et al., 2005; Schraw, 1995). The relationship between them is also interesting for the purpose of this study. When the two measures are equal, that means that accuracy and performance are independent, i.e., there is zero association between the two and performance is at chance level (Schraw, 1995). This implies that when there is no knowledge to generate performance above chance level, calibration accuracy is also random. This is the mathematical basis for the

model from a broad perspective to operationalise the link between calibration and performance in the task.

Absolute accuracy is usually calculated on an item-by-item basis as a squared deviation providing information about the degree of precision (Schraw, 2009b). The bias index indicates the direction of discrepancy between judgment and performance, signaling whether the lack of fit is due to under- or overconfidence (ibid.). In comparison, relative accuracy is an overall estimate on a set of judgments. Both performance and confidence on an item is compared to the respective means. Positive relationships indicate deviation to the same direction, thus internally coherent judgments. Relative accuracy focuses on the trend of confidence judgments relative to the trend in performance (ibid.). The discrimination index assesses the degree to which confidence judgments are able to classify correct and incorrect responses. It is a variation of relative accuracy (ibid.). In the case of a rating scale, the ability to discriminate does not depend on which part of the scale the participant prefers to use. What matters is whether they use consistently higher ratings for correct than for incorrect trials, producing a confidence threshold for the individual with respect to performance. A discriminability index, the *confidence judgment accuracy quotient*, is calculated as follows (Keren, 1991, p. 262):

$$\frac{Mean\,Confidence(correct : H\&M) - Mean\,Confidence(incorrect : FA\&CR)}{\sqrt{Pooled\,confidence\,variance\,of\,correct\,and\,incorrect}}$$

This measure is related to the *d' prime* in signal detection theory, however, d' prime is more of a hybrid score (Schraw, 2009b). The focus of this study, regarding calibration, is how well students can classify their correct and incorrect responses with their confidence ratings. This will be introduced later in more depth as the basis of the models. Discrimination indices capture confidence across items unlike absolute accuracy or bias which are estimated trial-by-trial. It is important to note that all these different accuracy measures describe monitoring skills together. Building computational models in this study involved a process of reverse-engineering calibration accuracy. When simulating behaviour through modelling relationships between variables, the introduced accuracy measures informed the architecture and the reasonings behind the models instead of using the calculations to describe collected data.

Plotting *calibration curves* is a common practice in describing data from monitoring tasks (Keren, 1991, p. 221). Percentage of correct responses indicating performance are plotted against confidence ratings (refer to Figure 2 as an example from Keren (1991, p. 222)). Overconfidence with low performance and underconfidence with high performance are biases found across multiple tasks including multiple-choice tasks

(Isaacson & Fujita, 2006; Keren, 1991; Nietfeld et al., 2005). For example, Figure 2 shows similar tendencies in two different task environments, comparing data from a perceptual task and a general knowledge task. Isaacson and Fujita (2006) argue that these effects of lower and higher achieving students may be due to self-protective perceptions to protect self-esteem or academic self-concept (Dembo Jakubowski, 2003 as cited in Isaacson and Fujita (2006); Bol, Hacker, O'Shea, and Allen (2005)). As mentioned, similar patterns to the ones shown in Figure 2 have also been found for question difficulty as well as overconfidence on easy items and underconfidence on more difficult ones, also termed the hard-easy effect (L. A. Brenner, 2003; Keren, 1991; Nietfeld et al., 2005; Suantak, Bolger, & Ferrell, 1996). In this study, to reason about the processes producing increasing overconfidence by increasing question difficulty, intially, my intuition was based on FOK theories and the availability heuristic (Newell et al., 2015, p. 87). To illustrate with an example, when a participant cannot remember anything concrete about a given question but one option seems familiar with cues they recognise but cannot entirely recall, overconfidence is likely due to availability heuristic: "Since that is familiar, that must be the correct answer." The second research question of this study is whether these hard-easy effects of question difficulty can be modelled, and if so, to what degree will this model be coherent and descriptive of behaviour in comparison to other alternative models.

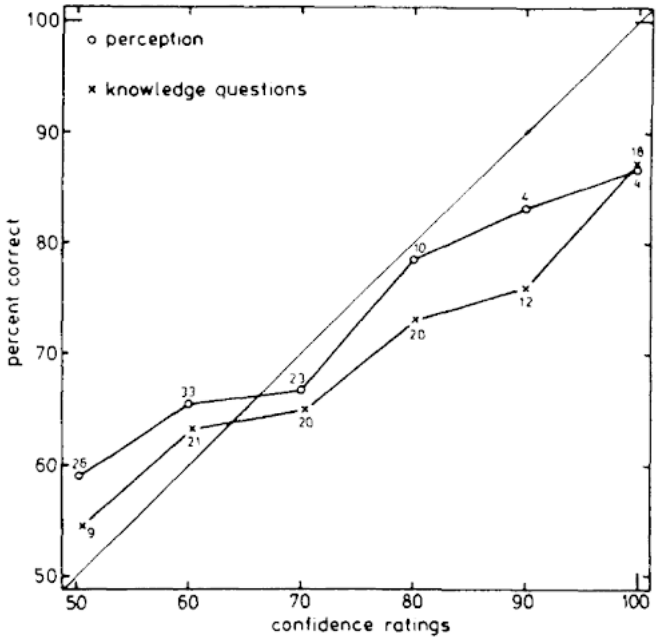

Figure 2. An example calibration curve. A common practice in assessing accuracy and patterns in confidence ratings in monitoring research is to plot performance on the y-axis (percentage correct) and confidence rating on the x-axis (as in percentage ratings). From this graph, under and overconfidence patterns can be easily gauged. Two datasets are plotted: one from a perceptual task and another from a general knowledge task. From: Keren (1991, p. 222)

Research questions

The overarching philosophical, conceptual question of the thesis is about the nature of metacognition, more specifically in the context of online, local monitoring in multiple-choice tasks. As proposed in the Introduction, metacognitive processes may be sequential upon prompting, repeating the cognitive activity from a 'third-person' perspective in an attempt to search for errors in the retrieval process based on feeling-of-knowing traces (Bertrand et al., 2016; Chua et al., 2014; Koriat, 1994). Prompting is given in the task environment as students are asked to rate their confidence. Repetition of the cognitive process monitored is conceptually built into two models described in the Methods section, inspired by the accuracy measures by Schraw (2009b) and the found link between monitoring and performance, to test in comparison to a null-hypothesis model (Isaacson & Fujita, 2006; Nietfeld et al., 2005). Findings may offer insights into the nature and process of metacognition and monitoring.

The present study specifically focuses on the local online calibration of students as part of their metacognitive monitoring skills. As described in the Introduction, two different qualities of monitoring have been found in previous research associated with global (offline) and local (online) calibration, suggesting two distinct processes of monitoring judgments (Nietfeld et al., 2005). Global judgments of performance, measured as predictions and postdictions, are likely based on an analytical cue-based probability judgment in which students consider indirect cues outside of, but related to, the material, such as the amount of time dedicated to study, prior successes in similar contexts, etc. (Isaacson & Fujita, 2006; Nietfeld et al., 2005; Schraw, 2009b). Local judgments, also referred to as online calibration, however, were suggested to stem from a different process based on the findings of Nietfeld et al. (2005). According to their results, while all students were rather accurate in their global judgments of their performance, those with higher performance were better at local online calibration when asked after each question to rate their confidence in their perceived correctness. Regardless of local or global monitoring skills, performance has been found to be the strongest predictor of monitoring and is the main pattern set out to be operationalised and investigated through cognitive modelling in this thesis (Isaacson & Fujita, 2006; Nietfeld et al., 2005). Thus, the first research question is: **1) Is it possible to build a model to meaningfully capture the relationship between performance and online calibration? To what degree will that model be internally coherent and fit real data in comparison to alternative models with no connection between performance and monitoring?**

The second pattern set out to be investigated here, found in previous research, was that the difficulty of questions seems to play a significant role in the accuracy of calibration and confidence bias of monitoring (L. A. Brenner, 2003; Isaacson & Fujita, 2006; Keren, 1991; Nietfeld et al., 2005; Suantak et al., 1996). More specifically,

increasing question difficulty decreases calibration accuracy, and responses become more overconfident in nature. The second research question directly targets the effects of question difficulty on monitoring: **2) Is it possible to build a model to account for systematic biases based on question difficulty? To what degree will that model be internally coherent and fit real data better than alternative, simpler models?**

Conceptually the present study attempts to synthesise findings, methods and concepts from three research areas of metacognition: 1) educational research in monitoring and self-regulated learning by students, 2) assessment methods of monitoring judgments as calibration accuracy in metamemory, and 3) signal detection paradigm to model recognition and associated metacognition. The thesis has two main aims:

- 1) to explore the potential of cognitive modelling methods in an educational setting for research (the first parts of the two research questions), and
- 2) to examine whether previous findings of calibration and monitoring in educational research can be systematically modelled and captured as universal processes underlying the observed behaviour in the task, supported by theoretical constructs and cognitive frameworks introduced (the second parts of the two research questions).

This will be achieved through building cognitive architectures that reproduce hypothesised behavioural patterns based on previous findings from research (as introduced earlier). The first aim relates to the process of modelling attempts (refer to the next section, Methods), while the second focuses on the outcomes of model building (refer to Results).

Methods

1. Data collection and task environment

Data was collected from freshmen bachelor students in Cognitive Science at Aarhus University, Denmark. The language of the Cognitive Science programme is English and 'English as a foreign language (EFL) students' need to present proof of language skills as part of the application process (Studieguide, 2021).Students of the course, Study Techniques offered to first-semester bachelor students at Cognitive Science, were asked to participate. In the beginning of the semester, 48 students indicated interest in participating in the study and filled out the demographics questionnaire. Of the 48 students, 71% of them were female and 29% were male, thus comprising around a 1:2.5 male-to-female ratio. They reported ages between 19 and 34 with a median of 21. Ten unique nationalities were identified. The majority, 75%, were Danish. At the time of data collection, the scope of this thesis was planned to be larger with multiple sources of data collected from the students. The demographics questionnaire was completed in the first class as a basis for all planned data collection. The object of the thesis was narrowed down to investigating only one type of data from one session with the most data (donation from 38 students). Participants gave written consent following the university's ethical guidelines (Lab, 2021). They were informed that they may withdraw their consent any time without any consequences. No personally identifiable information was collected from the participants who were completely anonymous to the researcher. They had to indicate a participant ID that they came up with themselves. Additionally, since the instructor was the researcher and the test was also part of the teaching activities, it was deemed important to apply an additional free data donation principle where students were reminded of their free choice in participation or withdrawal before each session of data collection. Participation in the study was allowed to be used as the mandatory experiment participation required in one of their first semester courses, Introduction to Cognitive Science.

The students completed several multiple-choice tests over the course of the semester as part of their Study Techniques course. There were no consequences of scores in these tests, thus no incentives to perform well, which will be discussed later as a potential limitation of the study. Over the semester, students completed three multiple-choice tests. After each question, participants were asked to rate how confident they were that their selected answer was the correct one on a 1-4 scale with 1 indicating the lowest confidence and 4 the highest (Fleming & Daw, 2017; Fleming & Lau, 2014). All multiple-choice questions were created by the students; they were asked to individually send a multiple choice question to the instructor before class in relation to the topics they had already covered at the course up until the time of the test. An

example of the multiple choice questions and the related confidence rating scale can be found in Appendix A. Participants were not given explicit feedback on their scores, only a PDF file containing all questions with the correct answers marked. From the pedagogical perspective of the course, all students in the class completed the tests to train their recall of knowledge in preparation for their exam. However, only study participants were asked to additionally rate their confidence for each question. The multiple-choice tests were carried out using SurveyXact. Additional tests were prepared for non-participating students without confidence scales. Data from these were deleted right after class.

Responses to the first multiple-choice test were included for analysis. As mentioned, 38 students donated their data from this test consisting of 24 questions (as not all students remembered to send a question before class). Data from the rest of the two tests were not analysed for the following reasons: Monitoring skills have been found to be rather fixed and unaffected by training in the task or feedback (Isaacson & Fujita, 2006; Nietfeld et al., 2005). At the same time, other course activities could have affected students' monitoring skills and thus introduced biases too complex to account for. Additionally to local judgments of confidence, participants were also asked to make global judgments before and after taking the test: subjectively rating their monitoring ability as a percentage (Schraw & Dennison, 1994), predicting and postdicting their score and giving estimates of their effort (hours of studying) similar to the study of Isaacson and Fujita (2006). However, as the scope of this thesis was narrowed down to only assessing local online monitoring these global judgments were not analysed.

The following main characteristics of the task environment are important to keep in mind for the upcoming sections. First, students came up with the multiple-choice questions themselves; thus, question difficulty varies immensely item-by-item and from person to person. 24 questions were collected from the students and data comes from 38 students. This indicates that some of the 38 students knew the answer to at least one question, but not everyone knew at least one. Second, students made their choices between three options, making the odds of a chance performance in this task 1/3. However, after making a choice they most likely implicitly assessed whether the answer was right or wrong (0 or 1) to rate their confidence on a 1-4 rating scale, which together constitute one step in the task.

2. Cognitive modelling

Computational modelling of behavioural data is the process of building precise mathematical models to generate observed behaviour in specific tasks (Wilson & Collins, 2019). In this study, observed behaviour is performance (correct/incorrect responses)

and choices (1, 2, 3, or 4 on the confidence rating scale). Models are algorithmic hypotheses of the ways behaviour might be generated. Underlying causes of behaviour are formulated through latent, unobservable variables or parameters and certain specific processes expressed by mathematical equations to produce behaviour that can be observed. The psychological processes and latent variables cannot be directly measured, only inferred by their impact on overt behaviour (Heathcote, Brown, & Wagenmakers, 2015). Task performance is usually an end result of multiple unknown combinations of processes that need to be decomposed to build a cognitive process model. Cognitive modelling is generally carried out based on the principles of Bayesian data analysis as the Bayesian paradigm naturally allows for complex, flexible, but also parsimonious, specific models that convey uncertainty (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). The uncertainty around parameter values is expressed by probabilities and probability distributions (Lee & Wagenmakers, 2014a, pp. 3-4). Observed or simulated data is then used to update prior distributions to become posterior distributions through abiding probability theory by using the data as evidence (Wagenmakers et al., 2008). Cognitive modelling was used in this study to describe latent processes underlying the behavioural data and reason with regards to the research questions associated with the hypothesized cognitive processes.

To investigate calibration skills of students in the present task environment in light of the research questions, the following principles and procedures were applied for ensuring a sound practice of cognitive modelling. First, all models built were based on previous work. Priors for parameters were chosen based on the origins of each model component from the selected previous studies, referenced respectively. Other unique additions were included based on the conceptual background of each processing module (model component) with regards to this study's specific task environment and the unique purposes of each model, highlighted and explained when relevant. Model components, their origins, choices made, and reasonings behind will be introduced in the next section. Second, each model was chosen based on the alternative, competing hypotheses they capture. These two principles together ensured that models are built explicitly to test processes in question without trying to capture everything that is ongoing in cognition throughout the task. Models were kept to be as simple as possible with their purposes in mind to test the potential hypotheses (Wilson & Collins, 2019, p. 6). Referring back to the research question, a model was built to account for random behaviour, suggesting either no engagement with the task or completely different processes than hypothesised. Another model captured the opposite of the first hypothesis, that there is no relationship between performance and monitoring, i.e., monitoring being independent of performance. The other two models operationalise the two main research questions based on previous research: 1) that performance and calibration are linked, and 2) the role question difficulty plays on confidence bias in calibration. The four models will be introduced later in this chapter in more detail. Inference was conducted using Markov chain Monte-Carlo

sampling with JAGS software, implemented through R using the R2jags package (Plummer, 2003, 2012). In the sampling, 5000 iterations were run, three chains were specified, and the first 1000 samples were discarded ('burn in').

Three studies were carried out to compare the four alternative models describing behaviour: parameter recovery, model recovery, and descriptive adequacy or posterior-predictive checks on real data. *Parameter recovery* is the procedure of simulating data from the widest meaningful ranges of parameter values, then fitting the simulated 'fake' data by the same model and checking whether set and inferred parameter values correlate. An internally coherent model should have recoverable parameters, i.e., set and inferred parameter should correlate (Wilson & Collins, 2019). For this, set and inferred parameters were plotted against one another and to quantify the correlation Pearson correlation tests were carried out in R. Each model was run forward and fitted 100 times. *Model recovery* is the process of again simulating data from all models with a meaningfully wide range of parameters, then fitting all those sets of data by all models to see whether the particular model used to simulate a data set is also able to fit the same data set the best out of all competing models. Models with the lowest deviance information criteria (DIC) scores were considered to 'win'. 100 simulations were run from each model and fitted by all models producing a confusion matrix with counts of how many times out of 100 each data set was fitted best by the four models. *Descriptive adequacy, or posterior-predictive checks*, is the process of taking the posterior distributions of the models trained on the real data and, based on them, predict the observed data to see how well the models fit the data. The aim of this procedure is to find the model with the best predictive performance (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Overfitting is not a concern in Bayesian cognitive modelling as it comes with a built-in automatic Ockham's razor (Wagenmakers et al., 2008, pp. 13-15). Since model selection is based on marginal probabilities of the data, given the model, simple models are always favored (Wagenmakers et al., 2008, pp. 13-15). In this study, the four models were used to make predictions of ratings (1-4) on each trial for each participant then compare with the actual observed data. Again the models' performances were evaluated in light of each other and their differences.

3. modelling behaviour in the task and signal detection theory

3.1 Simple signal detection theory on 'Yes/No' tasks.

The aim of the present study is to model students' calibration in multiple-choice tasks when making judgments about their momentary accuracy, i.e., how well they can differentiate between their correct and incorrect responses. To operationalise the processes in question, the signal detection paradigm was chosen. In this section, the

theoretical foundation for the models, based on the signal detection paradigm and certain modifications of it, will be introduced.

Signal detection theory (SDT) is used to model processes behind many tasks in cognitive psychology where participants need to discriminate between two kinds of stimuli (Stanislaw & Todorov, 1999). In a recollection memory task, participants are asked to study a list of words after which they are then tested by being asked on a trial-by-trial basis whether a given word was part of the studied list or not ('old' versus 'new' response) (Pratte, Rouder, & Morey, 2010; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). In this testing session, there are two kinds of trials of stimuli. *Signal trials* in which the presented word was in fact part of the studied list and the correct response is 'old', and *noise trials*, in which the word was not part of the list and the correct response is 'new'. Information about signal and noise trials are available to the experimenter but unknown to the participant, i.e., there is no feedback on whether they answered correctly or not.

In the task environment of the chosen multiple choice task, students are presented with a question and three alternative response options out of which they are instructed to find the correct one. Signal trials are considered to be those in which students chose correctly, while noise trials constitute the incorrect responses. In this case, it is more meaningful to call these two kinds of events *correct (C) and incorrect (IC) trials*. This information, similar to signal and noise trials, is only available to the researcher but not to participants as they do not receive direct feedback on their performance at any point during the task. They solely rely on their own judgment of accuracy. The main source of data underlying the process in investigation by the SDT is the confidence rating given by the students after their choice of answer. Task 1, the knowledge task, constitutes the event (*stimulus* in traditional SDT tasks) which is subsequently evaluated by participants in Task 2. Task 1 is not modelled explicitly, i.e., participants' choice of answer A, B, or C, is not investigated in depth, because it is not the focus and it is more complex, involving many processes irrelevant to the purposes of the current research. Only the outcome of task 1, indicating C or IC trials, is modelled and considered.

Signal detection theory hypothesizes that participants have internal appraisals to signal and noise trials, respectively (Selker et al., 2019). These are modelled by two overlapping distributions (refer to Figure 3). The extent to which they overlap indicates the participant's optimal sensitivity to signal trials in contrast to the noise trials. This is captured by the distance between the two means of the distributions, the *d' prime, the optimal discrimination ability*. Together with the *bias* (λ) term they capture the observed behaviour: the amount of correct responses and errors that the participant makes. The basic paradigm has four parameters: the *signal distribution*; the *noise distribution*; their distance from each other, reflecting the likelihood of the

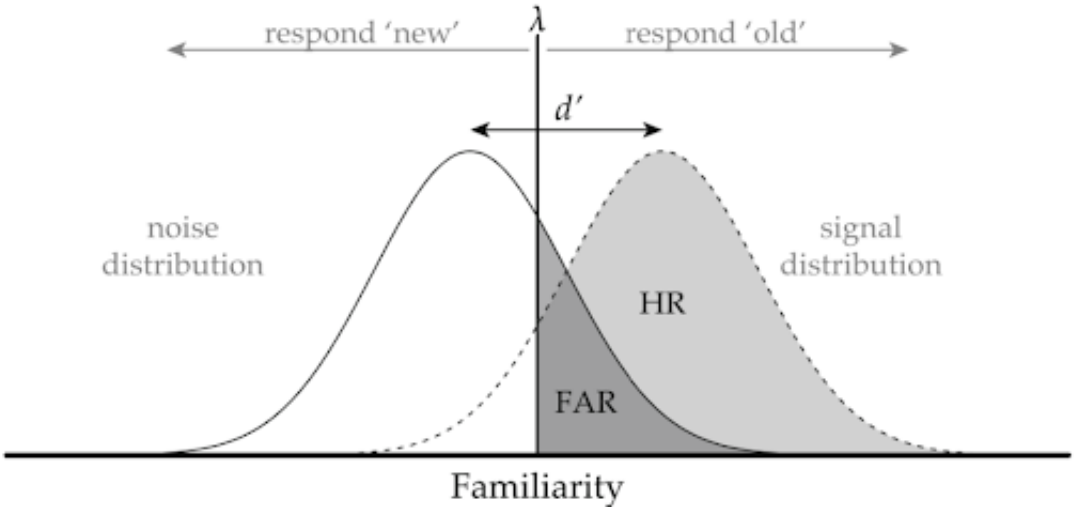

Figure 3. Basic Signal Detection Paradigm. Two distributions are shown as the two different internal appraisals towards signal and noise trials. Difference between the two means is the d' prime parameter reflecting optimal discrimination skills. Lambda is the individual bias parameter in relation to d', and together they describe behaviour in the task, observed hit and false alarm rates. Hit rate (HR) is the area under the curve of the signal distribution above the bias minus the false alarm rate (FAR) which is the area under the noise distribution curve above lambda. Illustration taken from: Selker, van den Bergh, Criss, and Wagenmakers (2019, p. 1954)

participant answering 'old' or 'new' given their 'true' sensitivity (i.e. how far the two means of the distributions are, how much they overlap; d' on Figure 3); and a bias term that describes the individual bias in the participant's propensity to answer 'old' or 'new' (λon Figure 3). The areas under the curves above the bias threshold account for the *hit rate* (HR = amount of correct 'old' divided by all 'old' responses) and *false alarm rate* (FAR = amount of incorrect 'new' divided by all 'new' responses). Hit rate is the probability captured by the area under the signal distribution curve above the bias minus the false alarm rate, the area under the noise curve above bias, together fully describing the data. The latent variables of optimal sensitivity and associated bias describe the psychological processes underlying the behaviour in the task environment. There are two kinds of correct responses to the stimuli: a recognition of signal trials (hits) and a rejection of noise trials (correct rejections is the area under the noise distribution curve below the threshold lambda). Accuracy of the participant is then described and visualised with the *receiver operating characteristic (ROC)* curve, where hit rates and false alarm rates are plotted against each other, and the *area under the curve (AUC)* gives an accuracy score between 0.5 (chance) and 1 (refer to Figure 4).

In the multiple choice task environment, whether students answered correctly or in-

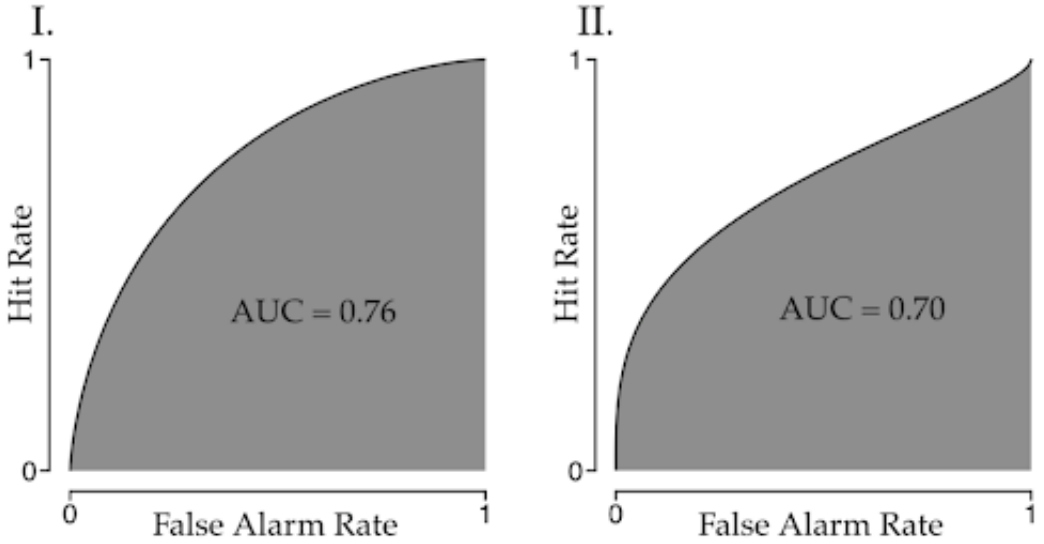

Figure 4. Illustration of the AUC (area under the curve) of the ROC (receiver operating characteristics) curve. By plotting hit against false alarm rates an accuracy score between 0.5 (chance) and 1 (perfect accuracy) can be estimated as the area under the curve. Illustration taken from: Selker et al. (2019, p. 1958)

correctly is known by the researcher but not the students. Students judging whether their choice was correct or not therefore corresponds to their optimal sensitivity to their performance, which in this case is the optimal accuracy of their calibration. In other words, how well were they able to distinguish between their correct and incorrect responses in theory regardless of bias, or how well they know what they know and what they do not know from the material that the questions were generated from. From a cognitive process perspective, this may be the stage when they recall their trails of thoughts answering the multiple-choice question in task 1. Since students do not receive feedback about their performance, they are making judgments under uncertainty, relying on their skills of task comprehension that include verbal comprehension of the questions and potential answers as well as understanding the concept of multiple choice tasks that they need to find the one correct answer. Most importantly for this study, students rely on their knowledge and memory of the material that they are tested on. For the purpose of simplicity, students are assumed to be competent in their task comprehension, i.e., their understanding of English and the multiple choice task. The model focuses on the students' awareness of their knowledge and related performance, the object of the monitoring paradigm, their optimal metacognitive sensitivity, and bias. This is operationalised in line with the signal detection paradigm. It is hypothesised in this thesis that there are internal appraisals to correct and incorrect trials that reflect students' skill level (knowledge of the material), directly observable in their performance. The further away the means of the two distributions are, the better they know the material; and the more they know,

the more they are aware of that, in theory, regardless of bias. This has been found by Nietfeld et al. (2005), showing that the better students knew the material, the better they were at knowing in detail what they did and did not know in online calibration. Thus this thesis speculates that participants' monitoring sensitivity, i.e., the distance between the two distributions, is directly related and proportionate to their skill level which has to be between 1/3 (guessing) and 1 (perfect performance) in the chosen task environment (with three options to choose from). The premise of this relationship is that students' perception of their knowledge directly corresponds to their 'true' underlying knowledge of the material. In cognitive process terms, outcome on the multiple-choice task—the performance—is generated through a cognitive activity that is repeated flawlessly in the judgment task. If an error occurred in the recall of knowledge relevant to answer the task then that will be repeated. This also indicates that without bias, participants are perfect calibrators, lying on the diagonal of the calibration curve (Figure 2). This is a reasonable starting point for operationalizing the patterns as over- and underconfidence can be considered individual biases. It is also safe to assume that the study's population can recall mental processes immediately. This would not necessarily be the case for other populations.

To connect with Schraw (2009a, p. 35) measures of metacognition, d' prime, optimal metacognition is closely related to the discrimination index, while the scatter index, variability of judgments for correct and incorrect responses, may be related to the variance of the two distributions. The confidence judgment accuracy quotient includes both as shown in the Introduction (Keren, 1991, p. 262). Bias can be captured by the bias lambda of this simple yes/no SDT confidence account, while relative accuracy becomes relevant with the rating scale of confidence. The d' prime, relative thresholds, and biases together form a hybrid account of calibration (Schraw, 2009b).

3.2 Equal versus Unequal variance SDT.

Signal detection with unequal variances of the two underlying distributions corresponding to correct and incorrect trials along with other alternative versions of signal detection have been considered. However, equal variance signal detection was chosen as a basis upon which to build the current architecture of calibration in relation to multiple-choice responding (DeCarlo, 2010; Wixted, 2007). The decision was made based on hypothesized underlying processes in the presented task in comparison to a recollection memory task. In recollection memory, two accounts exist for two distinct psychological processes driving the observed behaviour—*familiarity* and *recollection*— supporting an equal variance as opposed to an unequal variance account of SDT, respectively (Wixted, 2007). Recollection suggests that when participants see a word, they either consciously recollect it from memory—that it was indeed part of the studied list—or they immediately reject the word by not recognising it from the studied list (ibid.). Reaction to the item needs to exceed a certain threshold for the

conscious recall of the word to happen. This supports the unequal variance version of the SDT as only signal trials elicit conscious effort and thus are more prone to error (increased variance of the signal distribution). In comparison, recognition of noise trials involving only a clear-cut rejection is more stable (usually modelled as a standard normal distribution with a mean 0 and standard deviation of 1). Familiarity, on the other hand, hypothesises that participants make a decision based on a discrete level of familiarity response to the word (ibid.). Some words are more or less recognised as being 'old' or 'new', involving a similar amount of conscious effort, thus producing equal variances around both signal and noise distributions.

Cognitive processes underlying multiple-choice responding and monitoring, however, are more complex and not that clear-cut. Keren (1991, p. 262) also notes that stimulus recognition and related monitoring is not the same as response classification and monitoring. In the first task of choosing the correct answer there is definitely a conscious recollection of the material, but when judging one's own performance, an incorrect response can elicit just as strong of a correct, low-confidence response as a certainly correct one eliciting high confidence. Imagine, for example, the situation when, after making a choice, one realises that a different option was the right one and not the one they ended up choosing. This can happen because in the second task of making judgments under uncertainty, participants likely recall their thought processes and reasoning for a second time and might recognise errors in them. Especially when choices overlap to a high degree and they need to pay attention to and recollect the details around the topic in question. As the cognitive effort involved in the first task is increasingly high, the monitoring task likewise demands more cognitive capacity. There is no clear rejection of incorrect trials as opposed to correct trials as the recollection account of SDT suggests. Not to say that it can never happen that a student does not remember anything about a question, but it is not a rule for all trials that a question needs to elicit a certain threshold to allocate more processing power and be remembered, thus being more likely to be correct. It would be false to say that monitoring of an incorrect trial is a pure rejection in this context. Based on these reasons, two processes are hypothesised, both corresponding to an equal variance SDT model: 1) A recollection account that is equal to both kinds of trials (correct and incorrect), and 2) a familiarity account of less knowledgable (partially guessing) participants or with easier questions. In the latter, judgments are made on a continuous scale of familiarity like the familiarity account of SDT with equal variance. However, alternative versions of the paradigm can be appropriate, especially in light of Scraw's scatter index as a measure of variability towards correct and incorrect trials (Schraw, 2009b). Still, the traditional unequal variance account—fixing the noise distribution's variance at 1 and only allowing the signal distribution's variance to vary—is conceptually problematic. Then both should be allowed to vary. For the sake of simplicity, the equal variance SDT was chosen.

The recall of the thought processes involved in the first task and making judgments, as to potentially detect errors in them, is the focus and object of this study and what is being modelled, operationalising the first research question. The main hypothesis is that knowledge, measured as performance, drives optimal monitoring skills, i.e., the potential classification of own correct and incorrect responses. The thought processes underlying both tasks are essentially the same, only in the second task the participants' perspective likely changes from doing the task to a third-person evaluation of doing the task. The more they know about the context—the material itself—the better they are at judging their own performance (Nietfeld et al., 2005). This is operationalised in the model in a top-down approach. Accuracy, the area under the ROC curve (Figure 4), is usually estimated and plotted from data and subsequently inspected. In this model, however, accuracy is directly estimated from a hypothesised prior skill level. From this, a monitoring sensitivity parameter is calculated utilising the documented relationships between them in equal variance SDT in which such d' assumptions hold (DeCarlo, 2010; Stanislaw & Todorov, 1999). Traditional SDT model applications consider these parameters separately. However, in this model, participants themselves generate their own 'signal' and 'noise' trials; it is not a setting of the experiment, like in a memory recollection task, towards which objective all participants relate. They react to their own data which varies from person to person through how many correct or incorrect trials they make. The additional effects that participants can have on their confidence ratings are their self-esteem and related perception of skill, such as under- or overestimating their own knowledge and their mental model of the 4-point rating scale, all of which affect accuracy and are modelled with individual biases (scaling and shifting of the scale, discussed below). In summary and in light of research question one, if the model alternatives deriving the d' prime from skill perform better than other model versions, that may indicate that the hypothesised process repetition underlying task 1 and then repeated in task 2 might be a valid account of what is happening in cognition.

3.3 modelling the 4-point rating scale.

Metacognitive accounts of signal detection theory have been proposed by: Fleming and Lau (2014); Maniscalco and Lau (2014). In these paradigms, the 'old'/'new' responses of stimulus classification is referred to as the type 1 task, while the following confidence ratings given to the participants are the type 2 tasks, in which they need to indicate their confidence in their response in the type 1 task (Maniscalco & Lau, 2014, p. 26). The multiple choice task, however, is different. There is no type 1 task in this sense, as there is no stimulus to be classified on a binary scale reflecting one distinct process like recognition or recollection in the chosen example. Instead, processes in task 1 are more complex, involving verbal comprehension, attention, memory, and being more nuanced with three choices. On the other hand, task 2 involves both type 1 and type 2 tasks similar to their original sense. The part of the experiment oper-

ationalised by the SDT is the classification of own correct and incorrect responses. Students need to make a judgment whether they think they answered correctly or not, not as a binary choice but on a rating scale factoring in some additional uncertainties. In this process, type 1 task, classification of own correct/incorrect responses, is already a metacognitive account (type 2 task in the original paradigm), while the chosen confidence scale allows for more nuanced responding. The binary judgment of performance perception ('was I correct or incorrect') is implicit and not separately elicited but instead incorporated in the confidence ratings.

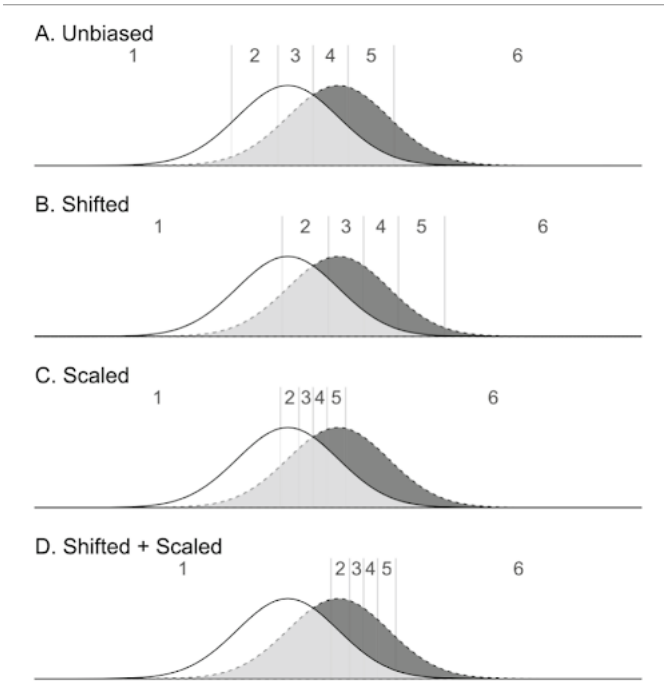

Figure 5. The variety of scaled and shifted confidence rating scales on the underlying signal/noise distributions generated by Selker et al.'s linear regression model. Parameters of the linear regression model (intercept and slope) can account for diverse biases on a rating scale, in this illustration a 6-point Likert-scale; shifts towards either side preferring lower or higher ratings and scale reflecting perception of the distance between individual ratings preferring the middle or the ends of the scale. Illustration taken from: Selker et al. (2019, p. 1957)

In a monitoring or calibration version of the signal detection recall task, after indicating whether the word was part of the studied list or not, participants are additionally asked in a separate question to rate on a scale how confident they are in their response. Such a metacognitive element has been explored in Maniscalco and Lau (2014), and Fleming and Lau (2014). The main challenge has been the modelling of sensitivity corresponding to each rating and related biases that requires the estimation of many parameters (a d' prime and a λfor each threshold). Selker et al. (2019) propose a linear regression to model the threshold parameters together and produce

different biased thresholds (λ-s) corresponding to each rating. Here, d' prime is an overall parameter that together with the multiple bias thresholds (λ-s) produce the calibration accuracy observed. The linear modelling approach can produce a diverse layout of confidence scales (refer to Figure 5 for examples with a 6 point rating scale) and it was found to be useful in the current monitoring paradigm as students can be very different in their individual mental model of the rating scale. The focus of this study, regarding calibration, is how well students can distinguish between their correct and incorrect responses regardless of whether they prefer to use the far ends of the scale (1 and 4), the middle (2 and 3), or either sides of the scale (1 and 2, or 3 and 4). As long as the participant is consistent with their ratings corresponding to either cases (correct/incorrect trials), their monitoring sensitivity should be optimal regardless of exactly what rating they have been choosing. It also captures individual differences in the perception of the distance between the ratings, which might likely vary. These ideas are also in line with Schraw's relative accuracy (correspondence between confidence and performance, internal coherence), discriminability (classification of correct and incorrect performance), and bias (tendencies for over- or underconfidence) (Schraw, 2009a). Bias is the shift parameter, the intercept of the linear regression, accounting for overall under- or overconfidence. In the model, accounting for question difficulty, it was added to modify this bias on a trial-by-trial unique basis (refer to next section).

The linear regression sets each threshold between the confidence ratings individually. Then, the probability of the certain rating, following probability theory—given that the outcome of task 1 was correct or incorrect—is the area under the correct/incorrect distribution curve below or between the threshold and the previous one, respectively. Similarly, hit and false alarm rates corresponding to each threshold will be the area under the correct and incorrect distribution curves above each threshold (Selker et al. 2019). These corresponding hit and false alarm rates plotted against one another produces the ROC curve (refer to Figure 6).

3.4 The competing models investigating the research questions.

MODEL 1: RANDOM RESPONDING

The first model, random responding, serves as a 'sanity-check' (or, the all-encompassing null hypothesis). Conceptually, such a random responding model is commonly used to account for participants not engaging with the task; not an uncommon phenomenon, especially in tasks with no incentives, like the presented one (Wilson & Collins, 2019, p. 7). Apart from implicit motivations there were no explicit incentives rewarding the cognitive effort required to engage with the task. Skill, the latent variable of knowledge of the material, was still specified to produce varying levels of performance between chance (1/3) and perfect performance (1) as the outcomes of task

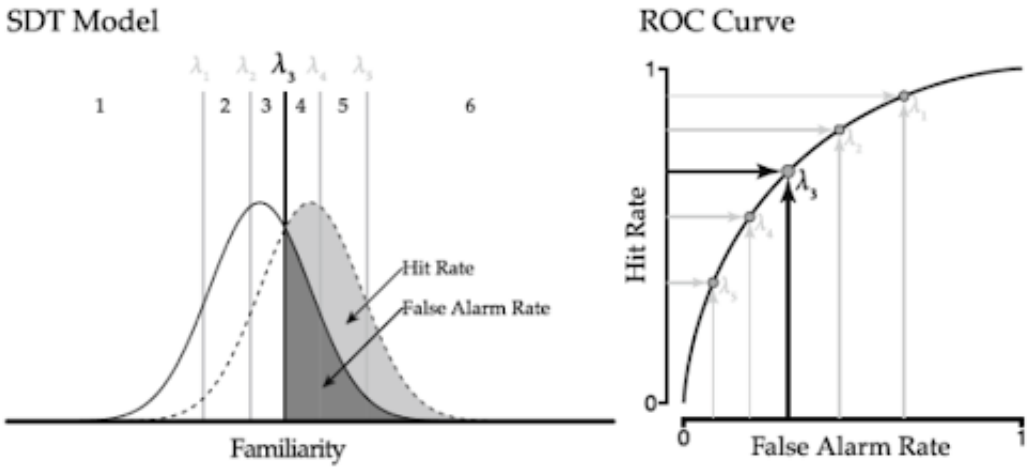

Figure 6. Thresholds produced by the linear regression, related hit and false alarm rates, and the ROC curve. Thresholds between ratings constitute the multiple lambdas in connection to the two underlying distributions. For each threshold a hit and a false alarm rate can be calculated and plotted against one another to observe the ROC curve. Illustration taken from: Selker et al. (2019, p. 1957)

1 (correct and incorrect answers). This model could also verify whether the first component, skill generating performance, is valid in terms of parameter recovery of the latent variable. Confidence ratings on the 4-point rating scale, however, was just a random number generator, a categorical distribution of the 4 ratings with equal underlying probabilities, disregarding even a preference for a certain rating. In model comparison, success of this model could also indicate a completely different underlying cognitive process and behavioural patterns than specified by the research questions and hypotheses.

MODEL 2: UNRELATED SKILL AND D' & MODEL 3: D' FROM SKILL

Figure 7 is a graphical representation of the structure of two models. The outcome of each trial (Ot) is directly generated from an underlying latent variable (S = skill), unique and descriptive of the individual's knowledge of the material. The prior chosen to initialise S is a uniform distribution between 1/3 (chance with three options) and 1 (perfect accuracy) reflecting the naïve guess that any skill level in-between is equally likely ($S \sim U(\frac{1}{3}, 1)$). S generates Ot following a binomial distribution to update the underlying latent skill variable trial-by-trial. Based on the Ot, a rating is generated on trial *'t'* (Rt) that maximises the probability. Whether the response was correct or incorrect, the corresponding probability distributions are selected for each rating and one of the ratings (1, 2, 3, or 4) is then chosen following a categorical distribution.

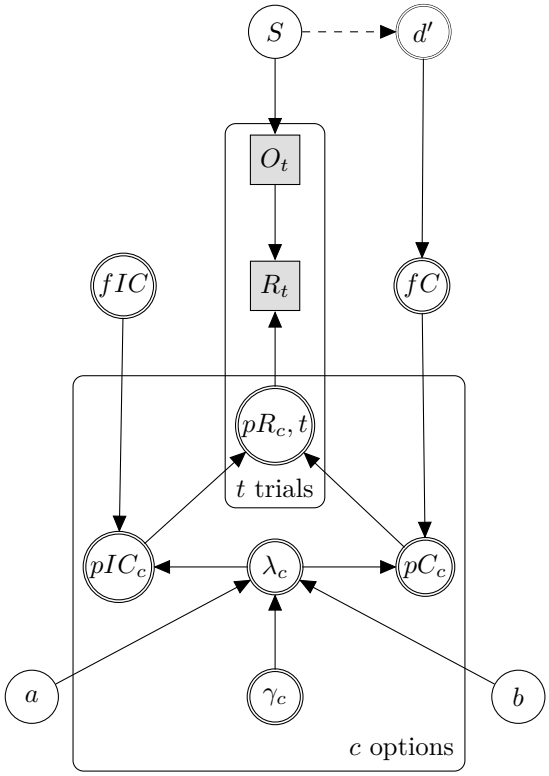

Figure 7. Graphical representation, plate notation of model 2 and 3. In model 2 skill and d' prime parameters are separately specified (no connection between them), in model 3 d' prime is expressed from skill (dashed line connecting them, double line border for d' when determined from S). Skill 'S' generates outcome 'O' on trial 't' (0 or 1 across t trials). D' prime is the mean of the correct distribution with a set variance of 1 (fC). fIC (incorrect distribution) is set at a normal distribution of 0, 1. The linear regression model consists of an unbiased set of threshold (gamma c), the intercept shift parameter (b) and slope scale parameter (a) to produce the actual thresholds (lambda c). These thresholds together with the two correct/incorrect underlying distributions produce probabilities for each rating (c options) given that an outcome is correct or incorrect (pC,c and pIC,c). Probabilites for the four ratings at each trial is then the corresponding set of probabilities given whether the trial was correct or incorrect. The actual rating is sampled from/fitted by a categorical distribution with respect to the given underlying probabilities (pRc,t). Notation: grey boxes mark observed data, circles indicate latent variables that may be determined (double line borders).

Mental models of the confidence scale are initialised (loop for C options, bottom part of the model). This follows Selker et al.'s linear regression model to estimate individual biases corresponding to each rating on the scale (Selker et al., 2019, p. 1956). Where:

$$\gamma_c = log(\frac{c/C}{1 - c/C})$$

$$\lambda_c = a\gamma_c + b$$

γ_c (for each option/rating) is intialised with a log transformation (C indicating total number of confidence ratings), then the linear regression model is applied on the γ_c to get λ_c (again for each threshold) by scaling (a) and shifting (b) them. Priors for a and b were chosen following the recommendations of Selker et al. (2019) (a gamma distribution of (2, 2) and a normal distribution of $\mathcal{N}(0, 1)$, respectively). Based on the estimated rating thresholds, probabilities corresponding to the two distributions underlying correct and incorrect trials (pC_c and pIC_c, size of the scale (4) times two for correct and incorrect trials) are calculated for each confidence rating following probability theory and areas under the correct and incorrect distribution curves between thresholds. So far, it is almost an exact adaptation of Selker et al.'s model and code (Selker et al., 2019, p. 163) (refer to Appendix B of this thesis).

In model 2, with unrelated S and d', the two parameters are individually modelled with no set connection or relationship between them, while in model 3, d' is expressed from S, the underlying latent variable of knowledge responsible for producing performance in task 1 (dashed line between S and d' latent variables reflects that this relationship is only true in model 3 but not in model 2). The rest of the models' architectures are identical. Model 2, unrelated S and d' constitutes the null hypothesis of research question one, that there is no link as hypothesized between skill and monitoring sensitivity. In model 2, the prior for d' is chosen to be a uniform distribution between 0 and 2 ($\mathcal{U}(0, 2)$). The maximum distance is reasoned here to be 2 variances (1+1, as the correct and incorrect distributions' variances). The incorrect distribution is set at a mean of 0 and variance of 1 following SDT trends of identifiability ($\mathcal{N}(0, 1)$). The mean of the correct distribution is d' itself while—because of equal variance SDT—its variance is also 1 ($\mathcal{N}(d', 1)$). If d' equals to zero it means that the two distributions entirely overlap and the participant is unable to distinguish between correct and incorrect trials. While the maximum distance between them could be +infinity it is most reasonable to expect values between 0, and 2, i.e., the two means two standard deviations apart from each other (Stanislaw & Todorov, 1999, pp. 139-140).

Model 3, d' expressed from S, captures the main hypothesis, the relationship between skill level (S), monitoring sensitivity (d), and the corresponding two underlying distributions (internal appraisals) towards correct and incorrect trials. The following equations are based on d' assumptions, its relationship with accuracy (area under the ROC curve), and inspired by Schraw's comprehensive account of calibration accuracy measurements and calibration curves (Keren, 1991; Schraw, 1995, 2009a, 2009b). The chosen approach involves three steps. First, it is assumed that the underlying skill level (knowledge of the material, S) directly reflects the accuracy of discriminating between correct and incorrect trials by the student. Since the area under the ROC

curve describes accuracy, and values need to be between 0.5 (chance) and 1, the skill is rescaled from 1/3–1 to 0–1 with just a regular linear interpolation, so that the cumulative distribution (area under the curve) of the rescaled skill would be between 0.5 and 1:

$$Accuracy \; = \; AUC \, of \, the \, ROC \, curve \; = \phi \, rescaled \, S$$

Since the prior for skill was chosen to be a uniform distribution between 1/3 and 1, this transformation generated a very simple 'step-function'-like distribution for accuracy a priori. Agents are initialized as optimal calibrators lying on the diagonal line of the calibration curve. Any deviance from this is captured by the confidence rating biases.

Second, relationships between accuracy and *d'*, established in previous literature, are utilised (DeCarlo, 2010; Stanislaw & Todorov, 1999). Certain *d'* assumptions need to be met, specifically that signal and noise distributions need to be both normal with an equal variance (Stanislaw & Todorov, 1999). As discussed in the previous section, both assumptions are strongly hypothesised for this model. Macmillan (1993) offers the following relationship between accuracy and *d'* (as cited in Stanislaw and Todorov (1999, p. 143)):

$$Accuracy \; = \; \phi \, (\frac{d'}{\sqrt{2}})$$

From this equation, *d'* can be derived from accuracy as:

$$d' \; = \; -\phi(accuracy) \; * \; \sqrt{2}$$

Thus, monitoring (calibration)—the distance between the two distributions—is the inverse cumulative or quantile function of accuracy multiplied by the square root of 2, when the incorrect distribution is fixed at a normal distribution of 0, 1. Variance of the correct distribution is also 1 and its mean is *d'* itself. These underlying distributions are then merged with the scaled confidence ratings and together generate the probabilities for each rating given correct or incorrect trials similarly to Selker et al.'s model; only here, the two distributions have equal variances (Selker et al., 2019). For example, probability of rating 2, given that the outcome was correct, is the area under the correct distribution's curve below the first threshold separating rating 1 and 2 (cumulative distribution of the correct distribution until the first threshold). The observed calibration accuracies are the products of an optimal calibration process and biases on the confidence scale. Psychologically, this may be interpreted as participants being perfectly aware of their abilities, that they are only affected by their biased perceptions of it with regards to the confidence scale. Cognitively, it may reflect that repetition of the recall process of knowledge is unaffected (no errors at this stage), but that the personal judgment of it is through biased perceptions again. The exact calculations and distributions can be examined in Appendix B where the jags file is presented.

MODEL 4: ADDED QUESTION DIFFICULTY

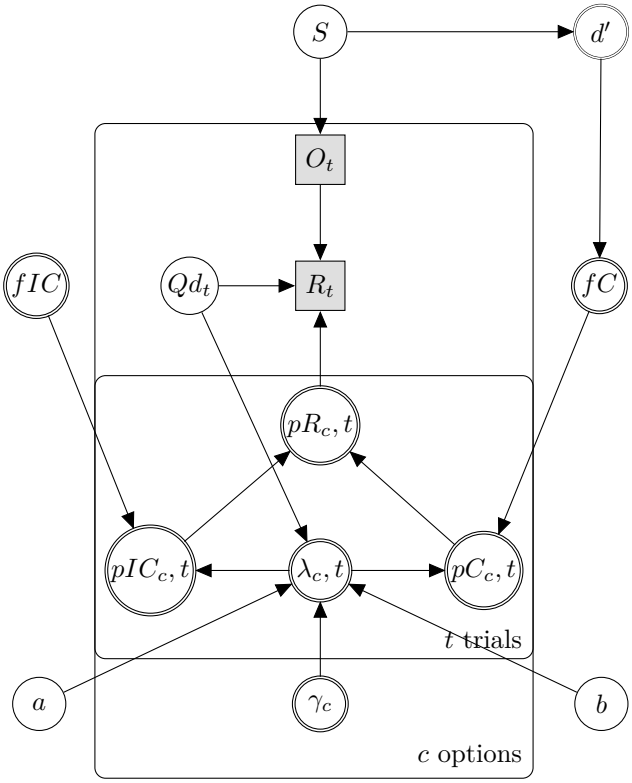

Figure 8. Graphical representation of model 4 with added question difficulty. As can be inspected on the plate notation, difficulty of questions is added to influence two components of the model, and alters the structure of the loops (rectangles). First, unique 'Qd' (question difficulty) on trial 't' together with overall skill ('S') produces the outcomes on trial 't', performance. Second, on each trial 'Qd' uniquely shifts the thresholds between the ratings (lambda c,t) which then in turn produces probability estimates uniquely for each trial, each choice and for the two cases whether the 'Ot' was correct or incorrect (pCc,t and pICc,t). From the corresponding probabilities depending on outcome (pRc,t) a rating is generated through a categorical distribution.

Question difficulty (*Qd* on Figure 8) is incorporated into the model in two ways. Following some ideas from Lee and Wagenmakers (2014b, pp. 71-74), outcomes on each trial are now a product of the participant's overall skill level (S) and the unique question difficulty level on trial 't'. *Qd*, similarly to the underlying skill parameter, is sampled from a uniform distribution between 1/3 and 1 for each trial. A probability of 1/3 reflects maximum difficulty where the probability of answering the specific question is equal to chance. While skill is an overall variable, question difficulty is unique to each trial. Performance outcomes are also affected by question difficulty, providing a level of uniqueness to each trial. In this model, performance is not only

generated from a latent skill variable but also affected by the specific question. The underlying distribution of each trial is now a Bernoulli trial since it is unique to each (refer to Appendix C for the jags code). To directly affect bias, Qd is then transformed to the same scale as thresholds to alter them with the same log transformation:

$$sQdt \leftarrow -log(\frac{1 - Qd_{\mathrm{t}}}{Qd_{\mathrm{t}}})$$

The overall thresholds corresponding to the confidence ratings set by the linear regression are then altered by adding the scaled question difficulty to directly shift the intercept, the bias term, and nothing else. Underconfidence reflects the use of the lower part of the scale with most of the distributions covering lower ratings (Figure 5). Adding a negative number means subtracting its absolute value which results in shifting all thresholds to the left giving more space for the higher thresholds on the distributions, increasing confidence. Thus, this manipulation should produce increasing confidence on more difficult questions, as the bigger the number (the more difficult the question is) the more thresholds are shifted to the left. This produces the pattern found in previous literature that is set out to be modelled and investigated in research question two (Keren, 1991; Nietfeld et al., 2005; Suantak et al., 1996). This alteration is unique to each trial and therefore, the estimation of probabilities for correct and incorrect trials needed to be brought into the trial loop as these will be unique, depending on the unique alterations by the new component of question difficulty. This model is computationally more expensive as probabilities are calculated for each trial and each option given whether the outcome was correct or incorrect, respectively. The basic architecture of the model is identical to model 3. Biases and optimal accuracy together produce observed data. This model, due to incorporating uniqueness of trials, allows for more variance in data patterns systematically by accounting for the general findings of hard-easy effects of question difficulty (Keren, 1991; Nietfeld et al., 2005; Suantak et al., 1996).

The presented models are evaluated in comparison to each other through: parameter recovery, model recovery based on simulations, and posterior-predictive checks on the collected data. This procedure is applied for model selection in terms of internal coherence and descriptive adequacy. The overarching aim is to find the model that best describes behaviour in the task while internally coherent and in parallel evaluate the hypotheses they capture. The next section (Results) lists the findings of the three studies and interprets them in the context.

Results

1. Parameter recovery study

The first step of investigation was a parameter recovery study for each model. As mentioned, the purpose was to see whether models are internally coherent and parameters are recoverable. Priors for parameters were sampled from meaningful distributions offering a wide range of possible values. Data was simulated on a trial-by-trial basis through sequential updating from posteriors to priors. The number of trials were set to 25, while the number of confidence ratings (C) were set to 4, imitating the task environment in which students participated. Simulated data was then fitted by the same model that produced it to infer parameters. Models were run forward and simulated data was fitted 100 times each. Sampled, set parameters were then plotted against their inferred counterparts and correlation coefficients were calculated to assess recoverability.

MODEL 1: RANDOM RESPONDING

Parameter recovery of the random responding model was successful. The skill parameter generating performance outcomes was recoverable with a significant positive correlation coefficient of 0.91 (p-value $<$ 2.2e-16). The correlation between set and inferred parameters from the range of (chance) and 1 (perfect performance) can also be inspected on Figure 9. Lower values are somewhat more scattered than higher values indicating that performance levels around chance are a bit more difficult to consistently infer correctly . This makes sense as guessing is not necessarily set at exactly chance performance level. Some are more lucky, while others are less. There is variance. With increasing knowledge, variance is less uncertain as the latent skill variable incorporates a more concrete cognitive process of information retrieval. The underlying latent skill variable was the only parameter of this model. Results of this first parameter recovery study supported the simplified process of task 1 generating performance.

MODEL 2: UNRELATED MODEL

To assess parameter recovery in the model with unrelated S and d', all four parameters of the model (S, d', a, b) were plotted on their respective spaces to inspect correlation between set and inferred values (Figure 10). The underlying skill (S) was recoverable similarly to the random responding model (*r: 0.92, p<2.2e-16*). Monitoring sensitivity, d', however, was not (*r: 0.65, p=2.639e-13*). d' was set to vary between 0 and 2 following a uniform distribution. Both the scale (a) and shift (b) parameters of the linear regression operationalising the mental model of the confidence

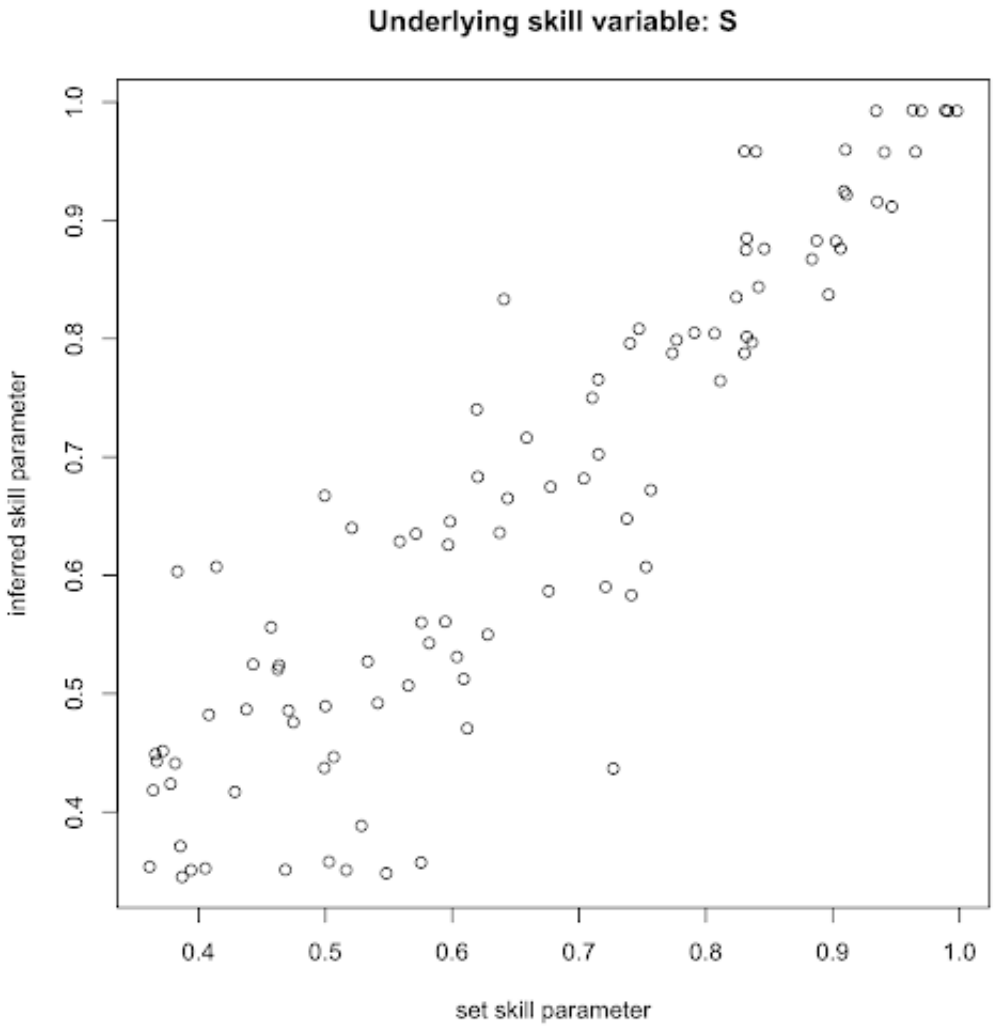

Figure 9. Recoverability of the skill variable in the random responding model. Set parameter values are on the x-axis sampled from a uniform distribution between and 1 to run the model forward. On the y-axis inferred parameters from the simulated data by the same model are plotted. Correlations between them can be visually inspected (r=0.91, p-value < 2.2e-16).

scale were recoverable (*r=0.88, p< 2.2e-16, and r=0.93, p< 2.2e-16*, respectively). The scaling of the ratings becomes more uncertain with higher values (refer to Figure 10). On the other hand, the thresholds set by the two parameters together were beautifully recoverable (*0.95, p< 2.2e-16*, refer to Figure 10). Each set and inferred threshold was individually compared (3 comparisons for each of the 100 simulations and inference). The underlying skill and monitoring variable, *d'*, showed no correlation in this model (*r=-0.01, p=0.89* between set parameters, and *r=0.1, p=0.32*

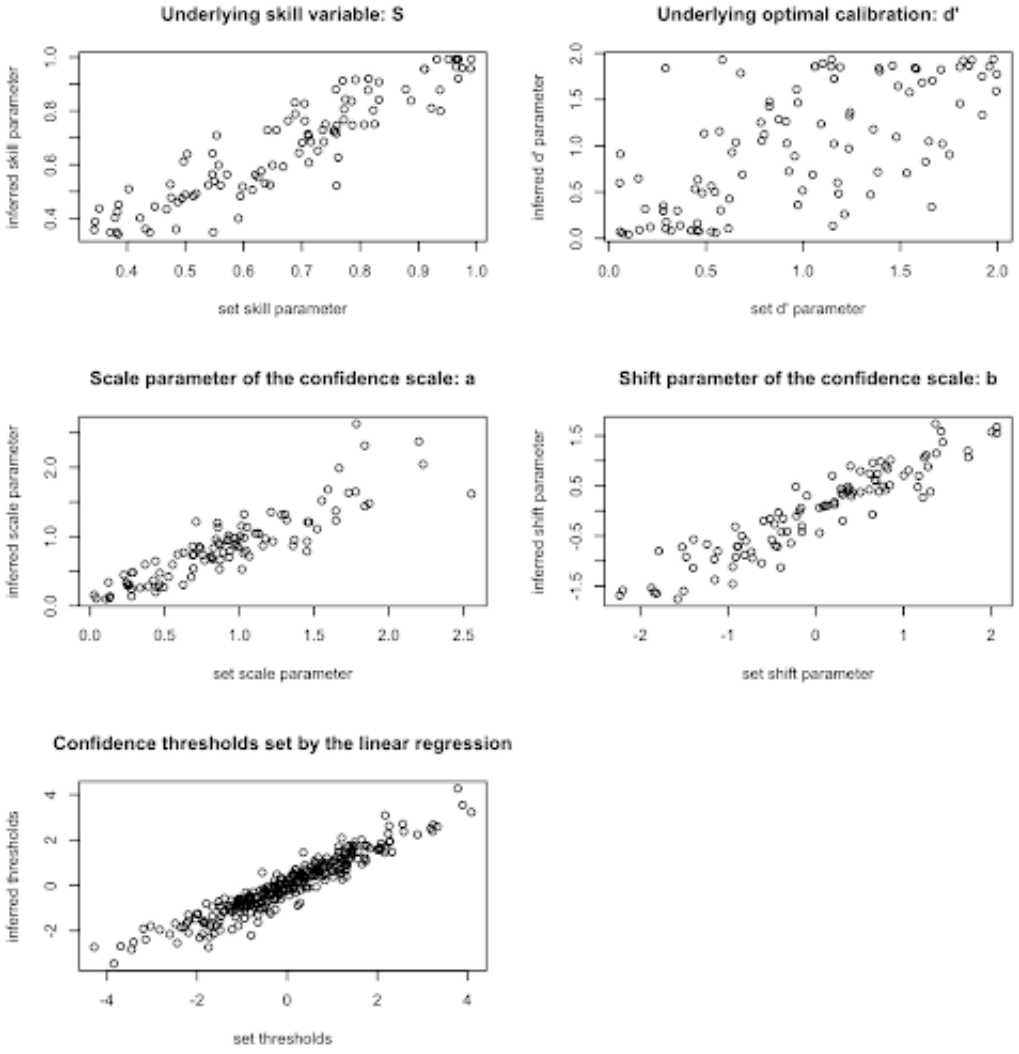

Figure 10. Parameter recovery of the model with unrelated S and d'. On this set of plots, correlations are visualized for the four free parameters and the confidence thresholds set by parameter 'a' and 'b'. Apart from d', monitoring sensitivity (r=0.65, p=2.639e-13), all the other parameters were recoverable (r between 0.88 and 0.93, p< 2.2e-16) including the thresholds (r=0.95, p< 2.2e-16).

between inferred parameters, refer to Figure 11). This is good because this model was supposed to incorporate the alternative null-hypothesis, that skill and monitoring sensitivity are completely unrelated. Interestingly, *d'* was not recoverable in this model.

MODEL 3: MODEL WITH ESTABLISHED LINK BETWEEN SKILL AND D'

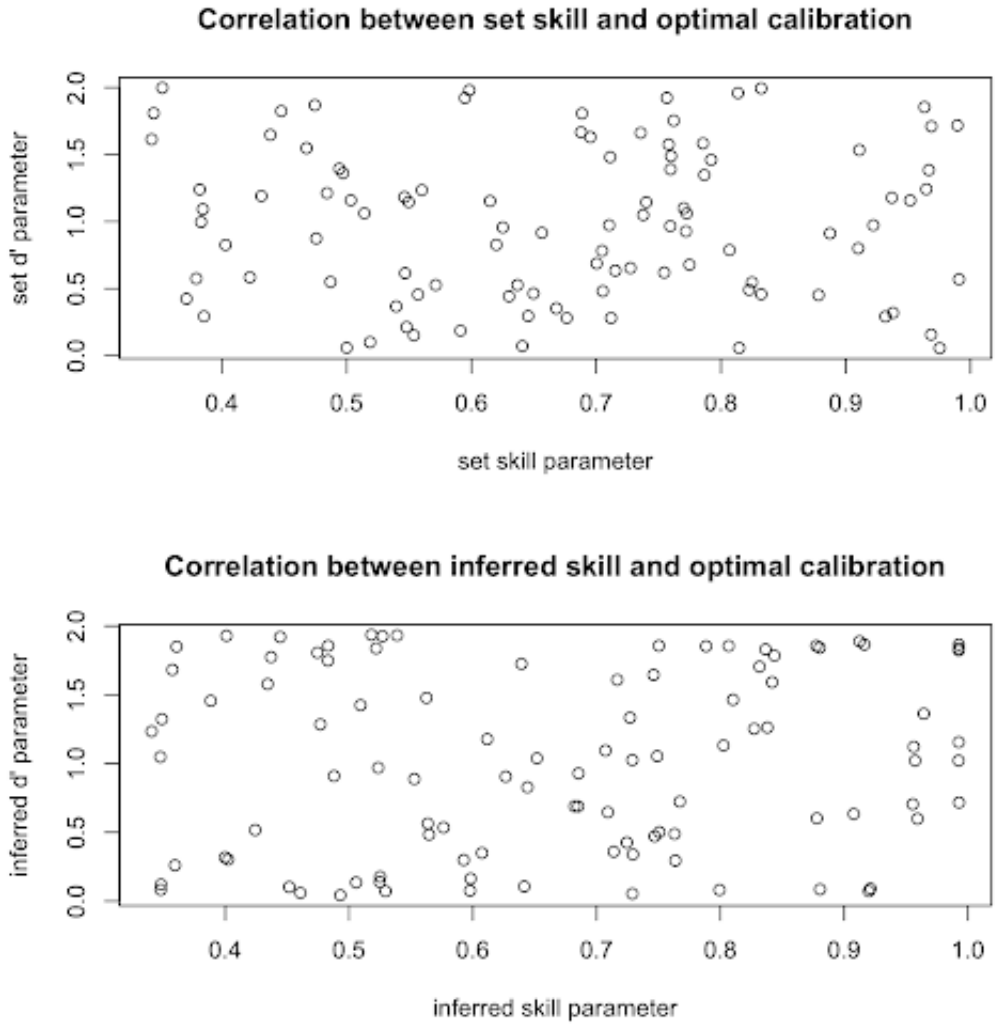

Figure 11. Correlation between S and d' in model 2. Model 2 establishes no relationship between underlying skill variable generating performance and d' underlying calibration accuracy. The fact that no correlation was found between these two variables either in set or inferred parameter space is expected here (r=-0.01, p=0.89 between set parameters, and r=0.1, p=0.32 between inferred parameters). This model was built to operationalise the alternative null-hypothesis in comparison to model 3 that expresses a direct link between S and d'. These plots show how scattered the values of these two parameters are with respect to each other.

Model 3 incorporated the hypothesis that monitoring sensitivity, optimal calibration, is directly related to the underlying skill. As shown in the model, *d'* is explicitly expressed from *S*. Thus, the parameter *d'* was expected to show the same correlation patterns as *S* on their respective space of possible values. This was true as both were

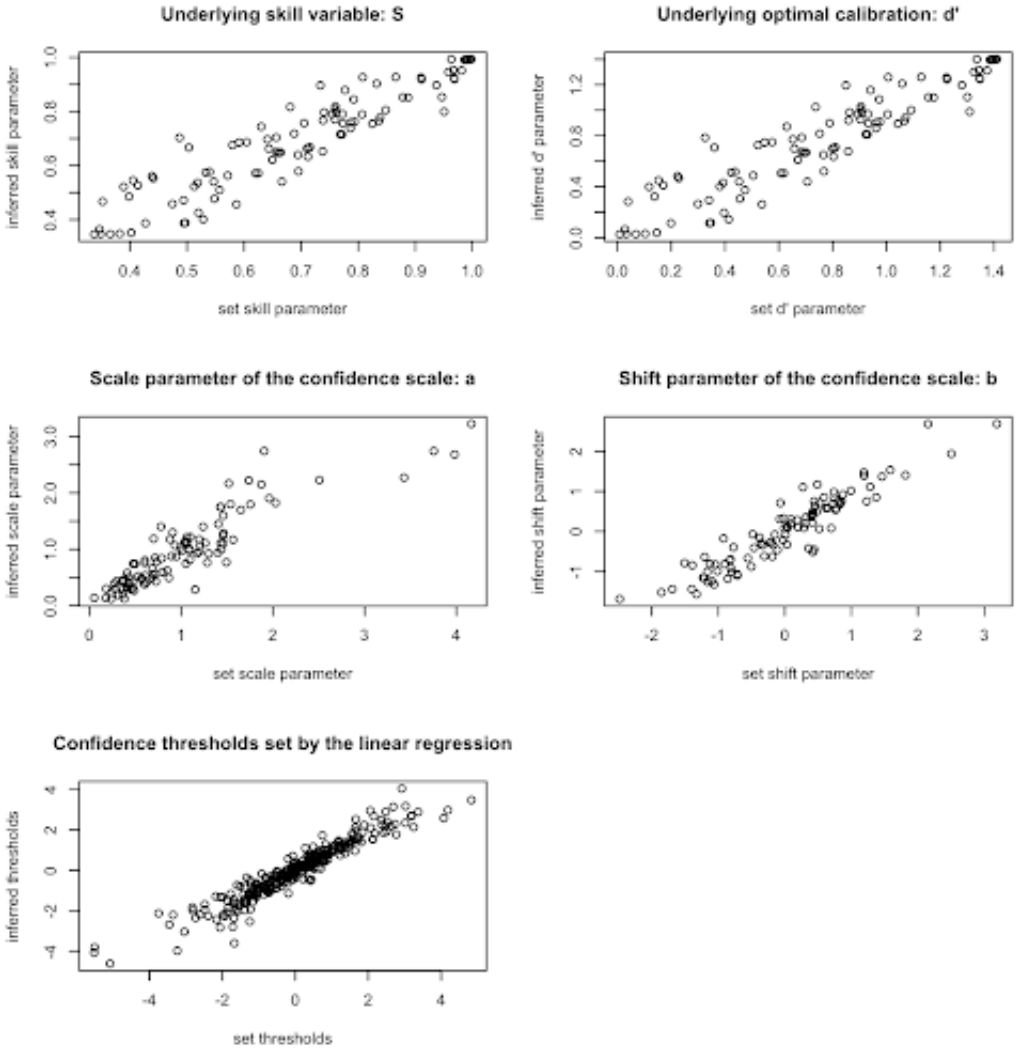

Figure 12. Parameter recovery of the model with d' expressed from S. As expected, values of d' follow the same pattern as the values of S with the same correlation coefficients showing recoverability (r=0.92, p< 2.2e-16). The rest of the two parameters and the resulting thresholds also correlate and are recoverable (r: 0.89-0.95, p< 2.2e-16). In comparison to model 2 (figure 10), only the scale parameter of the confidence scale 'a' is different, taking on higher values but also seeming slightly more recoverable.

recoverable with the exact same correlation coefficients and patterns when plotted (*r=0.92, p< 2.2e-16*, Figure 12). The rest of the parameters were also recoverable similarly to model 2 (*a: 0.89, b: 0.93, thresholds: 0.95, p< 2.2e-16*). The only difference was that the confidence scale's scale parameter (a) took on higher values occasionally, while at the same time correlated a slight bit better (*r=0.88 and r=0.89*

correlation coefficients in the two models respectively, Figure 10 and 12). As expected, *S* and *d'* on both the set and inferred parameter space show perfect correlation (*r=1*, Figure 13). This pattern was supposed to mimic the optimal diagonal line of the calibration curve (Keren, 1991, p. 222). It is interesting that *d'* was unrecoverable in model 2 suggesting that in this context it is reasonable to assume a link between underlying skill and monitoring sensitivity.

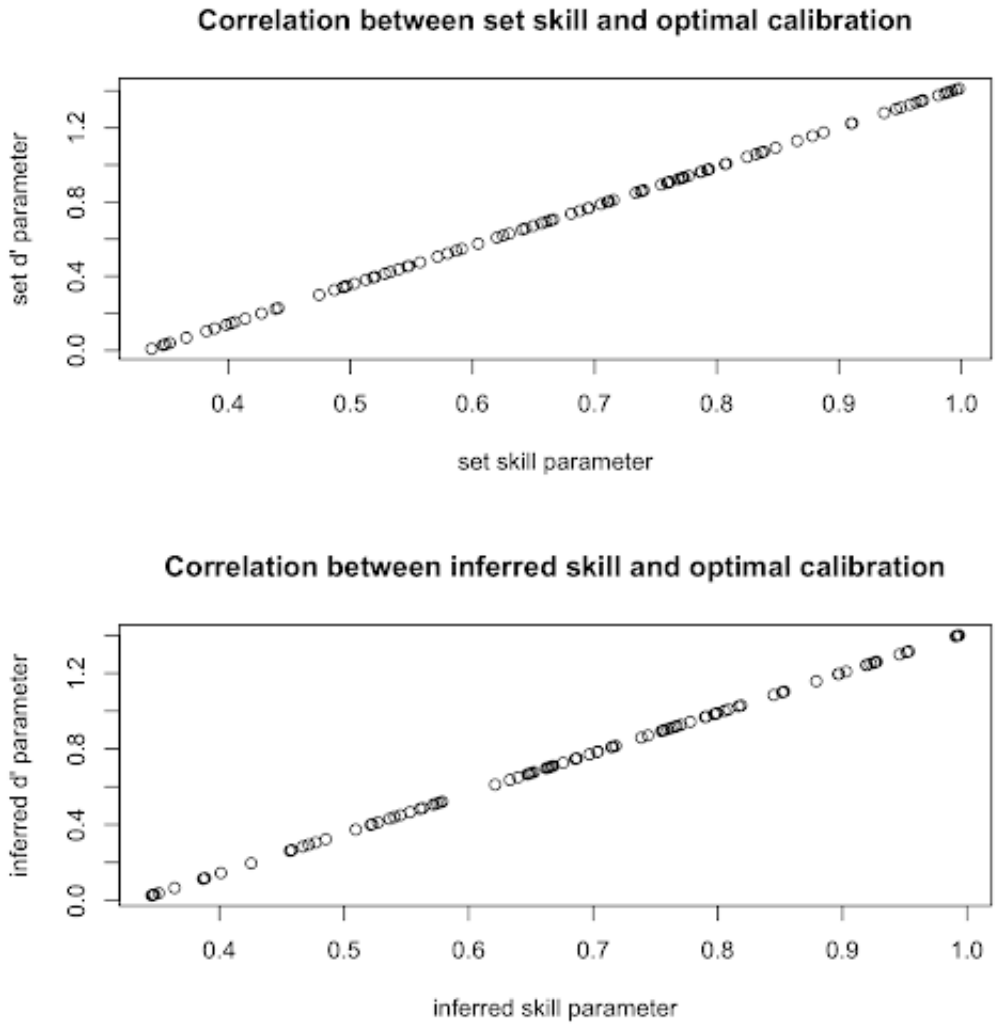

Figure 13. Correlation of skill and d'. As expected, since d' is expressed from S, they show perfect correlation of 1 on both the set and inferred parameter space. This mimics the optimal diagonal line of the calibration curve with perfect accuracy and incorporates the ideas of the first hypothesis.

MODEL 4: ADDED EFFECTS OF QUESTION DIFFICULTY

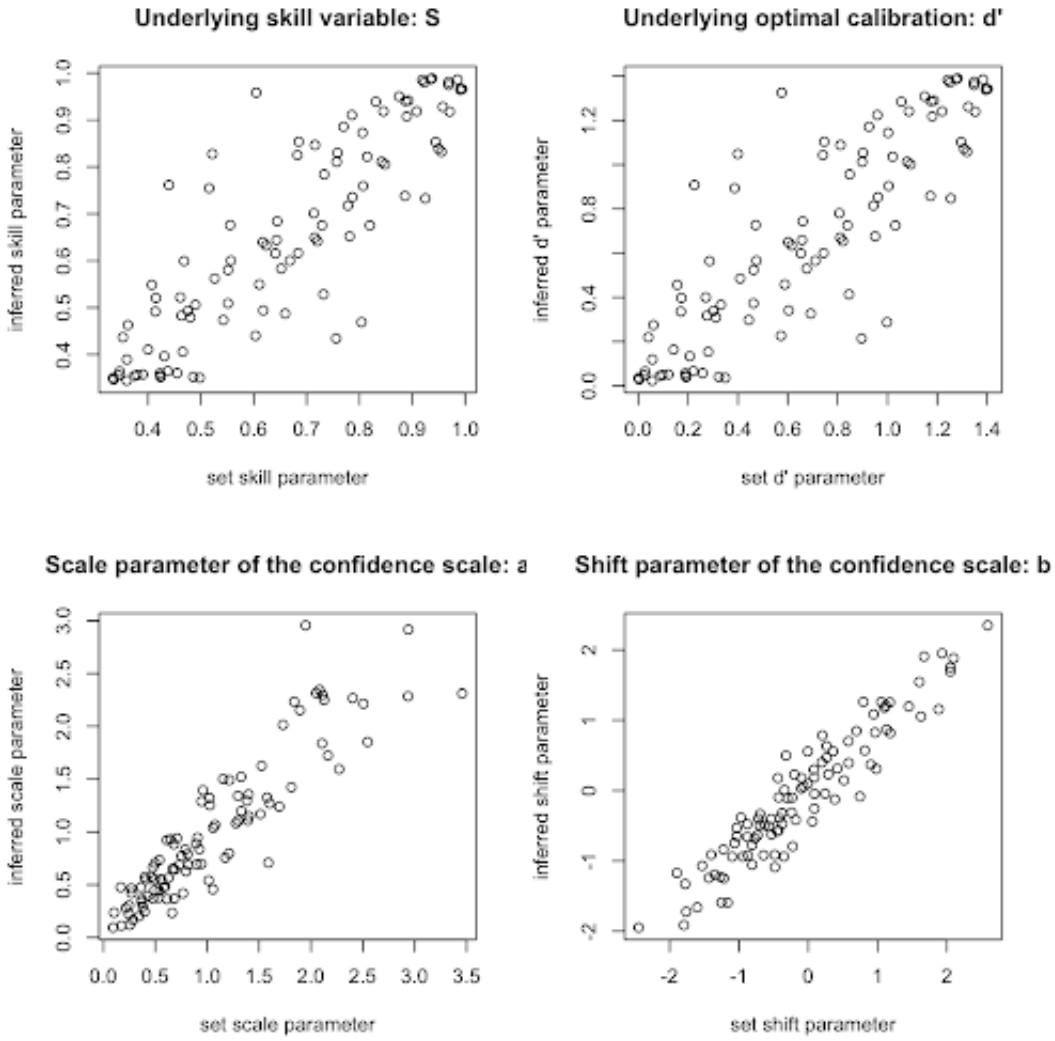

Figure 14. Parameter recovery of model 4 with added question difficulty. All parameters were recoverable as can be seen on these plots (S d': 0.86, a: 0.91, b: 0.94, p-value $<$ 2.2e-16). S and d' are more scattered than in model 3, as with added question difficulty an element of trial-specificity was introduced, allowing for more variance in the generated data. Thresholds were not assessed as they were calculated for each trial based on the alteration of the linear regression's intercept by the question difficulty.

Model 4 included the established relationship between S and d', but the architecture with added question difficulty allowed for more variance in the data as the effects of unique trials were considered. The results of this modification can be examined in Figure 14. Both S and d' are more scattered in comparison to model 3, but still follow the same patterns on their respective parameter space with the same

correlation coefficient (*r=0.86, p-value < 2.2e-16*). They were still recoverable but to a lesser degree. This is not surprising considering the variance of unique trials introduced into the patterns. In this model, outcome on trial 't' was not directly produced by underlying skill, but was also affected by the unique perceived difficulty of the question. The overall underlying confidence scale parameters a and b remained recoverable in this model as well (*a: 0.91, b: 0.94, p-value < 2.2e-16*). This indicates that overall biases still exist and are recoverable even though each unique trial shifts the mental model of the confidence scale slightly. Thresholds were not assessed here as they were calculated for each trial with respect to the shift in bias corresponding to the specific question's difficulty. Just like for model 3, a perfect correlation of 1 was found between S and d' on both the set and inferred parameter space.

2. Model recovery study

In model recovery, each model was run forward to simulate 100 data sets each. On these four times 100 data sets, each of the four models were fitted and assessed to determine which one performed the best on each round. The one with the smallest DIC (deviance information criteria) scores were considered to 'win'. At the end, a confusion matrix was produced to visualize the results (Table 1). A recoverable model should fit its own data the best most times.

	Datasets simulated from the models			
Models fitted	Model 1	Model 2	Model 3	Model 4
Model 1	81	5	4	17
Model 2	6	19	19	13
Model 3	9	72	70	39
Model 4	4	4	7	31

Table 1

Confusion matrix of the results of the model recovery study. Model 1 is with random responding, model 2 with no explicit link between S and d', in model 3 d' is expressed from S, and model 4 is with the added effects of question difficulty. The diagonal marks the data sets fit by the models from which they were generated. These values should be the greatest in each row and column to indicate perfect model recoverability. This is not the case here. Refer to the main text for a detailed interpretation of these results.

Data sets simulated by the random responding model were best fit by itself to a high degree (81%). The model also fit data sets simulated from the model with added question difficulty to some degree (17%). Data from the model with unrelated S and d' was best fit by the model with d' expressed from skill (hypothesised relationship) to a large degree (72%) in comparison to its own model (19%). Data sets of the model

with linked S and d' were best fit by itself 70% of the times. Interestingly, data from both model 2 (alternative null-hypothesis) and model 3 (with the hypotesised relationship) were best fit by model 3 and poorly fit by model 2 to the same extent. In terms of architecture (loops), these models are alike: Model 2 is rather unable to explain data from both itself and from model 3, while model 3 can account for both to a great degree. This may also suggest that the hypothesised link between S and d' is naturally present in this context and task environment. It is important to note here that model 3 is simpler as d' is expressed from S, thus essentially using fewer parameters. Consequently, model 3 is also more constrained on d'. So far, based on both parameter and model recovery, comparing model 2 and 3, model 3 is winning.

Data sets produced by the model with added question difficulty were best fit by model 3 as well (39%), but not significantly more than by itself (31%). To some degree, all models fit these data sets, the worst being model 2 with unrelated S and d' (13%) not far from the random responding model (17%). This model generated more varied data as thresholds change trial-by-trial. In model 2 and 3, thresholds are set as underlying, internal, true variables that do not change. In model 4, they change quite dynamically, but also in concert with performance as question difficulty affects outcomes on the trials as well. It is reasonable—based on the differences in architecture—that model 4 produced data sets that are more varied and thus fitted by all competing models to some degree. On the other hand, model 4 fits data sets generated from the random responding model and model 2 the worst and fares poorly also on data from model 3 while performing the best on its own data (Table 1: last row). It is important to note that both models 3 and 4 incorporate the hypothesised relationship between S and d' and are in that respect not that different from each other. Based on the first two studies of parameter and model recovery, I would conclude that the hypothesised relationship between S and d' makes sense in terms of naturally present patterns in the data. Model 4 with added question difficulty fared well on parameter recovery, but results from the model recovery were not convincing. The problems could be explained by the significantly different architecture generating more varied data and its underlying conceptual similarities with model 3.

3. Descriptive adequacy (posterior-predictive checks)

To assess the descriptive adequacy of the models, they were fitted on real data to assess their predictive powers. The real data came from 38 participants going through 24 multiple-choice questions rating their confidence for each on the 4-point Likert-scale. For each participant, their data was used to update the models' prior distributions sequentially and then—using the resulting posterior distributions to predict the observed data, the confidence ratings—to compare with the actual ratings. The question in this study was which model would best predict the observed data.

For the random responding model, out of the 38 sessions (38 participants) the amount of accurate predictions ranged from 0-19 out of 24. Mean accuracy was 4.29 and standard deviation 4.23. The median was 3.5. The probability of this model predicting participants' ratings accurately was 0.18 (correct predictions / all). These results suggest that it is unlikely that participants of this study responded randomly on the confidence scale. The predictive success of model 2 with unrelated S and d' ranged between 6-20 out of the 24 trials. Mean correct predictions were 12.74 (out of 24), median 12, and standard deviation of 3.68. The overall accuracy of this model was 0.53. Model 3 with the hypothesised link between S and d' likewise guessed ratings correctly on a range between 6-20 with a mean of 12.42, median of 12, and a standard deviation of 3.41. Prediction accuracy was 0.52, slightly worse than for model 2. In this study, model 2 actually fitted real data better to a small degree. This might be because real data is more varied than simulated and model 2 is less restricted than model 3. On the other hand, while model 2 performed slightly better, model 3 had a slightly smaller standard deviation suggesting that model 3 is more certain around its predictions. Model 4 with added question difficulty had a predictive success rate of 0.81, significantly outperforming all competing models. Correct guesses ranged between 14-24 with a mean of 19.39, median of 19.5, and standard deviation of 2.42. Apparently, variance that could not be captured by any of the other models is not random, but instead systematically varies with question difficulty, and this has the biggest effect on the behavioural patterns.

Based on descriptive adequacy of the models, model 4 with added question difficulty undeniably performed the best. Even from a general perspective, 0.81 accuracy is significantly high. Model 4 also fared well in parameter recovery. Only in model recovery it did not seem to be recoverable in light of the alternative competing models, which may have been the reason. Apart from the random responding model, all models were similar in one way or another. Model 2 and 3 shared an underlying architecture of loops and internal calculations, while model 3 and 4 both established the hypothesised link between skill and monitoring sensitivity. These similarities may have given rise to two distinct patterns in simulated data. The first stems from an architecture of overall unchanging mental models of the confidence scale, best fitted by model 2 or 3, in which case model 3 won based on parameter and model recovery. In descriptive adequacy they were almost the same, most likely due to their similar architecture, but in the other two studies model 3 significantly outperformed model 2. The second distinct pattern is related to S and d' fluctuating in concert by explicitly being linked and best fitted by model 3 or 4, in which case model 3 won in model recovery, but model 4 won based on descriptive adequacy with the additional architectural difference of changing biases with respect to unique question difficulty on trial 't'. As model 4 fared so well in posterior predictive checks, it is considered the ultimate best model in this study and discussed in more detail in the next section. Further, model 4 incorporates both the main hypothesised link between

S and *d'* and the hard-easy effect of question difficulty. It also performed well in parameter recovery, indicating internal coherence. In model comparison, choice of contestants (the other models) may have affected its performance. Nonetheless, even in model recovery, model 4 cannot be disqualified because, while datasets generated from model 4 were fitted by all models, datasets from the alternative models were not well fitted by model 4.

Discussion

In conclusion, findings of this study suggest that there is validity in assuming a link between underlying skill, knowledge of the material, and monitoring. While calibration accuracy is a product of both optimal discriminability and the potentially biased mental models of the confidence scale, optimal monitoring suggests a distinct cognitive activity of repetition in metacognitive activity. Both parameter and model recoveries supported model 3 expressing *d'* from *S*, even suggesting that the hypothesised relation may be a natural pattern. Model 4 with added question difficulty outperformed all competing models in posterior-predictive checks. This result suggests that the hard-easy effects of question difficulty account for most variance in the data. Regarding the two aims of this study (refer to Research questions), both were explored to some degree of success. Computational modelling of behavioural data seems to be a valid approach to investigate cognitive processes even in less-constrained educational settings. The hypothesised patterns were systematically modelled and captured, but the extent to which these attempts can be considered successful is up for debate. The following sections discuss the findings in more detail and list limitations of and alternative approaches to this study.

1. Interpretation of the results and contributions

The first research question posed in this thesis was: **1) Is it possible to build a model to meaningfully capture the relationship between performance and online calibration? To what degree will that model be internally coherent and fit real data in comparison to alternative models with no connection between performance and monitoring?** An attempt was made to build such a model in the specific context in a simplified way, in which responding to the multiple-choice task was not specifically modelled, only outcome and associated confidence ratings. However, whether it meaningfully captures the relationship between performance and calibration is up for debate. The model was internally coherent, although it may have been too general and simple, fitting all kinds of data as realised through the model comparison study where it outperformed all the other models at fitting all datasets, except data from the random responding model. Furthermore, it did not fit real data better than the alternative model with no connection between performance and monitoring. On the other hand, the model with added question difficulty also captured the hypothesised relationship and fitted real data significantly better than any other models. For this reason, and because of internal coherency issues with the model with no link between skill and monitoring both in parameter and model recovery, the hypothesised relationship is considered to hold some truth, but might be better operationalised in an alternative way as discussed later. The fact that—while

internally coherent across all studies—the random responding model fitted real data very poorly indicates that confidence judgments are definitely not random, and the approach taken in this thesis has some validity as a starting point for further investigations and model alterations.

Secondly, it was asked: **2) Is it possible to build a model to account for systematic biases based on question difficulty? To what degree will that model be internally coherent and fit real data better than alternative, simpler models?** Again, yes, it was possible to build such a model. The model with added question difficulty was internally coherent in parameter recovery, but showed some issues in model comparison. This might have been because the alternative model only linking skill and monitoring—model 3 operationalising research question one—was fitting data from all three competing models apart from random responding. In posterior-predictive checks on the real, observed data, this model significantly outperformed all other models. The hypothesis incorporated in this model seems to be valid. In this architecture, participants are initialised as optimal monitors, calibration accuracy perfectly mirroring and fluctuating with performance as the diagonal line of a calibration curve. When performance is at chance, calibration accuracy also is at chance. As performance increases, calibration accuracy also increases. However, participants have different mental models of the confidence scale which produces discrepancies in this perfect pattern and accounts for the actual observed data. Most importantly, these mental models differ across questions. Each trial is unique to each individual, depending on their perception of the question. While they all studied from the same book, they paid attention to and remembered certain sections more than others, thus all questions are uniquely more or less difficult for them. The implicit hypothesis in this model's architecture was that overall and trial-specific biases and optimal absolute discriminability ability together account for calibration accuracies that would be measured by Scraw's absolute and relative calibration measures (Schraw, 1995, 2009b).

This thesis contributes to the field of monitoring research through some bold ideas and a methodological, conceptual contribution: 1) What if, deep down, everyone is a perfect monitor of their own processes, and only our biased pictures of ourselves and our skills produce different patterns? What if metacognition is a sequential process, a recall of internal experience, upon prompting, and being biased by overall self-concepts? 2) This study shows that it is possible to account for participants' individual hard-easy effects of question difficulty and not as general question-specific noise. It is important to note, with regards to contribution 1, that it was safe to assume underlying optimal calibration in the context of the population of the present study. More specifically, students in this study should have no issues recalling a cognitive event right after it happened. This may not be the case for patient groups with neurological or psychiatric disorders where cognitive processes, such as repetition

of a mental event, might not be optimal due to deficits related to memory (Bertrand et al., 2016; Cosentino, 2014; David et al., 2014). Nonetheless, the first contribution is more philosophical in nature and certainly up for debate, but will be reflected upon both from a general philosophical and a personal perspective approaching the end of this thesis. The relevance of the second contribution, however, is important to discuss in more detail here. Models that incorporate individual differences may be more robust and thus offer stronger support for certain behavioural patterns.

Two accounts are considered here to dive deeper into an explanation of the hard-easy effect of question difficulty: the decision variable partition model and the theory of adjustment and anchoring (Suantak et al., 1996; Tversky & Kahneman, 1974), and the ecological model and the confidence-frequency effect (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Suantak et al., 1996). The pattern of underconfidence with regards to easy questions and overconfidence with regards to hard questions resonates with prospect theory of risk aversity for gains and risk seeking for losses (Newell et al., 2015, pp. 123-123). Certainly, not decision making but judgment under uncertainty were at play here which led me back to the fathers of the field, Tversky and Kahneman (1974). The anchoring effect posits that judgments made under uncertainty are insufficiently adjusted since there is always a relative point, towards which participants are calibrated, that bias them (Tversky & Kahneman, 1974, pp. 1128-1129). Judgments are relative to past experience of judgments. When there is a chain of experiences, a certain event is judged in light of the events preceding it. The decision variable partition model explains the hard-easy effect with participants' insufficient adjustments being due to a lack of information about the task environment and thus improper adjustment of decision partitions (Suantak et al., 1996, p. 204). Students may accurately judge one set of questions to be easier or harder than another set. However, when these constitute individual events adjustments become less precise. Insufficient adjustments might explain underconfidence with regards to easier and overconfidence with regards to increasingly more difficult questions, as previous events anchor the perception of the current event, and, when bigger adjustments would be needed, participants fail to properly update their decision partitions due to their bias (Suantak et al., 1996, p. 204). Without feedback on performance or any indication of question difficulty, students were in the dark and needed to rely on their own perceptions of difficulty. Since questions were uncontrolled and randomly shuffled for each student, the difficulty of them was even harder to assess and also impossible to track objectively as a researcher. No scale could be applied on the levels of question difficulty as they were unknown and rather random. Nevertheless, even in this rather fluid task environment the hard-easy effect seems to have played a significant, systematic role in the behavioural patterns. This implies that students were perceiving levels of difficulty in their biased ways, even if the task environment was not specifically constructed to reproduce behavioural patterns related to the hard-easy effect. Results of this study suggest that the hard-easy effect is even more universal on an individual

basis and not just as an overall question-specific noise.

The ecological account of the hard-easy effect is rooted in the Brunswikian theory that people learn things in their natural environment, thus, naturally, more frequent cues become more strongly associated with the object of learning (Gigerenzer et al., 1991; Suantak et al., 1996). This is in line with the availability heuristic that more frequent cues—as they are more strongly associated with the object of recall—will drive the choice selection and thus also the confidence judgment upon accuracy (Newell et al., 2015, p. 87). The ecological model, however, blames the experimenter's biased choice of questions and related levels of difficulty for the hard-easy effect (Suantak et al., 1996, p. 205). Objectively, according to the researchers, easier questions are considered those whose obvious cues work well, while more difficult ones are those in which obvious cues do not work. Quite logically, this produces underconfidence on easy and overconfidence on hard questions. Again, since students were coming up with the questions themselves for the test in this study, experimenter biases in question selection are non-existent. This indicates that even when questions are not consciously constructed to be easier or harder, level of difficulty is an inherent quality of these tasks that participants respond towards on an individual basis, meaning that a set of questions might be easier for some students, while the same set may be difficult for another group of students.

Keren highlights that calibration in stimulus classification is different from calibration in participants' own responses, the main difference being the nature of stimuli (Keren, 1991, p. 262). Stimulus classification operates on a predetermined set of similar events of two kinds, signal and noise. Participants judging their own prior responses, categorizing them as either right or wrong, are stimuli generated by themselves which requires a different process. There is more connection between the stimulus (making a decision) and response (judging that decision). Nonetheless, both involve events similar to one another. In comparison, general knowledge tasks that constitute questions from random subjects are considered unique events as each stimulus is more salient from one another (Keren, 1991, p. 228). This study involved questions in the multiple-choice task that may be considered similar events. However, saliency of a trial in comparison to the neighboring trials may be an individually different perception. Students sent a question from their course material that they had already covered up until that day. On the other hand, students may not progress with the material in perfect synchrony. Some might have already read the chapter for the upcoming week and assumed that others did, too, while another group of students may have been a week behind. For this reason, certain questions sent by well-prepared students could have been very salient for other students that had not looked at that part of the material yet. Saliency is thus considered in comparison to other questions and other students. The model sampled a question difficulty level for each trial and each student to account for these very diverse perceptions of question difficulty.

In summary, the biggest strength (and weakness) of this study is its rather uncontrolled, non-traditional task environment of multiple-choice tasks. Question difficulty was unique to each participant and each question, while common patterns were also found. The results of posterior predictive checks showing the universal effects of individual perceptions of question difficulty is thus a significant finding. The hard-easy effect is not only a task-dependent or experimenter bias, but a bias in participants' processing in light of their perceptions of the task. The model incorporated mechanisms indicative of biases in cognitive processes of calibration that are sensitive to individually different perceptions of task difficulty. Effects of the saliency of events responsible for insufficient confidence adjustments and differences in what available cues may be considered indicative of correct response are now not only patterns found in tasks constructed to allow for a common ground among participants, but are also general individual patterns. What this means is that even though students were taking the same test, their experience and perceptions were considered different, while at the same time, their data showed the same underlying patterns. It might not have been the same set of questions that students found difficult, but for all students there existed a set of questions that were harder, thus producing overconfidence. Such findings highlight the beauty of the approach of Bayesian cognitive modelling. With this method it is possible to test hypotheses accounting for individual and, in this case, even event-dependent differences (Wagenmakers et al., 2008). Frequentist approaches operate with rigid protocols, such as constrained task environments, to allow for (null-)hypothesis falsification through significance testing (ibid.). This was not feasible here with the rather random task environment. The high levels of uncertainty around the task was possible to account for in a Bayesian framework, as it is more flexible and not focused on significance testing (ibid.). Instead only the probability of certain ideas gets updated. With Bayesian cognitive modelling the focus is to build flexible potential hypotheses up for debate, including scientific uncertainty (ibid.). Through this approach it is possible to build multiple alternative truths, as also shown in this study, and update their probability in the specific context (ibid.). Such models are also robust in nature, built on parsimony, and able to account for complex systems as the findings of this investigation also highlighted. The main reason why Bayesian modelling is close to me is that it never claims to have found the truth. It is a humble technique that from its methodological nature always considers subjectivity in research and science.

2. Limitations and alternative approaches

2.1 Method of observation: task environment.

First and foremost, it is important to discuss the experimenter bias in task 2. I chose the question to assess confidence: "How confident are you in your response?" Then I

decided on a 4-point rating-scale where 1 was indicated with “low confidence”, and 4 with “high confidence” (refer to Appendix A, based on Fleming and Lau (2014); Maniscalco and Lau (2014)). These were my choices as the researcher, likely affecting participants’ responses and biases. The question could have been phrased many different ways, e.g., “how confident are you that you have found the correct answer?” for more specificity, or “how likely would you rate that you were correct?” for indicating probabilistic estimates. The first example shows that the question could have been made more specific. On the other hand, I was intentionally simplifying it and leaving more space for interpretation to minimize my bias. My reasonings were that using more words will inevitably put more of my words into their ‘head’. I also find the expression ‘found the correct answer’ anxiety inducing as it highlights the fact that there is a true answer that you should have found, and now how do you rate your abilities at succeeding in that. While, obviously, that is the case in a multiple-choice task, the alternative more specific question sounded too direct and aggressive to me.

Schraw distinguishes between probability judgments of performance and confidence judgments (Schraw, 2009b; Keren, 1991, pp. 245-248). Confidence judgments are most commonly used in calibration studies (Schraw, 2009b). Considering the meaning of the alternative second example question with the word ‘likely’, participants would probably have associated the question with a probability estimate of their performance. This would not have been a different approach. Let us consider a 2-choice task, true or false, after which the participant is asked to rate on a scale how likely they think they answered correctly. If the person in question does associate the word ‘likely’ with probability, they may be more aware of their probabilities for success. They may reflect on the fact that, even if they do not know anything, they might get lucky 50% of the time, thus participants may gamble. Confidence, on the other hand, has a connotation of inner judgment which is different and unrelated to external probabilistic laws. With probability judgments, an optimal guesser is expected to gamble as they already gamble to answer the first question. In a probabilistic alternative, participants would probably quantify uncertainty instead of an assessment reflecting the judge’s internal state (Keren, 1991, pp. 245-249). The word confidence may elicit more honest responses. With confidence ratings in calibration, a ‘pure’ guesser (with exactly 0.5 performance) is expected to be honest about their inability in the test and choose low ratings all the way in the task. This would result in data where misses (correct with low rating) and correct rejections (incorrect with low rating) are equal, because the participant guessing is honest about guessing, and their calibration accuracy (which in this case is perfect) should mirror that. Nevertheless, some students may have interpreted the question differently, even from a probabilistic perspective.

As mentioned, a 4-point rating scale was chosen to assess confidence. For discrimination tasks it is best to choose an even number scale, so that participants cannot be indecisive choosing the middle score. I find 4-point ratings scales to work the best

in most situations because they are the simplest to understand, in my opinion. The more options there are, the more levels participants need to subjectively interpret in their mental model of the scale. This choice was also in line with previous research of monitoring in signal detection theory (Fleming & Lau, 2014; Maniscalco & Lau, 2014). A continuous scale from 0 to 100 could have been an alternative, considering that most participants are comfortable with percentages, and is a common practice in calibration research (Keren, 1991; Schraw, 2009b). However, that would have produced more complex data and less clear-cut patterns to model. Selker et al.'s (2019) linear regression was built for categorical scales and worked very well on the chosen 4-point rating scale data. Nonetheless, while I preferred this scale, not all of the participating students could have been just as happy and comfortable with it as I was. In general, some may be more comfortable with 5 or 6-point rating scales or in fact any other alternative scales. There was some degree of cultural differences among students, and culture most definitely plays a role in preferring certain conventions for scales that we understand the most based on our experiences. In my own life, I noticed how people from different cultural educational backgrounds prefer the grading systems they grew up in and even after years cannot get used to a new one in comparison. I am certainly an example of that. After 6 years, I'm still interpreting my Danish grades on the 7 point scale to the Hungarian system of a 1-5 scale. Likewise, someone from the UK who is used to percentages probably prefers a continuous scale between 0 and 100 in any task if that is an option. Misunderstandings of the scale could have resulted in calibration accuracies below chance level, and this was not regarded in any of the models (Keren, 1991, p. 241).

Another limitation of the task environment was that students had no external incentives to perform well and be honest. They did not receive points, scores, or grades meaningful to their lives. Keren argues that incentives encourage honesty about 'true' underlying confidence levels (Keren, 1991, pp. 248-249). Massoni, Gajdos, and Vergnaud (2014) ran a study with three different methods of observing confidence judgments of performance to assess which environment fits best with SDT. They recommend using the matching probability paradigm in such studies that are based on SDT analysis, in comparison to simple confidence ratings. In matching probability, participants are asked whether they prefer to be paid according to the correctness of their answer or a specified lottery to reveal their 'true' confidence in their response (Massoni et al., 2014). The present study used a simple confidence rating on a numerical scale with no monetary or any alternative meaningful consequences. In an educational context, Isaacson and Fujita (2006) 'manipulated' their students with grades in meaningful ways in the task environment, such that students had to make wise decisions about which questions to choose to answer to maximize their scores. Such a manipulation involves regulatory, executive control processes in connection to metacognitive monitoring. On the other hand, with the addition of any incentives, participants are more likely to gamble and quantify uncertainties instead of

an assessment of internal states. Nonetheless, without consequences a task can become mundane or boring and result in a lack of engagement after a certain point. From the perspective of the course, it was important to not include study-relevant consequences. The overarching pedagogical mindset of Study Techniques was to engage students in finding their internal motivations to discover themselves and take responsibility for their learning within and not outside of themselves. As I instructed the course, doing my job and assisting students in a way that I believed to be good for them was more important than constructing the best task environment for my research. Sometimes such conflicts show up and one needs to take a stance.

No reaction time data was recorded in the experiment on SurveyXact. This may seem a major limitation, considering the possibilities in modelling choices. Data on time could have provided more information and opportunities for more fine-grained models of detailed, cognitive processes, such as Pleskac Busemeyer's two-stage dynamic SDT account of calibration (Pleskac & Busemeyer, 2010). The underlying conceptual framework is that reaction time is indicative of the recall process, as evidence or cues accumulate in decision making between choices. This was an unexplored area in the present investigation for two reasons. For one, working with the confidence scale was already a bigger task than expected. For two, responding on the first task of multiple-choice was disregarded as the environment was rather uncontrolled in its question selection and due to uncertainties around the preparedness of students with regards to the material. Additionally, as questions and answers were shuffled and three alternative choices were used in a random order, modelling students' choices would have been complex with a high degree of uncertainty, which was not the dominant focus of this study. True/false or similar binary questions are better suited for modelling choices. However, reaction times on the confidence ratings or overall between trials could have been utilised in an indirect, presumptuous way. For example, they could be indicative of engagement with the task or difficulty of the question and the cognitive processes involved in task 1. These assumptions, however, would have been big jumps in reasoning as responding in the first task was disregarded.

As mentioned in the previous section, the applied non-traditional task environment was not only a strength but also a weakness in this study. In a traditional metamemory task, the memory acquisition phase is controlled (Pratte et al., 2010; Selker et al., 2019; Yonelinas et al., 1996). Participants receive a list of words chosen by the experimenters and are given a set time to study it before the testing session. I did not control the set of questions nor the amount of time students spent studying the material. While these uncertainties offer greater support for the validity of the findings about monitoring and the hard-easy effect of question difficulty, it is also more difficult to reason about the cognitive processes involved. One participant may exhibit all kinds of different mental activities due to these uncertainties and surrounding factors. They might guess due to not yet having read the chapter from which the

question was made, or they might be certain about the answer because they came up with the specific question. They might be honest on the confidence scale because they just started the test, or they could be bored after the 10th or 15th trial and become less engaged. To some questions, they might definitely know the answer, but after making a decision realize that they chose the wrong option and rate their confidence low. Alternatively, they could be deciding on the basis of familiar cues because they had attended the course lectures where they paid attention but not have read anything. All of this is possible even for a single participant. It would have been a clearer situation to have given the students something to read, then asked related multiple-choice questions and acquired ratings of confidence, similar to reading comprehension studies of metacognition (Schraw & Dennison, 1994). Here, there would be an information acquisition phase, a recall phase, and a monitoring phase, neatly following one another sequentially. Experimenter bias of question selection could still be eliminated in this paradigm by asking students to come up with a question themselves during reading the material, then in the recall phase only receive other students' questions. I think such modifications would be useful for the field of calibration and monitoring research to implement, to limit biases of experimenters' selection of questions and artificial ordering of question difficulty.

2.2 Method of investigation: modelling.

Selker et al.'s linear regression model of the confidence rating scale is parsimonious but also constrained (Selker et al., 2019). It does not allow for internally incoherent, in any way reversed, mental models of the confidence scale. Optimally, such reversals should not happen unless participants misunderstand the task or interpret the scale the wrong way. I was lucky to have a very optimal population to study, but with either very culturally diverse groups or patient groups to fit their data, such a constraint may cause issues. The same goes for performance levels below chance which was outside of the scope of the models applied in this study. An alternative solution, to account for significantly different behaviours on the scale, could be to build an additional random responding model with different rating preferences irrespective of questions. However, that would not be able to account for nor explain systematic scale reversals.

The psychological meaning of *d'* is optimal monitoring sensitivity. As it is now, d' is quite tightly connected to the underlying skill variable generating performance. It is expressed from skill. Consequently, the two internal distributions for correct and incorrect responses are also very constrained. Their variance is set at 1, the mean of the incorrect distribution is 0, and the correct distribution's mean is d'. This is a rather simplified account. Even with the equal variance SDT, variance of the two distributions could be sampled from a meaningfully broad prior. With added question difficulty, it could be reasonable to switch to an unequal variance paradigm

of SDT based on the familiarity versus recollection account (DeCarlo, 2010; Dunn, 2004; Wixted, 2007). More difficult questions may result in a process of familiarity, while easier ones may be associated with recollection and unequal variance. However, in this recollection process, unequal variance may be around any of the distributions not only the 'signal', correct distribution because, as mentioned, incorrect trials are not necessarily clearly rejected, but may be also recalled, and at a certain point a mistake could be realised and then rejected.

A huge limitation of the methods is not accounting for the biased patterns of under- and overconfidence in relation to performance shown on the calibration curve in the Introduction. This would require a hierarchical, latent mixture model, and if I were to start my thesis now I would definitely do this (Lee & Wagenmakers, 2014b). Basically, two groups of participants can be hypothesised, one with lower performance with regards to some overall group mean, and another with higher performance. The lower performing group would be systematically overconfident, while the higher performing group would be underconfident. Performance of this model on real data could then be compared with the performance of the question difficulty model and a merged version of the two. Additionally, the lower performing group could be initialized with a bigger variance around the distribution means and vice versa for the higher performing group. Closer to guessing, the familiarity account of SDT seems a more intuitive choice, while for the higher performers the alternative recollection theory may be more valid with less uncertainty (based on Wixted (2007)). Familiarity of a question comes from a process of remembering that is prone to errors, while knowing in recollection is more concrete (Dunn, 2004).

As an alternative to the one confidence scale, a modified meta-d' approach could have been applied (Fleming & Lau, 2014; Maniscalco & Lau, 2014). Meta-d' usually refers to signal detection tasks where participants' first task is to determine whether the stimulus was a signal or noise (Maniscalco & Lau, 2014). Task 1 is a 2-choice alternative task with 2 different options that are modelled together and in association with confidence ratings. A trial may be a signal trial towards which participants respond correctly or incorrectly, or the trial can be a noise trial eliciting again correct or incorrect responses (Maniscalco & Lau, 2014). In this paradigm the options are modelled separately, together with signal detection and on top of that with confidence. Here, modelling of the responding process in task 1 with three alternative choices was disregarded for simplicity as it would have been too complex. On the other hand, after responding, students could first have been asked to indicate their believed correctness on a binary scale and then rate their confidence. In this paradigm from a psychological interpretation, next to performance a perceived performance component would be generated with the first binary choice which would then direct the choice on the scale instead of the objective outcome. Calibration is now directed towards perceived skill instead of actual skill. While this idea sounds intuitive, and one could definitely

play around with it, it is also rather redundant. Both the perceived skill and the confidence ratings would be part of monitoring, but the confidence ratings would not be connected to the actual performance monitored. Although, in calibration studies everything seems a little redundant as it is assessing thinking about thinking and when one thinks too much about that, sometimes it all collapses into a spiral of redundancy. If I were to start my PhD on this topic of metacognition in monitoring processes and calibration studies, I would definitely experiment with the task environment, collect data from all kinds of settings, and try all kinds of models, including possible mixes and extensions of them. An idea I had at the end of this thesis process was to model correct and incorrect trials separately and initialise variances of the two distributions to account for extremity in choice patterns, such as using the farther ends of the scale (L. Brenner, Griffin, & Koehler, 2005; L. A. Brenner, 2003; Suantak et al., 1996). However, these models were all modelling choices in the first cognitive task. In my imaginary PhD project this would not be an issue, though, because I would experiment with binary tasks and alternative task environments where choices can be modeled.

2.3 Conceptual issues and alternative potentials.

One of the main conceptual problems with linking performance and calibration lies in the interpretation of incorrect trials and identification of those. The more one knows, the fewer incorrect trials one generates. Thus, there is less opportunity to correctly identify them as incorrect (correct rejections). At the same time, the more incorrect trials one makes, the more opportunities there are to classify them correctly, but the number of misses should increase in concert (correct trials classified as incorrect). Both Isaacson and Fujita (2006); Nietfeld et al. (2005) argue that better performing students are better at judging what they know and what they do not know, as their calibration accuracy is better, based on their descriptive analyses. Fewer mistakes in task 1 suggest—and in the models of this thesis even result in—fewer mistakes in task 2. Now, what they do not know decreases as their performance increases, leaving less opportunity to 'prove' that they know what they do not know. Monitoring of knowledge depends on knowledge, the object of monitoring. There is once again some redundancy and a conceptual paradox, especially due to shifting to modelling in this thesis from descriptive studies. The resolution of this thesis was using Bayesian methods accounting for and carrying over uncertainties providing an opportunity for variance. Optimal behaviour is not necessarily the real behaviour, as a hypothetical mental model is not necessarily the real mental models of participants (Keren, 1991, p. 264). This is not to say that the models are false, but participants are not perfect: Even if a model truthfully describes their processes, there are always uncertainties and variances. A big limitation of all monitoring research is that metacognition cannot be isolated from cognition, i.e., performance cannot be stripped from calibration. A basic cognitive task is always needed so that participants can monitor and reveal their

skills.

The present thesis followed a rather top-down approach. Findings were synthesised to establish patterns, models were built based on such patterns and existing models and then fit on real data to check validity. If I were to start my life's work on metacognitive monitoring, I would start with a bottom-up approach. I would run multiple qualitative studies to explore processes as individual experiences in monitoring all kinds of tasks (e.g., problem solving, monitoring of emotional states). Through these I would inform a reasonable selection of general process components and design specific tasks and models to systematically study them. This thesis suggests three such components: 1) being prompted to elicit the process, 2) repetition of the cognitive event, and 3) switching to a third person perspective of oneself. The first and a third component implies that metacognition is not a parallel but a sequential process that needs to be elicited. This resonates with Hawkins's ideas of the predicting mind (Friedenberg & Silverman, 2012, pp. 370-372). He suggests that, out of an intelligent habit, our mind continuously predicts the next actions and states, both externally and internally, and learning happens with surprise, a bottom-up process when a novel event occurs. This may be how the mind naturally gets prompted to reflect and change. Prompting to engage is needed, and while asking someone 'how confident are you in your answer?' for the 20th time may not be very surprising, (it might not be surprising even the first time), the question still prompts the participant to engage in the metacognitive task although in an artificial, constructed way.

Working on this thesis made me reflect to a high degree on the repetition account throughout the thesis. I have been playing the confidence scale game in my complex lines of thoughts and connections between materials and previous research over and over again. I would continuously question my ideas, 'how confident am I that this idea is right'. I would go through the same complex lines of thoughts over and over again reaching similar conclusions sometimes with more or less confidence. I had been making judgments under uncertainty which were connected to the cognitive activity under monitoring, and there is only my internal knowledge with which to compare my cognitive processes and validate or reject the ideas. Monitoring of knowledge depends on knowledge, and the process may very well be a repetition of cognitive events.

The third component of switching to a third person perspective may be the most interesting and challenging to investigate, at least in similarly rigid tasks. Theory of mind is based on the idea that we learn about ourselves through others (Meltzoff, 1990, 2007b). Thus, it is reasonable to assume that metacognitive processes carry some social characteristics even in non-social contexts. Such a shift may be conceptualised as looking at oneself through the 'eyes' of others that may or may not exist. Self-concept is also associated with social interactions as others might confirm or disconfirm them (Hattie, 1992, p. 37). In my life I found that social interactions prompt

me to reflect on my self-concept. I have been working in interdisciplinary teams of researchers for the past few years where, sometimes, I would be the only one from the field of Cognitive Science. These experiences made me reflect on who I am as a cognitive scientist and all the knowledge, skill, and methods I internalised throughout my degree without even explicitly realising.

3. Concluding remarks

In conclusion, results of the present thesis suggest some validity of the repetition account of metacognitive monitoring, and universality of the hard-easy effects of question difficulty that may vary from person-to-person and unaffected of experimenter-bias. There are many alternative approaches and possibilities for investigation already at the level of online monitoring, and even more in metacognition. The present thesis may not be flawless, but still showed some promising avenues for further research: investigating metacognition through the implementation of the techniques of computational modelling of behavioural data in less-constrained educational settings. This thesis started with the notion of agency, self-concept, and metacognition. More specifically with my inspiration and motivation for studying metacognition in the context of engaging students in discovering themselves in learning contexts. Through monitoring it is possible to become aware of cognitive processes, and more consciously use them or even change our views of ourselves. I have definitely changed through writing this thesis. Now, I am realising a need for reflecting and updating my self-concept. Thus, I am closing with a quote from Franken about self-concept that I find empowering as I am letting go of my 'student self-concept':

"there is a growing body of research which indicates that it is possible to change the self-concept. Self-change is not something that people can will but rather it depends on the process of self-reflection. Through self-reflection, people often come to view themselves in a new, more powerful way, and it is through this new, more powerful way of viewing the self that people can develop possible selves" (Franken, 1994, p. 443)

Acknowledgments

First of all, I would like to take a moment and appreciate my opportunity of instructing Study Techniques for three years and all the inspiring experiences I got from this job and the students. I am indepted to the 38 students for donating their data to my thesis. I feel lucky to have stood on the shoulders of so many great scientists, researchers, and their valuable works. I am very grateful for my two amazing supervisors, Daina Crafa, who has been with me in thick and thin, and Joshua Charles Skewes, whose voice I will never stop hearing when modelling, personally my greatest teacher. I wish to also thank my stand-in proof-readers, and last-minute Latex helpers, the best colleagues at the Center for Computational Thinking and Design at AU. I especially owe a huge thanks to Bjarke Vognstrup Fog and Line Have Musaeus. Last but not least, thanks to my co-workers from the bar for taking my weekend shifts last minute in this last month before deadline, and my family and friends for cheering me on.

References

Armstrong, T. (2012). *Neurodiversity in the classroom: Strength-based strategies to help students with special needs succeed in school and life*. ASCD.

Bacon, F. (1597). *Meditationes sacrae*. Londini.: Excusum impensis Humfredi Hooper. Retrieved 2021-11-06, from http://gateway.proquest.com/openurl?ctx$_ver$ = $Z39.88-2003res_id = xri : eeborft_val_fmt = rft_id = xri : eebo : image$: 194755 (OCLC: 766939025)

Bertrand, E., Landeira-Fernandez, J., & Mograbi, D. C. (2016). Metacognition and Perspective-Taking in Alzheimer's Disease: A Mini-Review. *Frontiers in Psychology*, *7*. Retrieved 2020-08-27, from https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01812/full (Publisher: Frontiers) doi: 10.3389/fpsyg.2016.01812

Blakemore, S.-J., & Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry*, *47*(3-4), 296–312. Retrieved 2020-10-30, from https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7610.2006.01611.x (_eprint: https://acamh.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7610.2006.01611.x) doi: 10.1111/j.1469-7610.2006.01611.x

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, *73*(4), 269–290. (Publisher: Taylor & Francis)

Brenner, L., Griffin, D., & Koehler, D. J. (2005, May). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, *97*(1), 64–81. Retrieved 2021-11-10, from https://www.sciencedirect.com/science/article/pii/S0749597805000051 doi: 10.1016/j.obhdp.2005.02.002

Brenner, L. A. (2003, January). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes*, *90*(1), 87–110. Retrieved 2021-11-10, from https://www.sciencedirect.com/science/article/pii/S0749597803000049 doi: 10.1016/S0749-5978(03)00004-9

Chua, E. F., Pergolizzi, D., & Weintraub, R. R. (2014). The cognitive neuroscience of metamemory monitoring: Understanding metamemory processes, subjective levels expressed, and metacognitive accuracy. *The cognitive neuroscience of metacognition*, 267–291. (Publisher: Springer)

Cosentino, S. (2014). Metacognition in Alzheimer's Disease. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 389–407). Berlin, Heidelberg: Springer. Retrieved 2021-11-30, from https://doi.org/10.1007/978-3-642-45190-4_17 doi: 10.1007/978-3-642-45190-4_17

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2014). Failures of Metacognition and Lack of Insight in Neuropsychiatric Disorders. In S. M. Flem-

ing & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 345–365). Berlin, Heidelberg: Springer. Retrieved 2021-11-30, from https://doi.org/10.1007/978-3-642-45190-4$_1$5 doi: 10.1007/978-3-642-45190-4$_1$5

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*(3), 304–313. (Publisher: Elsevier)

Dunn, J. C. (2004). Remember-Know: A Matter of Confidence. *Psychological Review*, *111*(2), 524–542. Retrieved 2021-11-10, from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.111.2.524 doi: 10.1037/0033-295X.111.2.524

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological review*, *124*(1), 91. (Publisher: American Psychological Association)

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. Retrieved 2020-09-17, from https://www.frontiersin.org/articles/10.3389/fnhum.2014.00443/fullh3 (Publisher: Frontiers) doi: 10.3389/fnhum.2014.00443

Franken, R. (1994). *Human motivation . Pacific Grove, CA: Brooks.* Cole Publishing Co.

Friedenberg, J., & Silverman, G. (2012). *Cognitive science : an introduction to the study of mind* (2nd ed.). United States of America: SAGE Publications, Inc.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological review*, *98*(4), 506. (Publisher: American Psychological Association)

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (2009). A growing sense of "agency". In *Handbook of metacognition in education* (pp. 13–16). Routledge.

Hattie, J. (1992). *Self-Concept.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An Introduction to Good Practices in Cognitive Modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience* (pp. 25–48). New York, NY: Springer. Retrieved 2020-05-13, from https://doi.org/10.1007/978-1-4939-2236-9$_2$ doi: 10.1007/978-1-4939-2236-9$_2$

Isaacson, R. M., & Fujita, F. (2006, August). Metacognitive Knowledge Monitoring and Self-Regulated Learning: Academic Success and Reflections on Learning. *Journal of Scholarship of Teaching and Learning*, *6*(1), 39–55. Retrieved 2021-04-21, from https://eric.ed.gov/?id=EJ854910 (Publisher: Indiana University)

Karpicke, J. D., Butler, A. C., & III, H. L. R. (2009, April). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. Retrieved 2020-09-06, from https://doi.org/10.1080/09658210802647009 (Publisher: Routledge _eprint: https://doi.org/10.1080/09658210802647009) doi: 10.1080/09658210802647009

Kelley, C. M., & Jacoby, L. L. (1996, April). Adult Egocentrism: Subjective Experience versus Analytic Bases for Judgment. *Journal of Memory and Language*, *35*(2), 157–175. Retrieved 2021-04-25, from

https://www.sciencedirect.com/science/article/pii/S0749596X96900091 doi: 10.1006/jmla.1996.0009

Keren, G. (1991, October). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. Retrieved 2021-11-10, from https://www.sciencedirect.com/science/article/pii/000169189190036Y doi: 10.1016/0001-6918(91)90036-Y

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639. (Place: US Publisher: American Psychological Association) doi: 10.1037/0033-295X.100.4.609

Koriat, A. (1994). Memory's knowledge of its own knowledge: The accessibility account of the feeling of knowing. *Metacognition: Knowing about knowing*, 115–135.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, *124*(3), 311–333. (Place: US Publisher: American Psychological Association) doi: 10.1037/0096-3445.124.3.311

Lab, C. a. B. (2021). *Ethical Approvals.* Retrieved 2021-11-20, from https://bss.au.dk/en/cognition-and-behavior-lab/for-researchers/checklist-for-new-r

Lee, M. D., & Wagenmakers, E.-J. (2014a). *Bayesian cognitive modeling: A practical course.* Cambridge university press.

Lee, M. D., & Wagenmakers, E.-J. (2014b, April). Chapter 6: Latent Mixtures. In *Bayesian Cognitive Modeling: A Practical Course.* Cambridge University Press. (Google-Books-ID: Gq6kAgAAQBAJ)

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: meta-d, response-specific meta-d, and the unequal variance SDT model. In *The cognitive neuroscience of metacognition* (pp. 25–66). Springer.

Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014, December). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*, 1455. Retrieved 2021-06-30, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4263084/ doi: 10.3389/fpsyg.2014.01455

Meltzoff, A. N. (1990). Foundations for developing a concept of self: The role of imitation in relating self to other and the value of social mirroring, social modeling, and self practice in infancy. In *The self in transition: Infancy to childhood* (pp. 139–164). Chicago, IL, US: University of Chicago Press.

Meltzoff, A. N. (2007a). Infants' causal learning: Intervention, observation, imitation. In *Causal learning: Psychology, philosophy, and computation* (pp. 37–47). New York, NY, US: Oxford University Press. doi: 10.1093/acprof:oso/9780195176803.003.0003

Meltzoff, A. N. (2007b, January). The 'like me' framework for recognizing and becoming an intentional agent. *Acta Psychologica*, *124*(1), 26–43. Retrieved 2020-08-06, from http://www.sciencedirect.com/science/article/pii/S0001691806001211 doi: 10.1016/j.actpsy.2006.09.005

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2015). *Straight choices: The psychology of decision making.* Psychology Press.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive Monitoring Accuracy and Student Performance in the Postsecondary Classroom. *The*

Journal of Experimental Education, *74*(1), 7–28. Retrieved 2021-04-22, from https://www.jstor.org/stable/20157410 (Publisher: Taylor & Francis, Ltd.)

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. (Publisher: American Psychological Association)

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, pp. 1–10). Vienna, Austria. (Issue: 125.10)

Plummer, M. (2012). *JAGS Version 3.3. 0 user manual.* Lyon, France.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 224. (Publisher: American Psychological Association)

Pressley, M., & Ghatala, E. S. (1990, January). Self-Regulated Learning: Monitoring Learning From Text. *Educational Psychologist*, *25*(1), 19–33. Retrieved 2020-09-07, from https://doi.org/10.1207/s15326985ep2501_3 (Publisher: Routledge _eprint: https://doi.org/10.1207/s15326985ep2501_3) doi: 10.1207/s15326985ep2501_3

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*(3), 114–121. (Publisher: Wiley Online Library)

Schneider, W., & Löffler, E. (2016). The development of metacognitive knowledge in children and adolescents. In *The Oxford handbook of metamemory* (pp. 491–518). New York, NY, US: Oxford University Press.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology*, *9*(4), 321–332. Retrieved 2021-04-25, from https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2350090405 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350090405) doi: https://doi.org/10.1002/acp.2350090405

Schraw, G. (2009a, April). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33–45. Retrieved 2021-10-11, from http://link.springer.com/10.1007/s11409-008-9031-3 doi: 10.1007/s11409-008-9031-3

Schraw, G. (2009b). Measuring metacognitive judgments. In *Handbook of metacognition in education* (pp. 427–441). Routledge.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary educational psychology*, *19*(4), 460–475. (Publisher: Academic Press)

Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019, October). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*, *51*(5), 1953–1967. Retrieved 2021-06-30, from https://doi.org/10.3758/s13428-019-01231-3 doi: 10.3758/s13428-019-01231-3

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137–149. (Publisher: Springer)

Studieguide. (2021). *How to apply for a Bachelor's degree.* Retrieved 2021-10-21, from

https://bachelor.au.dk/en/admission/c1926890

Suantak, L., Bolger, F., & Ferrell, W. R. (1996, August). The Hard–Easy Effect in Subjective Probability Calibration. *Organizational Behavior and Human Decision Processes*, *67*(2), 201–221. Retrieved 2021-11-17, from https://www.sciencedirect.com/science/article/pii/S0749597896900746 doi: 10.1006/obhd.1996.0074

Tobias, S., & Everson, H. T. (2009). The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In *Handbook of metacognition in education* (pp. 107–127). Routledge.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. Retrieved 2021-11-23, from https://www.jstor.org/stable/1738360 (Publisher: American Association for the Advancement of Science)

Veenman, M. V. J., Hout-Wolters, B. H. A. M. V., & Afflerbach, P. (2006, April). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. Retrieved 2020-07-22, from https://link.springer.com/article/10.1007/s11409-006-6893-0 (Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Kluwer Academic Publishers) doi: 10.1007/s11409-006-6893-0

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). Springer.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010, May). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. Retrieved 2020-05-15, from http://www.sciencedirect.com/science/article/pii/S0010028509000826 doi: 10.1016/j.cogpsych.2009.12.001

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., ... Blakemore, S.-J. (2013, March). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271. Retrieved 2020-10-30, from http://www.sciencedirect.com/science/article/pii/S1053810013000068 doi: 10.1016/j.concog.2013.01.004

Wilson, R. C., & Collins, A. G. (2019, November). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. Retrieved 2020-05-13, from https://doi.org/10.7554/eLife.49547 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.49547

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review*, *114*(1), 152. (Publisher: American Psychological Association)

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and cognition*, *5*(4), 418–441. (Publisher: Elsevier)

Appendices

Appendix A: An example of the task environment: A multiple-choice question and the confidence scale used

Appendix B: The formal JAGS model definition of model 2 and 3 with unrelated and linked skill and d', respectively.

```
model{


#priors of interest
#Skill distr.
S ~ dunif(1/3,1)
#mean for skill distr. (1/3 chance with 3 options)


#d' in model 2
d ~ dunif(0,2)


#d' in model 3
rs <- (S-1/3)/(2/3)
#=phi(S): rescaling S to be between 0 and 1
auc <- pnorm(rs,0,1)
#AUC=cum(rs), acc. between 0.5 and 1
d <- qnorm(auc,0,1)*sqrt(2)
#from AUC=cum(d'/sqrt(2))


#Creating the thresholds for the ratings
#Unbiased on the [0,1] line and the real line [-,]
```

```
for (c in 1:(nCat-1)){
gam[c] ← c/nCat
#[0,1] interval for uninf equal thresholds
gamReal[c] ← -log((1-gam[c])/gam[c])
#[-inf,inf] interval for uncertainty
}

#Parameters to create biased thresholds: a (slope or scale), b (intercept or shift)
a ~ dgamma(2,2)
b ~ dnorm(0,1)
#linear regression to estimate thresholds
for (c in 1:(nCat-1)){
dReal[c] ← a * gamReal[c] + b
}

#Probability of each rating given correct or incorrect answer
#for rating 1 given incorr. (0,1 for incorr. distr.) and corr (mu,lambda)
pIn[1] ← pnorm(dReal[1],0,1)
pC[1] ← pnorm(dReal[1],d,1)
#prob for rest of the ratings given incorr. and corr.
for (c in 2:(nCat-1)){
pIn[c] ← pnorm(dReal[c],0,1) - sum(pIn[1:(c-1)])
pC[c] ← pnorm(dReal[c],d,1) - sum(pC[1:(c-1)])
} #prob for last rating
pIn[nCat] ← 1 - sum(pIn[1:(nCat-1)])
pC[nCat] ← 1 - sum(pC[1:(nCat-1)])


#Data draws by trials
for (t in 1:ntrials){


# outcome of the trial based on S
O[t] ~ dbin(S,1)
#outcome of trial t (0/1) follows binomial actual skill distr.


# ratings:
#based on the outcome the corresponding probability distributions are chosen
for (c in 1:(nCat)){
pR[t,c] ← ifelse(O[t]==1, pC[c], pIn[c])
}
```

```
R[t] ~ dcat(pR[t,])
#a rating is selected following a categorical distribution


}
}
```

Appendix C: The formal JAGS model definition of model 4 with the added hypothesized effects of question difficulty.

```
model{


S ~ dunif(1/3,1)
#Sensitivity d': also expressed from S (same as before)
rs ← (S-1/3)/(2/3)
auc ← pnorm(rs,0,1)
d ← qnorm(auc,0,1)*sqrt(2)


#Creating the thresholds for the ratings
for (c in 1:(nCat-1)){
gam[c] ← c/nCat
gamReal[c] ← -log((1-gam[c])/gam[c])
}

#Parameters to create biased thresholds
a ~ dgamma(2,2)
b ~ dnorm(0,1)
#linear regression to estimate thresholds
for (c in 1:(nCat-1)){
dReal[c] ← a * gamReal[c] + b
}

#Data draws by trials
for (t in 1:ntrials){


# outcome of the trial based on S and Qd
Qd[t] ~ dunif(1/3,1)
#difficulty of question on trial t = prob of answering correct
prob_c[t] ← S*Qd[t]
```

```
#prob of answering right = product of skill and Qd
O[t] ~ dbern(prob_c[t])
#outcome of trial t (0/1) is a bernoulli trial

# rating based on adjusted scale on Qd
#transform Qd to a different scale
sQd[t] <- -log((1-Qd[t])/Qd[t])
#decreasing bias along with increasing Q difficulty
for (c in 1:(nCat-1)){
dRealt[t,c] <- dReal[c] + sQd[t]
}


#probability distributions for each rating is brought into the trial loop

pIn[t,1] <- pnorm(dRealt[t,1],0,1)
pC[t,1] <- pnorm(dRealt[t,1],d,1)


for (c in 2:(nCat-1)){
pIn[t,c] <- pnorm(dRealt[t,c],0,1) - sum(pIn[t,1:(c-1)])
pC[t,c] <- pnorm(dRealt[t,c],d,1) - sum(pC[t,1:(c-1)])
}

pIn[t,nCat] <- 1 - sum(pIn[t,1:(nCat-1)])
pC[t,nCat] <- 1 - sum(pC[t,1:(nCat-1)])


for (c in 1:(nCat)){
pR[t,c] <- ifelse(O[t]==1, pC[t,c], pIn[t,c])
}

R[t] ~ dcat(pR[t,])


}
}
```