Meta-metacognition

The regulation of confidence realism in episodic and semantic memory

Sandra Buratti

UNIVERSITY OF GOTHENBURG

Printed in Sweden by Ineko AB
Göteborg 2013

ISSN: 1101-718X
ISRN: GU/PSYK/AVH--280—SE
ISBN: 978-91-628-8711-7

DOCTORAL DISSERTATION IN PSYCHOLOGY

ABSTRACT

Buratti, S. (2013) *Meta-metacognition: The regulation of confidence realism in episodic and semantic memory.* Department of Psychology, University of Gothenburg, Sweden

The aim of this thesis was to investigate whether people have the ability to make their confidence judgments for episodic and semantic memory tasks more realistic. How realistic a person's confidence judgments are reflects how well their confidence judgments for their memory reports correspond to the actual correctness of the reports. The regulation of first-order confidence judgments by making successful second-order judgments can be seen as a form of meta-metacognition, since it aims at regulating a metacognitive process. **Study I** consisted of two experiments, and investigated whether people could increase the realism in their reports by excluding the confidence judgments they believed were unrealistic. The participants were shown a video clip and, in the Confidence task, were told to answer questions about the video and rate how confident they were that they had answered the questions correctly. Half of the participants answered two-alternative questions (recognition), and half had to come up with their own answers (recall). The participants then performed the Exclusion task, in which they were asked to exclude the 15 answers they believed had the most unrealistic confidence judgments. In Experiment 1 the recognition condition *decreased* their level of realism in their report, and in Experiment 2 the recall condition increased their level of realism. In **Study II**, the aim was to investigate whether people could increase the realism in their report by modifying the confidence judgments they believed were unrealistic. The relationship between realism of confidence and two possible memory cues, the phenomenological memory qualities *Remember*/*Know* and processing fluency, was investigated as well. The procedure was similar to that in Study I, with the exception that all participants answered recall questions and that the participants in the so-called Adjustment task were told to modify the confidence judgments they believed were unrealistic. Results showed that the participants were able to increase the realism of their confidence judgments, even though the effect was small. **In Study III**, the aim was to investigate whether people had the possibility to increase their confidence realism in semantic memory reports and whether individual differences, personality and cognitive styles, could help explain differences in this ability. The procedure was very similar to that in Study II, and the results showed that the participants only managed to increase the realism for correct items in the Adjustment task. **In Study IV**, the aim was to investigate whether the improvements in realism found in Study II could be further enhanced by giving people advice during the Adjustment task and asking them to "try more" in an Extra Adjustment task. However, results showed that although the participants managed to improve their realism like in Study II, they were not able to further improve it when given advice or by "trying more". In all, Studies II, III and IV (and to some extent also Study I) lend support to the idea that people are able to regulate the realism of their confidence judgments by making successful second-order judgments.

Keywords confidence judgments, realism of confidence, calibration, debiasing, episodic memory, semantic memory, second-order judgments, metacognition, meta-metacognition

ISSN 1101-718X ISRN GU/PSYK/AVH--280—SE

Sandra Buratti, Department of Psychology, University of Gothenburg, Box 500, SE-405 30 Gothenburg, Phone: +46 31 786 1641. Email: sandra.buratti@psy.gu.se

Preface

This thesis is based on the following four studies, which will be referred to by their Roman numerals:

I. Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive Processing, 13*(3), 243-253. doi: 10.1007 /s10339-012-0440-5

II. Buratti, S., & Allwood, C. M. (2012). Improved realism of confidence for an episodic memory event. *Judgment and Decision Making*, *7*(5), 590-601. Retrieved from http://journal.sjdm.org/11/111121a/jdm111121a.html

III. Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First- and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and Learning*, *8*(1), 79-102. doi: 10.1007/s11409-013-9096-5

IV. Buratti, S., & Allwood, C. M. (2013). *The effects of advice and "try more" instructions on improving realism of confidence*. Manuscript submitted for publication.

Sammanfattning på svenska (Swedish summary)

Varje dag gör människor så kallade säkerhetsbedömningar (i fortsättningen *konfidensbedömningar*) av hur säkra de är på olika minnes- och bedömningsuppgifter. Många människor gör dessa konfidensbedömningar som en viktig del av sin yrkesroll, ett officiellt uppdrag eller helt enkelt som en del av vardagliga göromål. En domare måste t.ex. avgöra hur säker hen är på att en patient inom rättspsykiatrin inte återfaller i brottslighet i samband med särskild utskrivningsprövning. Ett annat exempel är en läkare som måste göra en riskbedömning angående troligheten att en patient kommer att drabbas av en hjärtattack. Även ett vittne till ett brott måste t.ex. avgöra hur säker hen är på att det var den misstänkte hen såg begå brottet. I ett mer vardagligt sammanhang kanske vi överväger hur säkra vi är på att vi låste ytterdörren. Dessa säkerhetsbedömningar tillhör det man brukar kalla för metakognition, dvs. bedömning och reglering av kognitiva processer, så som t.ex. minne.

Hur väl känslan av säkerhet för att ett minne är korrekt, dvs. konfidensbedömningen, stämmer överens med korrektheten i minnesprestationen kallas för *realism i konfidens.* Realismen gäller med andra ord hur realistiska våra konfidensbedömningar faktiskt är. Studier har visat att många personer är mer säkra på sitt minne, än vad de är korrekta, de uppvisar *överkonfidens*. Detta överkonfidensfenomen har man funnit hos personer som svarar på kunskapsfrågor men även hos personer som besvarar frågor om hur de minns olika händelser.

I den här avhandlingens fyra studier undersöks huruvida människor har förmågan att förbättra realismen i sina konfidensbedömningar genom att utesluta eller ändra tidigare gjorda konfidensbedömningar genom att göra så kallade andra-ordningens bedömningar. Eftersom dessa bedömningar gäller reglering av metakognitiva bedömningar, undersöker alltså avhandlingens studier om människan har förmåga att framgångsrikt utföra meta-metakognitiva bedömningar.

I Studie I, som bestod av två experiment undersöktes huruvida människor har förmåga att förbättra realismen i sina konfidensbedömningar genom att exkludera de konfidensbedömningar de tror är de mest orealistiska. I båda experimenten fick deltagarna först se en kort film. Därefter fick de en kort instruktion om begreppet *realism i konfidens*. Under den så kallade Konfidensuppgiften fick deltagarna först besvara 50 frågor om den film de just sett. Hälften av deltagarna svarade på två-alternativs frågor (igenkänningsgruppen) och den andra hälften fick komma på svaret själva (erinransgruppen). Efter varje fråga fick de sedan skatta hur säkra de var på att de svarat rätt, dvs. de gjorde en konfidensbedömning av sitt svar. Om deltagarna inte visste svaret på frågan så skulle de gissa. I Exkluderingsuppgiften

skulle deltagarna försöka att höja realismen i sina konfidensbedömningar genom att utesluta de 15 konfidensbedömningar som de trodde var de mest orealistiska. Deltagarna blev tillsagda att den person som hade den bästa realismen för de kvarvarande 35 frågorna skulle få en extra biobiljett. Resultaten från Studie I visade att igenkänningsgruppen i Experiment 1 till och med försämrade realismen i sina konfidensbedömningar. Endast erinransgruppen i Experiment 2 lyckades statistiskt signifikant förbättra realismen i sina konfidensbedömningar, dock var effekten väldigt liten.

Studie II undersökte huruvida människor har förmågan att förbättra realismen i sina konfidensbedömningar genom att justera de konfidensbedömningar de tror är de mest orealistiska. Precis som i Studie I fick deltagarna först titta på en film samt sedan besvara 40 frågor angående filmen och för varje fråga konfidensbedöma hur säkra det var på att de svarat rätt (Konfidensuppgiften). Sedan gjorde deltagarna den så kallade Justeringsuppgiften i vilken de skulle välja ut de konfidensbedömningar de trodde var orealistiska och försöka modifiera dessa så att de blev mer realistiska. Resultaten av Studie II visade att deltagarna lyckades med att signifikant förbättra realismen i sina konfidensbedömningar även om effekten var liten. Vidare analyser visade även att förbättringen i realism inte berodde på att deltagarna använde en enkel tumregel där de endast sänkte den allmänna säkerhetsnivån. Istället visade sig deltagarna ha förmågan att identifiera konfidensbedömningar med sämre realism och sedan höja realismen i de utvalda konfidensbedömningarna.

I Studie III undersöktes om deltagarna även kunde förbättra realismen i konfidensbedömningar för kunskapsfrågor. Vidare undersöktes om det fanns någon relation mellan olika personlighetsvariabler och kognitiv stil, dvs. olika stilar för hur man väljer att processa information, och förmågan att förbättra realismen i konfidensen. Deltagarna fick svara på 40 kunskapsfrågor gällande olika kunskapsområden så som geografi, historia och liknande samt konfidensbedöma hur säkra de var på att de svarat rätt. Precis som i Studie II så fick de sedan göra Justeringsuppgiften i vilken de blev instruerade att försöka förbättra realismen i sina konfidensbedömningar genom att ändra de konfidensbedömningar som de trodde var orealistiska. Därefter fick deltagarna besvara olika enkäter angående personlighet och kognitiva stilar. Resultatet visade att deltagarna endast lyckades öka realismen för korrekta svar, dvs. de lyckades höja konfidensen för svar som var korrekta men lyckades inte sänka konfidensen för felaktiga svar. Endast svag koppling hittades mellan olika personlighetstyper och kognitiva stilar å ena sidan och realism i konfidens å andra sidan.

Studie IV, gick ut på att undersöka om effekterna i tidigare studie gick att öka genom att ge deltagarna tips om hur de ska utföra Justeringsuppgiften som t.ex. att

försöka höja konfidensen för svar de tror är korrekta och sänka den för svar de tror är inkorrekta. Deltagarna fick förutom Konfidensuppgiften och Justeringsuppgiften göra en extra Justeringsuppgift. Resultatet visade att även om deltagarna lyckades att förbättra realismen i sina konfidensbedömningar så lyckades de som fick tips inte bättre med detta än de som inte fick några tips. Inte heller lyckades deltagarna bli ännu bättre i sin realism när de fick göra den extra Justeringsuppgiften.

Sammantaget visar resultatet från avhandlingens fyra studier att människor har förmåga att förbättra realismen i sina konfidensbedömningar även om förbättringen är liten. Detta ger stöd åt att människor kan göra en andra-ordningens bedömningar av realism i konfidens, samt att de har en meta-metakognitiv förmåga.

Acknowledgements

Writing a PhD thesis is hard work, and whoever tells you otherwise is either lying or hasn't tried writing one themselves. Therefore, there are several people I would like to acknowledge, because without their help an already difficult task would have been impossible. Although numerous people have contributed in different ways to this thesis, only a few will be mentioned by name.

First, I would like to thank my supervisor, Professor Carl Martin Allwood. Your guidance and patience have been vital for my work. Thanks for always pushing me towards excellence and for never letting me settle for easy solutions. Your brilliant thoughts on different matters have challenged me to look at problems from different perspectives. You never let even a single word of mine go unchallenged, and for that I am truly grateful. Thanks for always having the time to discuss different problems with me. I honestly believe you are one of the best supervisors a PhD student can have.

I would also like to thank my second supervisor, Associate Professor Linda Hassing, for supporting me and encouraging me in my work. Your career advice has been invaluable.

I would also like to thank Professor Peter Juslin, for your valuable comments on an earlier version of this manuscript.

I would like to extend my gratitude to Dr. Sabina Kleitman for inviting me to visit the School of Psychology at Sydney University for five months. I learned a lot during this time, and the experience was very valuable.

Thanks to my statistical gurus and colleagues, Leif Strömwall and Valgeir Thorvaldsson, for your great advice on different problems.

I would also like to thank all my wonderful PhD colleagues at the department. While working on this thesis I have gotten to know many of you very well, and the sharing of experiences we have had has been invaluable for me.

I would especially like to thank Lisa Rudolfsson and Lisa Olsson, for being there for me during difficult periods. We have shared both tears and laughter, and words can't describe how grateful I am to have you as my friends.

I would also like to thank my wonderful colleagues Anne Ingeborg Berg, Martin Geisler, Jennifer Strand, Karin Grip, Elisabeth Punzi, Maria Wängqvist, Marcus

Praetorius, Amelie Gamble, Angelica Hagsand, Sara Landström and Pär Bjälkebring for the numerous pep talks, and for always having the time for a chat.

I would also like to thank my beloved parents and sister for always believing in me. Without your unconditional love and support, I would never have been able to do this. You have been there all the way, and for that I am truly grateful.

I would also like to thank Johan. I love you, and I am so grateful that we met. Your love and support have carried me this last period of my thesis work, and for that I am truly grateful. Thanks for always reminding me what is really important in life.

I would like to thank Rev. Dr. John Hirt for always believing in me, even when I did not. When my work felt impossible, I would remember how you stood at the pulpit in church and shouted to me: “Sister, you’ll get there”. And guess what? I finally got there. I made it!

I would also like to thank the congregation at Leichhardt Uniting Church and the members of Christian Students Uniting in Sydney for their support. You carried me through one of my hardest periods of being a PhD student, and for that I am truly grateful.

I would also like to thank all my friends outside the department for your patience, prayer and support. Many of you have also helped me in my work by commenting on the instructions and participating in pilot versions of the study. I would especially like to thank my friends in the Student Christian Movement (G-kriss) for offering prayer and support. You surely kept me busy during these four years.

This research would not have been possible without the approximately 800 adults who participated in the studies, so a special thank you goes out to them for showing up and for taking the time to participate.

Special thanks to the Swedish Research Council (VR) for funding this project.

Finally, I would like to thank the Lord. For as Paul said in his letter to the Philippians 4:13, I can do all things through Christ who strengthens me.

Sandra Buratti

Göteborg 21 March, 2013

Contents

Introduction

A confidence judgment expresses the level of confidence a person has for different types of performances. Every day, people make these confidence judgments in different types of contexts.

Confidence judgments of semantic memory information (knowledge memory) are often made in different learning contexts, and are an important factor in optimizing learning outcome (Wang, Haertel, & Walberg, 1990). For example, students need to judge whether the answer they have provided on a test is correct or whether a particular paragraph they have written in an assignment is good enough. Unrealistic confidence judgments may hinder them, leading them to not revise their answers on the test or rewrite their paper when needed. This may ultimately keep them from passing a course.

People also make these judgments in their profession (Allwood & Granhag, 1999) or as part of a formal duty. For example, judges need to decide how confident they are that the offender they are about to release will not commit new offenses. Physicians need to judge how likely it is that their patient will have a heart attack based on the symptoms the patient displays.

In the context of episodic memory confidence judgments are often made, more specifically, in eyewitness situations, and several studies have shown that a witness' confidence is an important factor when jurors assess the credibility of the testimony (Cutler, Penrod, & Stuve, 1988; Lindsay, Wells, & Rumpel, 1981; Wells, Ferguson, & Lindsay, 1981). It is therefore important that the witness' level of confidence that it was the accused he or she saw committing the crime should correspond well with whether or not it actually was the accused who committed the crime. In real life, the testimony of overly confident witnesses has often led to the conviction of innocent people (Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998).

Needless to say, confidence judgments can have an enormous impact on the person making the confidence judgments as well as on the people facing the consequences of the judgments. It is therefore very important that these confidence judgments be as *realistic* as possible. The realism of confidence judgments depends on their relation to the correctness of the actual performance. Confidence realism is also called confidence accuracy.

Generally speaking, there is scientific support for a persistent overconfidence phenomenon in many types of situations; that is, people are more confident than accurate about their performance (McClelland & Bolger, 1994). At a minimum, research shows that individuals often show a lack of

realism in their confidence judgments, including that some may show underconfidence (they are more correct than confident).

Given a general lack of realism in confidence judgments, it is of interest to examine the extent to which people have the ability to improve their realism in their confidence judgments of semantic and episodic memory after they have been made. This question is investigated in the present thesis. The making of confidence judgments is a metacognitive enterprise that can be defined as "any knowledge or cognitive activity that takes as its object, or regulates any aspect of any cognitive enterprise" (Flavell, Miller, & Miller, 1993, p. 150). The improvement of realism in confidence judgments by regulating previously made confidence judgments can be seen as a meta-metacognitive ability. The regulation of a first-order confidence judgment could be considered a second-order metacognitive judgment, whereby the object of the second-order judgment is to regulate a metacognitive judgment. In order words, second-order judgments could be considered meta-metacognitive judgments as they aim at regulating metacognitive judgments.

Aim of the thesis

The aim of this thesis was to investigate whether people are able to increase the realism of their confidence reports when given the possibility to regulate their confidence judgments. That is, can people increase the realism of their confidence judgments when given the freedom to exclude or adjust confidence judgments they believe are unrealistic? If it is reasonable to conclude that people have the ability to regulate the realism of their first-order confidence judgments by making successful second-order confidence judgments, this would provide support for the existence of a meta-metacognitive ability.

Before the presentation and discussion of a summary of the empirical studies forming the foundation of this thesis, different ways of assessing and measuring realism of confidence are presented. Following this is a presentation of research on confidence judgments. Then, different models and theories attempting to explain the overconfidence phenomenon are reviewed. This section is then followed by a review of previous attempts at debiasing people's confidence judgments. Finally, research regarding second-order metacognitive judgments is reviewed.

Assessing and measuring the realism of confidence

For assessing a person's level of confidence in an experimental setting, so-called confidences scales are often used. When a person is answering two-alternative recognition questions, the confidence scale often ranges from 50% to 100%, with six different confidence classes to choose from (50%, 60%, 70%, etc.). Here, 50% indicates that the person is guessing and that he or she is equally confident regarding both alternatives, and 100% means that he or she is absolutely confident that the correct answer has been chosen. In the case of directed recall tasks (when a person is to come up with his or her own answer to a memory question), the full-range confidence scale is often used. This scale ranges from 0% to 100%, and can have 11 different confidence classes to choose from (0%, 10%, 20%, etc.). In the full-range scale, 0% is often defined as a person being absolutely confident that the answer he or she has given is incorrect. According to probability theory, confidence judgments should be the same regardless of which confidence scale is used. Contrary to this, several studies have shown that overconfidence is higher when the full-range scale is used than when the half-range scale is used (Juslin, Olsson, & Björkman, 1997; Juslin, Wennerholm, & Olsson, 1999; Juslin, Winman, & Olsson, 2000).

The Brier score

Numerous measures exist for measuring realism of confidence. A majority of these measures are decomposed components of Brier's (1950) mean probability score ($\overline{PS}$), which assesses the relationship between subjective and objective probability:

$$\overline{PS} = 1/N \sum_{i=1}^{N} (r_i - c_i)^2 \quad (1)$$

In (1) N is the total number of items, r_i is the confidence judgment and c_i is the binary outcome (0 or 1) or the correctness of item i. The closer the $\overline{PS}$ score is to 0 the better the realism of confidence. Since the difference between the confidence judgment and the correctness score is squared, this tells us nothing about the direction of the deviance from perfect realism, except when the realism is 0.

The Murphy decomposition

By decomposing the Brier score, Murphy (1973) showed that it could give information on different aspects of realism of confidence:

$$(2) \qquad \overline{PS} = c(1-c) + 1/n \sum_{t=1}^{T} n_t(r_t - c_t)^2 - 1/n \sum_{t=1}^{T} n_t(c_t - c_t)^2$$

In (2) *n* is the total number of items, *T* the number of confidence classes used when assessing the probability, c is the average proportion correct (accuracy), c_t is the mean accuracy of all items in confidence class t, n_t is the number of times a specific confidence class was used, and r_t is the mean of the confidence judgments in class *t.*

The component on the far left [c (1 – c)], known as the *uncertainty* component (or the *knowledge* component, which was the term used by Lichtenstein, Fischhoff and Phillips, 1982), is seldom assessed, but measures a person's ability to choose the correct answer.

The second component, however, is commonly known as *calibration* and is the squared deviance of confidence from accuracy. It differs from the Brier score, evaluating subjective probability against a probabilistic norm. For the Brier score, the norm is deterministic (0 or 1). Thus, calibration can be seen as an aspect of realism of confidence. Furthermore, it is calculated for each separate confidence class whereby each confidence class is weighted by the number of times it has been used and then the product for each separate confidence class is summed. A value of 0 would indicate perfect calibration, and the higher the value the worse the realism of confidence. Although it is a widely used measure for assessing the realism of confidence, the calibration measure is not regarded as a very reliable measure (Bruine de Bruin, Parker & Fischhoff, 2007). One of the reasons for this is that the deviance is squared. Consequently, larger deviances have a proportionally much larger effect than smaller deviances, compared to the deviances not having been squared. The calibration measure also has other negative aspects; for instance, it is hard to intuitively comprehend and offers no information about the direction of the deviance between confidence and accuracy. However, a positive aspect of the calibration measure is that underconfidence effects on one part of the scale cannot cancel out overconfidence on the other part within a single person.

The third component in the Murphy decomposition is known as *resolution*, and gives us information about a person's ability to discriminate

between correct and incorrect responses by means of confidence judgments. A resolution score of 0 would indicate no discrimination ability at all, and the higher the resolution score the better the discrimination. Discrimination is best explained with an example. Consider Person A and Person B, who both have a mean accuracy level of 80%. Person A always gives 80% confidence judgments when assessing the probability that an event will occur, while Person B uses 60% and 100% confidence judgments and is correct 60% of the times he uses 60% confidence judgments and is correct all the times he uses 100% confidence judgments. Clearly, Person B is better at discriminating between confidence judgments for correct and incorrect items than Person A (Person B's resolution value is 0.04, while Person A's value is 0). Note, however, that even though they show different degrees of separation, Persons A and B are equally realistic in terms of the calibration measure (both have a calibration value of 0). Although resolution gives information about a different and important aspect of realism of confidence than calibration does, a problem with this measure is that it offers no information about the direction in which a person discriminates between incorrect and correct items. A person who is 100% confident every time the target event occurs and 0% confident every time the target event does not occur would have the same resolution score as a person who is 100% confident every time an event does not occur and 0% confident every time the target event actually occurs. However, this problem can be solved at least to some extent by plotting the data in a calibration diagram in which the level of correctness for each specific confidence class is plotted.

The covariance decomposition

Yates has presented an alternative decomposition of the Brier score, known as the covariance decomposition (Yates, 1982, 1988, 1994). This decomposition is basically another way to partition the calibration and resolution terms in the Murphy decomposition. Presented below is a formula in which the names of the different components have been used to shorten it:

$$\overline{PS} = c\,(1 - c) + \mathrm{MinVar}(r) + \mathrm{Scat} + \mathrm{Bias}^2 - 2[\mathrm{Slope}][c\,(1 - c)] \quad (3)$$

The measures of interest in the covariance decomposition are *Bias* (also known as over-/underconfidence) and *Slope*, which are two commonly used measures of realism of confidence. Bias can be found in the fourth term of the formula (3); however, the non-squared version is often used. Bias is written as $r - c$, in which r is the mean level of confidence for all items and c

is the mean level of accuracy for the same items. A bias value of 0 would indicate perfect realism, whereas a bias value above 0 would indicate that a person is overconfident and a value below 0 would indicate that the person is underconfident. Therefore, the bias measure gives information about the direction of deviance in confidence judgments from accuracy.

One version of the bias measure was presented by Bruine de Bruin et al. (2007), in which the absolute deviance between average confidence level and accuracy level is subtracted from 1. This is one way to discard the direction feature, which is not always useful, and still have a measure that differs from the calibration measure, as the deviance is not squared in the bias measure. A version of the bias measure, similar to the one presented by Bruine de Bruin et al., is the so-called *absolute bias*. This measure is simply the absolute deviance between the average level of confidence and the average level of accuracy from 0. This makes it easier to interpret the possible improvements in realism of confidence, since no negative values have to be considered. As 0 indicates perfect realism, the direction of the measure is the same as for the calibration measure. This absolute bias measure was used in the present thesis, since it was necessary to have a measure without the direction feature that still measures the average deviance between the level of confidence and the level of accuracy.

Slope, which is a measure of separation, can be found in the fifth term of the formula (3) and is written as $r_1 - r_2$, in which r_1 is the mean level of confidence when the target event occurs and r_2 is the mean level of confidence when the target event does not occur. A value of 1 indicates perfect separation; that is, a person assigns 100% confidence judgments to events that occur and 0% confidence judgments to target events that do not occur. A value of 0 indicates no separation at all, and a value below 0 would indicate separation, but not a very good form of separation since a value below 0 occurs when a person assigns high confidence judgments to events that do not occur and low confidence judgments to events that do occur.

Correlational measures

Other measures of confidence realism that do not originate from the Brier score are the *Goodman-Kruskal gamma correlation*, which has been used frequently within the field of educational psychology (Nelson, 1984), and the *point-biserial correlation*, which has been very popular within forensic psychology, especially in lineup research (e.g., Sporer, Penrod, Read, & Cutler, 1995).

The Goodman-Kruskal gamma correlation has been criticized for not being a reliable measure of realism of confidence (Masson & Rotello, 2009; Spellman, Bloomfield, & Bjork, 2008). Likewise, the point-biserial correlation has been criticized for its dependence on the degree of spread of the confidence judgments (Juslin, Olsson, & Winman, 1996). A person can be well calibrated regardless of the degree of spread of his or her confidence judgments on the confidence scale.

Absolute and relative accuracy measures

One can divide the measures of realism of confidence into two dimensions based on whether they measure absolute or relative accuracy (for a thorough discussion of these dimensions, see Lichtenstein & Fischhoff, 1977; Nelson, 1984, 1996; Nelson & Dunlosky, 1991). The absolute accuracy measures, such as calibration and bias, measure whether a predicted value of an item is followed by the occurrence of that same value. In contrast, relative accuracy as indicated by, for example, resolution, slope, point-biserial correlation and gamma, measures the discrimination between correct and incorrect items in level of confidence.

Since different measures assess different aspects of confidence realism, Schraw (2009) recommends that researchers use several measures when investigating it. Therefore, two different realism of confidence measures were used in this thesis, namely, the two absolute measures absolute bias and calibration. Absolute bias was used because it measures the average difference from 0 without squaring the deviance. Calibration was used because it assesses the squared deviance at each confidence level and, as mentioned earlier, hinders overconfidence effects and underconfidence effects within a person from cancelling each other out.

Research on confidence judgments

Research on the realism of confidence judgments has been conducted within somewhat separate fields of research, namely educational psychology and the psychology of judgment and decision making (Koriat, 2002). In addition, this has been an active issue in witness psychology as well (Allwood, 2010).

In Koriat and Goldsmith’s memory model (1996), confidence judgments are used as an accuracy criterion for when and when not to report a certain memory. This accuracy criterion differs depending on the accuracy

demands of the context. For example, when reporting to the police about a witnessed event, the accuracy criterion will be high and a person is likely to withhold information he or she is not confident is correct. But when telling a story to a group of friends at the local pub, the accuracy criterion may be lower and the person will feel free to report more information even if he or she is not very confident about the accuracy of the information he or she is reporting. In addition, Koriat and Goldsmith (1996) proposed that when people have the possibility to choose which information to report (free report) they can increase the accuracy of their report as opposed to when they are asked to report everything (forced report). The model has been tested empirically and has received empirical support for both adults and children performing both event memory tasks and general knowledge tasks (e.g., Koriat & Goldsmith, 1996; Koriat, Goldsmith, Schneider, & Nakash-Dura, 2001).

A persistent finding within the different fields of research on confidence judgments is that people tend to be more confident than correct, the so-called overconfidence phenomenon (e.g., Griffin & Brenner, 2004; McClelland & Bolger, 1994). This phenomenon has been found for general knowledge tasks (e.g., Kleitman & Stankov, 2001; Lichtenstein et al., 1982) as well as event memory tasks (e.g., Allwood, Innes-Ker, Holmgren, & Fredin, 2008), and has also been found for several professions, such as physicians making diagnoses and lawyers predicting the outcome of a trial (for a review see Allwood & Granhag, 1999). Often not mentioned, an underconfidence effect has been found when people make confidence judgments regarding their performance of sensory and perceptual tasks (Baranski & Petrusic, 1999; Björkman, Juslin, & Winman, 1993; Stankov, 1998).

Another common phenomenon in the research field of confidence judgments is the so-called hard-easy effect, which means that people show overconfidence for difficult tasks and underconfidence for easy tasks (Lichtenstein & Fischhoff, 1977; Merkle, 2009). A reason for this could be the so-called scale-end effect. If the task is easy, a person’s accuracy level will be high (e.g., 100%) and consequently the confidence level is highly likely to fall beneath the accuracy level, causing underconfidence. If instead the task is difficult, a person’s accuracy level will be low (e.g., 0%) and the confidence level is likely to fall above the accuracy level, causing overconfidence (Juslin et al., 2000).

Confidence judgments as an aspect of metacognition

When we make confidence judgments we make metacognitive judgments; that is, judgments about cognitions, e.g. memory reports. This term, coined by Flavell in the 1970s, was initially said to mean "one's knowledge concerning one's own cognitive processes or anything related to them" (Flavell, 1976, p. 232). With time, however, metacognition has become a highly multifaceted concept and its definitions vary extensively among researchers (for a review see Lai, 2011).

In general, metacognition can be said to constitute two different components, namely cognitive knowledge and cognitive regulation (Flavell, 1979). According to Flavell (1979), cognitive knowledge concerns, for example, knowledge about one's cognitive strengths and weaknesses, and he categorizes this type of knowledge into three categories: the first is "person" knowledge, which constitutes the beliefs we have about human beings as cognitive processors; the second is "task" knowledge, which constitutes knowledge of difference in task demands; and the third is "strategy" knowledge, which is knowledge of the types of strategies that are suitable to employ.

Other metacognitive researchers have offered a different classification of cognitive knowledge whereby it is divided into declarative and procedural knowledge (Cross & Paris, 1988, Kuhn, 2000, Schraw, Crippen, & Hartley, 2006). Here, declarative knowledge concerns, for instance, the knowledge a student has about factors that may affect his or her thinking and knowing in general, and procedural knowledge concerns the awareness and management of cognition and different cognitive strategies. The second component of metacognition, procedural knowledge, is cognitive monitoring, which concerns the planning, regulation and evaluation of one's cognition (Cross & Paris, 1988; Shraw et al., 2006). The planning part can include goal setting and the selection of adequate strategies for obtaining the goal at hand, as well as the allocation of resources. The regulation aspect, on the other hand, consists of being aware of task performance and can include self-testing, while the evaluation aspect appraises the product of the cognitive enterprise and may include revisiting or revising one's goals (Shraw et al., 2006, p. 114)

One of the most famous models of metacognition is Nelson and Narens' two-level model (1990, 1994). In this model, the meta level controls and monitors the object level; that is, the cognition level. Through the control

process the meta level *modifies* the object level, but not vice versa, and is said by Nelson and Narens to be analogous to speaking into a telephone handset. This leads to one of three actions on the object level: (1) initiating an action; (2) continuing an action; or (3) terminating an action. However, the control process does not result in any information from the object level; instead there is another process, namely the monitoring process that informs the meta level of what is occurring on the object level. This process may change the meta level's model of the situation at hand, but does not necessarily have to. Nelson and Narens proposed that this monitoring from the object level to the meta level is analogous to listening to the handset. To further explain Nelson and Narens' model, the following example can be considered: in order for a student taking a test to answer a question concerning some topic, the student needs to self-direct his or her search for the answer and thus select a search strategy for the answer. This selection of search strategy and the termination of the search are control processes. The confidence the student expresses in this answer is part of a monitoring process that will determine whether the answer is at a satisfactory level to be presented during the test or if a new search for a better answer candidate should be initiated. This is very similar to the regulation aspect of Koriat and Goldsmith's memory model (1996). Nelson and Narens claimed that the two-level model they presented could easily be generalized to more than two levels, in such a way that the meta level may be the object level of a higher meta level. In this way, some metacognitive processes dominate other processes via control and monitoring. Although the model has been highly influential within the field of educational metacognitive research, it is somewhat abstract and few of the specific processes pertaining to the model have been addressed by Nelson and Narens. Also, the strict distinction into the two processes of control and monitoring can be seen as somewhat arbitrary, and they offer no valid argument for why it should be divided into only control and monitoring and not one or more types of processes.

Cues for making confidence judgments

A number of cues can influence confidence judgments (e.g. Koriat, Nussinson, Bless, & Shaked 2008). One is the so-called processing fluency, which is the subjective ease with which a cognitive task is performed. An example of this is the subjective feeling a person has when trying to retrieve a memory (Alter & Oppenheimer, 2009). Studies have shown that confidence judgments largely seem to be based on a processing fluency cue, in which easily recalled items in knowledge tasks are given high confidence judgments (Kelley & Lindsay, 1993; Koriat, 1993). High correlations have

also been found between processing fluency and confidence judgments in studies investigating eyewitness situations (Robinson, Johnson, & Herndon, 1997; Robinson, Johnson, & Robertson, 2000).

Another cue that can be important in determining the realism of confidence is phenomenological memory quality. Two such memory qualities are “Remember” and “Know”. A memory is considered to belong to the “Remember” quality if a person recollects concrete details of the memory and to “Know” if he or she has a feeling of familiarity with the retrieved memory (Tulving, 1985). In a study investigating the realism of confidence in an eyewitness situation, a higher degree of realism was found for “Remember” answers than for “Know” answers (Seemungal & Stevenage, 2002).

Individual differences in realism of confidence

A factor that might help explain differences in the ability to improve the realism of confidence judgments is individual differences such as differences in cognitive ability, personality and cognitive styles. Stankov, Lee and Paek (2009) found low to moderate correlations between the level of realism and cognitive ability, when measured based on high school grade point average (GPA), the Scholastic Aptitude Test (SAT) and the American College Test (ACT). This indicates that a person with higher cognitive ability can be expected to show a higher level of realism in their confidence judgments than a person with lower cognitive ability. However, cognitive ability is a coarse concept, and it could be that certain aspects of cognitive ability play a more important role in the realism of confidence than other aspects do. One such aspect could be short-term memory (STM). It may be that the number of items a person can hold in his or her STM is positively correlated with realism of confidence and with the ability to increase the realism of his or her confidence. STM is easily measured through the digit span task, in which the person is asked to hold an increasing number of digits in his or her memory and then shortly after this report them. Some researchers have argued that STM and general intelligence (cognitive ability) are basically the same concept, but several researchers have argued against this. In a study by Ackerman, Beier and Boyle (2005), the authors found only a moderate correlation (r = .49) between STM and general intelligence; and when using different statistical methods on the same data, other researchers found a high correlation (r = .85) between STM and general intelligence (Oberauer, Schulze, Wilhem, & Süß, 2005). Regardless of the controversies regarding the relationship between STM and general intelligence, digit span could be

considered a coarse but suitable measure of cognitive ability that can easily be applied when investigating realism of confidence.

Little research has investigated the relationship between personality and realism of confidence. However, one can easily imagine that differences in personality would lead to differences in the expressed level of confidence and consequently in realism of confidence. However, the few studies that have investigated this issue have only found a weak relationship between different personality aspects and realism of confidence (Dahl, Allwood, Rennemark, & Hagberg, 2010; Kleitman & Stankov 2007; Pallier et al. 2002; Schaefer, Williams, Goodie, & Campbell, 2004; Want & Kleitman 2006). Some of these results will be reviewed below.

A small, but statistically significant, relationship has been found between overconfidence and individuals high in extraversion (Dahl et al., 2010; Pallier et al., 2002; Schaefer et al., 2004). This might be explained by research indicating that extraversion is associated with individuals who are active and optimistic (Costa & McCrae, 1988) and who are consequently less likely to doubt their competence in confidence judgments tasks. Similar results have been found for people high in narcissism (Campbell, Goodie & Foster, 2004); that is, people with a grandiose sense of self-importance and competence.

Studies investigating feelings of self-doubt have found negative correlations between these and confidence in different judgment tasks (Mirels, Greblo, & Dean, 2002). There are several measures that investigate slightly different types of feelings of self-doubt, such as the Judgmental Self-doubt Scale, which measures perceptions of self-doubt in one's ability to make decisions. The Self-doubt Subscale (Oleson, Poehlmann, Yost, Lynch, & Arkin, 2000) captures feelings of self-doubt concerning one's ability in general, while the Clance Imposter Phenomenon Scale assesses subjective fears of evaluation (Clance, 1985).

Individuals high in conscientiousness have a tendency to show self-discipline and act dutifully. High levels of conscientiousness have been shown to have a relation, albeit small, to overconfidence (Dahl et al., 2010, Schaefer et al., 2004). Although other studies have failed to replicate this correlation (e.g. Kleitman, 2008).

Openness is defined as a tendency to be open to possibilities and different solutions and to have intellectual curiosity. People high in this trait have been found to show a higher proportion of correct answers in different tasks and higher levels of confidence when making confidence judgments (Dahl et al., 2010; Kleitman, 2008).

Another type of individual difference measure is cognitive styles. Cognitive styles basically concern individual differences in preferences for processing information, although the meaning of the concept has been under great controversy (e.g. Riding & Cheema, 1991). One of the most well-known cognitive styles is Need for Cognition (Cacioppo & Petty, 1982), which is associated with the previously mentioned personality facet of Openness (Sadowski & Cogburn, 1997). People high in Need for Cognition enjoy engaging in and solving complex problems. The Need for Cognition style includes three components: cognitive persistence, cognitive confidence, and cognitive complexity (Tanaka, Panter, & Winterborne, 1988). Some studies have found a positive association between Need for Cognition (Wolfe & Grosch, 1990) and overconfidence, while others have not (Jonsson and Allwood, 2003). However, it may be that aspects such as enjoying engaging in complex tasks are positively related to the ability to improve the realism of confidence, since the task of regulating confidence might be a complex task that demands a willingness to engage in such tasks.

Another cognitive style is the Need for Closure (Webster & Kruglanski, 1994), which measures individuals' preference for predictability, preference for order, decisiveness and closed-mindedness. People high in Need for Closure dislike ambiguity and seek decisive and predictable outcomes. Not surprisingly, this cognitive style has been found to have negative associations with Openness (Kleitman, 2008). Thus, it is likely that people who are low in Need for Closure are more likely to succeed with the regulation of realism, as it is likely that these tasks demand an open mind to new solutions. On the contrary, though, it may also be that they, like people high in Openness, are less likely to doubt their ability and will thus not engage in the regulation of realism task in a satisfactory way.

Theories and models explaining unrealistic confidence judgments

The complexity of the overconfidence phenomenon has been discussed in recent years, and researchers have found that there are different types of overconfidence depending on which types of measures are used (Moore & Healy, 2008). The overconfidence phenomenon as such has not been generalizable over different types of measures, e.g. confidence intervals, global judgments, and confidence judgments. However, since this thesis investigates the regulation of confidence judgments in particular, the review

below will focus on attempts at explaining unrealistic confidence judgments. There are numerous theories and models that try to explain why people make unrealistic confidence judgments (for extensive reviews see Griffin & Brenner, 2004 and McClelland & Bolger, 1994). The following section will briefly review some of the most common ones. This is followed by a section on studies reporting attempts to debias people's confidence judgments. In a majority of the cases, these studies adhere to one or more of the different models and theories that try to explain why people make unrealistic confidence judgments.

The confirmatory bias model

One widely known theory about the overconfidence phenomenon is the confirmatory bias model (Griffin & Brenner, 2004), a version of which is also known as the stage model (McClelland & Bolger, 1994). According to this model, people mostly seek arguments that support their beliefs and neglect those that oppose their beliefs. This leads to inflated confidence judgments (Arkes, 1991). The best known advocates for this model are Koriat, Lichtenstein and Fischhoff (1980), who presented a three-stage model, the first stage of which entails a person who is answering two-alternative general knowledge questions searching his or her memory to locate relevant information and choose an answer. The authors proposed that in this stage people selectively tend to activate information that is in favor of a proposed answer. In the second stage, when the level of confidence in the answer is assessed, the person making the confidence judgment will attend to the activated information and continue to disregard information that is not consistent with his or her hypothesis. In the third stage, in which the person translates the confidence judgments into a numerical response, he or she will have a tendency to generally assign too-high numerical values to these judgments.

A recent study by Sieck, Merkle and Van Zandt (2007) analyzed a situation with two answer alternatives, and suggested option fixation as a contributor to the overconfidence phenomenon. This can be said to adhere to the spirit of the confirmatory bias model, but the approach is somewhat broader. According to Sieck et al., overconfidence is an effect of bias in the systematic processing of alternatives so that people tend to fixate on only one option (the favored one) in a two-alternative general knowledge task.

The ecological model

There are different versions of the ecological model, but they all share the idea that overconfidence is an artificial effect due to that representative stimuli (i.e., questions) are seldom used in experimental settings investigating overconfidence. In accordance with this, advocates of the ecological models claim that people are good judges when it comes to assessing their own knowledge (Griffin & Brenner, 2004; McClelland & Bolger, 1994). The most well-known ecological model is the probabilistic mental model (PMM) by Gigerenzer, Hoffrage, and Kleinbolting (1991, but see also, Juslin, 1993, 1994; Juslin et al., 2000).

According to the PMM theory, if the answer to a question cannot be easily derived from memory through logic, the person trying to answer it will set up a probabilistic mental model. This is done by putting the task that needs to be solved in a larger context and drawing inductive inferences. In the example given by Gigerenzer et al. (1991), a person is asked to determine which of two cities in Germany has the largest population. If a person cannot derive the answer from memory or through logic, he or she will generate a reference class, for example “Cities in Germany”, containing both answer alternatives. From the reference class the person will generate a valid probability cue, such as the soccer-team cue. It is probable that a city with a soccer team playing in the Bundesliga has a larger population than one without a soccer team in the Bundesliga. The ecological validity of this cue is 91%; that is, in 91% of the cases in which one of the cities has a soccer team in the Bundesliga and the other does not, it is the city with the soccer team that has the highest number of inhabitants. Thus, according to the PPM theory, when a person interacts with the environment the observed frequencies of facts and events become internalized and can be used as valid cues.

If the selection of questions is not representative but consists of “tricky” questions, the ecological validity cues will no longer be valid. An example of what Gigerenzer et al. (1991) call a representative sample would be the questions having been randomly drawn from an artificially but systematically generated pool of questions concerning a certain area. Ideas similar to those presented by Gigerenzer et al. in the context of PPM theory were also presented by Juslin (1993, 1994) who, like Gigerenzer et al., argued that people are good judges of their knowledge and that a non-representative sample of items will cause overconfidence.

The error model

The error model, presented by Erev, Wallsten and Budescu (1994), accounts for overconfidence as a consequence of random response error. According to the error model, the overt confidence judgments consist of the internal "true" confidence judgment and random error. Thus, even though the underlying confidence judgment is unbiased, an increase in the random response error will lead to biased confidence judgments. Since its formulation by Erev et al., the error model has been incorporated into the ecological models (Juslin & Olsson, 1997; Juslin et al., 1997).

The weight and strength model

Griffin and Tversky (1992, 2002) presented the so-called strength and weight model to explain over- and underconfidence. According to Griffin and Tversky, overconfidence is caused by people's tendency to focus on the strength or extremeness of the evidence (e.g., one very bad review of a restaurant) rather than on its weight (e.g., how many bad reviews vs. how many good reviews). The strength of the evidence could be affected by the "representativeness" heuristic. In an example provided by Griffin and Tversky, an employer judges an interviewee's ability to be a successful manager based on whether or not he or she looks like one. The weight of the evidence may then be used to adjust its strength according to the "anchor-and-adjust" heuristic. That is, the employer may realize that whether or not a person looks like a successful manager may not be the best predictive cue, and that other cues such as education and work experience should be observed. Since this adjustment process will be insufficient, the employer in the example above, will still pay the most attention to how the interviewee looks to determine his or her ability to be a successful manager. According to Griffin and Tversky, underconfidence, on the other hand, would emerge when the focus on weight is too high and/or the focus on strength is too low.

Since strength and weight are not easy to control in an experimental setting, it is somewhat complicated to test the model. Griffin and Tversky solved this issue by constructing a number of experiments with a chance set-up. In one of the experiments, participants were to judge how likely it was that a spinning coin was biased towards falling heads up. The participants were told that the coin was biased towards landing three out of five times on one side, and were given a table with a number of samples with different sample sizes (number of times the coin had been spun). The results showed that participants tended to focus more on the proportion of heads observed

(the strength) than on the sample size (weight), which led to overconfidence in their judgments when strength was high and weight was low.

Attempts at debiasing people's confidence judgments

The different attempts at increasing the realism of people's confidence judgments often appertain to different models and theories regarding the overconfidence phenomenon. Therefore, first, a review of the debiasing attempts adhering to the different models and theories just reviewed will be provided. This will be followed by a review of the debiasing attempts that are not associated with any special model or theory.

The confirmatory bias model

If the confirmatory bias model is correct, then an appropriate debiasing technique would be to make people acknowledge information that argues against their answers. Koriat et al. (1980) reported two experiments that supported their theory and technique. In the first experiment, the participants were told to come up with arguments both for and against the two answer alternatives in a general knowledge task. Results showed that the group that had to come up with these arguments had a lower level of overconfidence than a control group that did not have to come up with any arguments. In the second experiment the participants in one condition were to come up with arguments against their answers, and in the other condition the participants were to come up with arguments favoring their answers. Only the condition in which the participants came up with arguments against their answers showed a lower level of overconfidence compared to a control condition. Although these results speak in favor of the model, attempts at replicating them have failed (Allwood & Granhag, 1996; Fischhoff & MacGregor, 1982). The option fixation theory, which, as noted above, spiritually adheres to the confirmatory bias model, was empirically tested by its authors (Sieck et al., 2007). In two experiments, overconfidence was reduced when participants were asked to evaluate their answer options separately in a two-alternative answer format general knowledge task.

The ecological model

In accordance with the ecological model, studies have shown that when a randomly selected sample from a pool of general knowledge questions is

used to assess the realism of confidence, overconfidence is significantly lowered or even disappears (Gigerenzer, et al., 1991; Juslin, 1993, 1994; Juslin et al., 2000). However, several studies trying to employ a representative design have still found overconfidence, indicating an overconfidence effect above and beyond what can be explained by the ecological model (Keren, 1997; Griffin & Tversky, 2002), although this issue remains controversial.

The error model

Although it is questionable whether Erev et al.'s (1994) study can be considered an explicit debiasing attempt, the authors showed that increasing the error variance also increased the level of overconfidence and underconfidence in a data set. Erev et al. further showed that underconfidence and overconfidence could be derived from the same data set, depending on which statistical analysis was used. Minimizing random error by aggregating multiple confidence judgments of an outcome has also proved to be a successful debiasing method (Johnson, Budescu & Wallsten, 2001; Wallsten & Diederich, 2001). Another debiasing attempt that adheres to the error model is the dialectical bootstrapping model, although this debiasing attempt has not been investigated for confidence judgments specifically but rather for numerical estimations (Herzog & Hertwig, 2009). In dialectical bootstrapping a second estimate is made after questioning the accuracy of the first estimate, and then the average estimate of the two judgments is used. Herzog and Hertwig found this average estimate to be more accurate than asking participants to simply make a second judgment without instructing them to question their first judgment.

The weight and strength model

According to the weight and strength model, perfect realism would occur when there is a balance between weight and strength (Griffin & Tversky, 1992; 2002). Although the experiments testing the weight and strength model by Griffin and Tversky (1992; 2002), mentioned above, cannot be considered explicit debiasing attempts, they implicitly investigate how to increase realism of confidence by investigating when over- and underconfidence occur. The experiments lend some support to the notion that perfect realism seems to occur when there is a balance between weight and strength.

Other attempts at debiasing confidence judgments

Several other studies have tried, without success, to increase the realism of people's confidence judgments (for a previous review see Fischhoff, 1982). For example, several studies have tried fruitlessly to warn people of the overconfidence phenomenon (e.g., Fischhoff, 1982; Gigerenzer et al., 1991; Hedborg, 1996).

Some studies have tried to train people to make more realistic confidence judgments. The more successful of these studies have been concerned with giving participants feedback during training sessions. In a study by Lichtenstein and Fischhoff (1980), participants took part in 11 consecutive training sessions, each consisting of 200 general knowledge questions that the participants were to answer and then assign confidence judgments to those answers. After each session the participants were given extensive feedback regarding how they had performed on the 200 questions with respect to the realism of their confidence judgments. The feedback consisted of, among other things, the level of over-/underconfidence shown, how often they used a certain probability assessment and the mean level of confidence for correct and incorrect items. The result of each training session was also discussed with the participants for 5-20 min after each session. The study showed that a majority of the participants did improve the realism of their confidence, and that this improvement took place early in the experiment (between the first and second training sessions). In a follow-up experiment, Lichtenstein and Fischhoff (1980) used only three training sessions and still found a significant increase in realism of confidence for the participants after the training sessions. The effects of the training did not generalize very well, for example, they did not generalize to a task in which the participants were to discriminate between European and American handwriting, a task considered very similar to the one the participants had been trained in. Other attempts involving giving participants performance feedback have also led to increased realism of confidence (Benson & Önkal, 1992). A study by Stone and Opel (2000) investigated whether different types of feedback could help participants increase their confidence judgments for two-alternative art questions. The participants were given either performance feedback (feedback on their level of realism in the session) or environmental feedback (information on the event about which they are making confidence judgments). Since the participants in this study answered questions on art history, the environmental feedback consisted of a small lecture regarding art history. Whereas the performance feedback led to increased realism, the environmental feedback led to higher levels of

overconfidence. However, environmental feedback also led to the participants being able to use their confidence ratings to better distinguish between correct and incorrect answers. In a similar study by Arkes, Christensen, Lai and Blumer (1987), the participants were given feedback on questions they had answered that appeared to be easy but were actually quite difficult. After the feedback, when told to answer a new set of questions, the participants' realism had improved, although they were now slightly underconfident.

Conclusions regarding models, theories and attempts at debiasing confidence judgments

Even though many empirically supported models and theories have been presented, no particular one can be said to hold the whole solution to the challenge of increasing the realism of confidence. Instead, they each provide us with a piece of the puzzle to better understand how to make confidence judgments more realistic. In light of this, the aim of this thesis is to provide further pieces to this puzzle. The studies in this thesis are not founded on any of the previously mentioned models or theories concerned with explaining unrealistic confidence judgments in general and the overconfidence phenomenon in particular. Rather, the basis of this thesis is to investigate whether people have the ability to regulate their confidence judgments in order to become more realistic, and the overconfidence as such may have been a consequence of several of the explanations offered by the different models above. However, the studies in this thesis assume that when people attempt to regulate the realism of their confidence judgments they use different types of cues. These often derive from the retrieval of the answer to the memory questions.

Second-order metacognitive judgments

What we mean by the term second-order metacognitive judgment is a judgment that regulates a first-order metacognitive judgment (e.g. the accuracy of confidence judgment, judgments of learning, etc.). Just as metacognition can be referred to as "any knowledge or cognitive activity that takes as its object, or regulates any aspect of any cognitive enterprise" (Flavell et al., 1993, p. 150), an activity that targets the regulation of a

metacognitive process, which is the nature of second-order judgments, can be referred to as meta-metacognition.

Although Nelson and Narens (1990,1994) never used the term meta-metacognition when describing the generalization of their two-level model to incorporate more levels, they laid the foundation for how this term could be used. The idea that metacognitive judgments can also be regulated is interesting, and in recent years a humble amount of research has provided support for the idea of a meta-metacognitive ability (Cesarini, Sandewall & Johanneson, 2006; Dunlosky, Serra, Matvey, & Rawson, 2005; Miller & Geraci, 2011). Cesarini et al. (2006) asked their participants to adjust their original confidence interval in assessing how many questions they would answer correctly, and this adjustment of confidence intervals resulted in a significant decrease in overconfidence. In the study by Dunlosky et al. (2005), the authors reported successful second-order assessments of judgments of learning of paired associates. Miller and Geraci (2011) found that low-performing students showed higher overconfidence in their first-order prediction of their total exam performance than did high-performing students. However, interestingly enough, the low-performing students were more accurate in their second-order judgments of these first-order predictions.

The above studies are all examples of successful second-order judgments, and provide support for the idea of a meta-metacognitive ability in the sense of a process that regulates a metacognitive process. This process can generally be said to be divided into two different parts: first, the *identification* of a metacognitive judgment in need of adjustment; and second, the *adjustment* of the metacognitive judgment. The identification part may be an example of Nelson and Narens' (1990, 1994) monitoring process while the adjustment part may exemplify Nelson & Narens' control process.

Summary of the studies

The four studies of this thesis share many common features. Therefore, with the aim of reducing repetitiveness in their presentation, common methodological features in the procedure will first be addressed.

General method

Introduction to realism of confidence. In all four studies, all participants were given an introduction concerning the concept *realism of confidence*. As part of these instructions, they were given examples of people

being underconfident, perfectly realistic and overconfident when answering a knowledge quiz. They were also given examples of a person being underconfident, perfectly realistic and overconfident when answering specific questions regarding a film he or she had just seen in Studies I, II and IV and specific questions regarding geography in Study III. This introduction was basically the same in all four studies, with the exception of Study I Experiment 1, which did not contain as many examples as the latter version did.

The *realism of confidence* knowledge test. To measure how well the participants understood the concept *realism of confidence*, they were all given a test on this. This test was administered after the introduction of the concept, mentioned previously, and was used in all studies with the exception of Study I Experiment 1, in which no such test was administered. The test consisted of ten items in Study I Experiment 2, and Studies II and III, and of 11 items in Study IV.

Confidence task. After the realism of confidence knowledge test, the participants were asked to answer questions concerning different types of memory. In Studies I, II and IV the participants were presented with episodic memory questions concerning a video clip they had seen at the beginning of the experiment. In Study III the participants instead answered semantic memory knowledge questions on a wide variety of topics. If they did not know the answer, they were asked to guess.

After each answer, the participants were told to rate on a confidence scale, described further below, how confident they were that their answer was correct. They were told that the person with the highest proportion correct after this task would receive an extra movie ticket.

Confidence scale used in the Confidence task. The confidence scale in Study I Experiment 1 ranged from 0% (“I’m absolutely sure my answer is incorrect”) through 50% (“Guessing”) to 100% (“I’m absolutely sure my answer is correct”). In Study I Experiment 2 , for the recognition questions the participants’ confidence scale ranged from 50% (“Guessing”) to 100% (“I’m absolutely sure my answer is correct”), and for the recall questions the confidence scale ranged from 0% (“Guessing”) to 100% (“I’m absolutely sure my answer is correct”). In Studies II, III and IV the confidence scale ranged from 0% (“I’m absolutely sure my answer is incorrect”) to 100% (“I’m absolutely sure my answer is correct”).

The Exclusion task. After the Confidence task the participants proceeded to the regulation task, in which they were to try to improve their realism of confidence. The regulation task used in Study I was called the

Exclusion task. In this task, the participants were asked to exclude the 15 items from the Confidence task that they believed had the most unrealistic confidence judgments. They were also informed that the person with the best realism of confidence after the Exclusion task would receive an extra movie ticket.

The Adjustment task. The regulation task used in Studies II, III and IV was called the Adjustment task. In this task, the participants were asked to modify the confidence judgments they believed were unrealistic in the Confidence task by assigning new confidence judgments to their answers. They were told that they could modify as many confidence judgments as they wanted, but it was recommended that they change at least 20 (Studies II and III) or 25 (Study IV). Yet, if the participants felt they had been perfectly realistic, or wanted to change fewer or more than this, they were to feel free to do so. The reason the participants were asked to change a certain number of confidence judgments (20 or 25, depending on the study) was that there should be a substantial number in order to detect a difference. They were also told that if they decided to change a confidence judgment they had to change it by at least ten percent units, but if they wanted to they could make a greater change than this. They were also informed that the person with the best realism of confidence after the Adjustment task would receive an extra movie ticket.

Study I

Aim. Study I consisted of two experiments, and the aim was to investigate whether it was possible for a person to increase the realism of his or her confidence judgments by choosing which confidence judgments to report. This idea can be seen as a generalization of Koriat and Goldsmith's memory model (1996), in which confidence judgments are used to regulate the accuracy of memory reports. The study also investigated whether there was a correlation between cognitive ability, as measured in the digit span task, and the ability to increase the realism of confidence.

Method. In Experiment 1 the participants were 135 adults. The experiment had four conditions: two for recognition and two for recall. The recognition and recall conditions were analyzed in separate 2×2 mixed ANOVAs. Two variables were investigated: the *within-participant* variable Task (Confidence task vs. Exclusion task) and the *between-participants* variable Group (confidence judgments with/without realism rating).

The participants were first shown a short video clip of a women being kidnapped, and after a filler task they were given instructions on the concept

realism of confidence. They performed the Confidence task, in which they answered 50 questions regarding the video clip they just had seen. In two of the conditions (the recognition conditions) they answered two-alternative questions, and in two others (the recall conditions) they answered directed recall questions. In one of the recognition conditions and one of the recall conditions they also, after each confidence judgment, rated how realistic they believed their confidence judgment was.

In Experiment 2, with 135 adults, the procedure was the same as in Experiment 1, with the exception that the filler task was switched to a digit span task in order to measure cognitive ability. The participants were given instructions on the concept *realism of confidence* and a test concerning how well they understood this concept. They then performed the Confidence task and the Exclusion task. The design of Experiment 1 was the same as in Experiment 2, with the exception that there were only two conditions: one for recognition and one for recall.

Results. In both Experiments 1 and 2, the participants showed overconfidence. In the recognition conditions in Experiment 1 the participants significantly *decreased* the realism of their confidence, as measured with the absolute bias measure, when asked to exclude the 15 confidence judgments they believed were the most unrealistic. In the recall conditions, there was no significant difference in the confidence realism between the Confidence task and the Exclusion task. The analyses showed that the participants in fact increased the level of confidence from the Confidence task to the Exclusion task, indicating that they chose to keep answers with high confidence judgments. In Experiment 1, no differences were found between the conditions with respect to the realism measures between the participants who made an additional realism rating of their confidence judgments and those who made no additional rating.

In Experiment 2, there was no significant difference in the recognition condition between the Confidence task and the Exclusion task. However, the recall condition managed to increase the realism of their reports, measured with the calibration measure. Furthermore, the analyses showed that the participants increased the level of confidence from the Confidence task to the Exclusion task, indicating that they chose to keep answers with high confidence judgments. Analyses conducted with only the participants who had gotten all items correct on the knowledge test concerning the concept *realism of confidence* did not render any different results. Furthermore, no correlation was found between cognitive ability, as measured in the digit

span task, and the ability to increase the realism of confidence in the Exclusion task.

Study II

Aim. The study investigated whether people had the ability to increase the realism of their confidence judgments by modifying the confidence judgments they believed were unrealistic. Two cues for making confidence judgments were also investigated, namely processing fluency and the phenomenological memory quality (Remember/Know).

Method. The participants consisted of 200 people. The study had a mixed 3×2 design with the between-participant variable Condition (control, Fluency, Remember/Know) and the within-participant variable Task (Confidence task and Adjustment task). The participants were shown a short video clip depicting a theft in a park, after which they performed a filler task. After this they were given instructions about the concept *realism of confidence* and then filled out the knowledge test concerning the concept. During the Confidence task, they answered 40 recall questions. In the control condition, the confidence scale was followed by the next question. In the Fluency condition, the participants were told to rate how easy/difficult it had been to retrieve the answer. In the Remember/Know condition the participants were told to rate the phenomenological memory quality ("Remember", "Know" or "Guessing") associated with the answer. Next, they proceeded to the Adjustment task.

Results. The results showed that on average the participants were overconfident after performing the Confidence task. Furthermore, the results showed that the participants increased the realism of their confidence when asked to modify the confidence judgments they believed were unrealistic, as measured with the absolute bias and calibration measures. However, it was only the control condition that managed to do this. The analyses also indicated that the increase in realism was not just the consequence of a simple heuristic, by which the participants merely decreased their average level of confidence from the Confidence task to the Adjustment task. Rather, further analyses showed that the participants had the ability to choose the confidence judgments that were more biased than the confidence judgments not chosen for modification. In addition, they then managed to increase the realism in the chosen confidence judgments. Moreover, the results showed that the participants targeted more incorrect than correct items. No differences in the results were found when the analyses were conducted with

only participants who had gotten all items correct on the knowledge test regarding the realism of confidence concept.

The results also showed that the reported processing fluency for the confidence judgments chosen to be modified was low, indicating that processing fluency may have been used as a cue for choosing which confidence judgments needed to be modified. The results from the phenomenological memory quality analyses indicated that confidence judgments belonging to "Know" responses were chosen to be modified more often than those belonging to the other two phenomenological memory qualities. The analyses also indicated that the degree of realism was the worst for the "Know" responses. With respect to the level of realism of confidence in the "Remember" and "Guessing" answers, there was no difference in level other than that there was a slight overconfidence for "Remember" answers and a slight underconfidence for "Guessing" answers.

Study III

Aim. This study investigated whether people could successfully regulate the realism of semantic memory reports by modifying the confidence judgments they believed were unrealistic. Furthermore, we also investigated whether differences in personality and cognitive styles could explain differences in the ability to make successful second-order judgments.

Method. The 151 participants were given a short presentation regarding the concept *realism of confidence*, and then answered the knowledge test measuring this concept. They then performed the Confidence task, answering 40 general knowledge questions on different topics such as sports, geography, etc. Next, they proceeded to the Adjustment task. After this they filled out several personality questionnaires, such as Goldberg's Big Five Markers (Goldberg, 1992), the Judgmental Self-doubt Scale (Mirels et al., 2002), the Self-doubt subscale (Oleson et al., 2000), the Clance Imposter Phenomenon Scale (Clance, 1985) and the Narcissistic Personality Inventory (Raskin and Terry, 1988). They also filled out two cognitive styles questionnaires, namely the Need for Cognition (Cacioppo & Petty, 1982) and the Need for Closure Scales (Webster and Kruglanski,1994).

Results. The results showed that on average the participants were slightly overconfident after performing the Confidence task. Furthermore, the results showed that on average the participants did not manage to increase their realism, as measured with the absolute bias measure. They did, however, manage to significantly increase their confidence for correct answers, thus improving the realism for these answers. Further analysis

showed that the participants identified an average of 46% of the potential deviation from perfect realism for correct answers and 39% of the potential deviation for incorrect answers. Less than 1% of the correct items with 100% confidence and 1% of the incorrect items with 0% confidence were chosen for adjustment.

The factorial structure of the personality measures and cognitive styles measures was determined using exploratory factor analysis. A five-factor solution, explaining 71.6% of the total variance, was suggested by the latent root criterion. The factors were: *Self-doubt, Openness, Conscientiousness, Extraversion/Narcissism,* and *Agreeableness.* The factors did not correlate with the ability to identify items in need of adjustment. However, the Openness factor predicted the participants' level of confidence in both the Confidence and Adjustment tasks. Interestingly, people high in Openness showed high confidence for incorrect items in both the Confidence and Adjustment tasks. Extraversion/Narcissism was also found to correlate positively with first-order confidence judgments for correct items.

Study IV

Aim. This study investigated whether participants had the ability to improve the realism in their confidence judgments for an episodic memory task, and whether this effect could be further improved by giving them advice. This advice concerned asking them to attend to the confidence level of both incorrect and correct items during the Adjustment task, as well as telling them to focus on the phenomenological memory quality the recalled answer had. The effect of asking the participants to engage further in the task was also assessed.

Method. The participants, 220 adults, were shown a video clip of a kidnapping. Next, they were instructed on the *realism of confidence* concept and then took the knowledge test concerning this concept. They then performed the Confidence task, in which they answered 50 directed recall questions on the video clip. After this, they performed the Adjustment task. In the control condition, participants were not given any advice regarding how they should go about making these adjustments. In the Correct/Incorrect condition, participants were advised to lower the confidence for the answers they believed were incorrect and increase the confidence for answers they believed were correct. In the Remember/Know condition, it was explained to participants what phenomenological memory quality meant and they were advised to focus on modifying confidence judgments to answers with a "Know" quality since these are often associated with less realism than

“Remember” answers. All participants then proceeded to the last task, the Extra Adjustment task, in which they were told that previous studies have shown that when people are told to try to adjust the realism in their confidence judgment the effect is small, and that there is often room for improvement after the Adjustment task. Thus, they now had a further opportunity to improve the realism in their confidence judgments. They were told that the participant with the best realism in confidence after this task would receive an extra movie ticket.

Results. After performing the Confidence task, the participants were on average overconfident. Furthermore, they increased the realism in their confidence judgments, as measured with the absolute bias measure, after performing the Adjustment task. The improvements had no effect on the calibration measure. Furthermore, when confidence was investigated for correct and incorrect items separately, there was only a significant decrease in confidence for incorrect items and no difference in level of confidence for correct items. There was no difference between the three conditions; thus the participants in the two advice conditions did not manage to further improve the realism in their confidence. The Extra Adjustment task did not help to improve the participants’ realism in confidence; rather, it significantly worsened the realism as compared with the Adjustment task. However, after the Extra Adjustment task the participants were still significantly more realistic than after the Confidence task.

When it came to identifying the items in need of adjustment, calculations showed that the participants, depending on condition, managed to identify a proportion of 41%-48% of the total confidence deviation from perfect realism for correct items after the Adjustment task. For incorrect items, the proportions were 31%-36%. In the Extra Adjustment task, the participants identified 25%-28% of the total deviation in the Adjustment task for correct items and 17%-18% of the total deviation from realism for incorrect items.

Analyses showed that the participants were better at identifying both correct and incorrect items in need of adjustment in the first Adjustment task than in the Extra Adjustment task.

However, they were successful in not choosing items that were not in need of adjustment, since only a small proportion of correct items with 100% first-order confidence judgments were chosen for modification (3% in the Adjustment task and only 2% in the Extra Adjustment task). For incorrect items, 6% of the items with 0% first-order confidence judgments were

chosen for modification in both the Adjustment task and the Extra Adjustment task.

Regarding the regulation aspect, analyses showed that participants were better at modifying the confidence for incorrect items in the Adjustment task compared to the Extra Adjustment task. No difference was found for correct items.

General discussion

This thesis explored people's ability to regulate the realism of their confidence judgments of their memory reports by being able to make successful second-order judgments of their first-order confidence judgments. The making of successful second-order confidence judgments could be seen as a form of meta-metacognition.

Two different methods of making second-order confidence judgments were used in this thesis. The first employed an Exclusion task whereby participants were asked to exclude the confidence judgments they believed were the most unrealistic. The second employed an Adjustment task whereby participants were asked to adjust the confidence judgments they believed were unrealistic.

Below is an integrated discussion of the results of the four studies. First, results regarding the main aim of the thesis are discussed, namely the making of successful second-order confidence judgments. This is followed by a discussion of the differences found between the regulation of episodic and semantic memory. Then, results regarding the use of recognition and recall questions are discussed, followed by a discussion regarding the effect of cues and individual differences on the improvement of realism. After this, attempts to further improve the realism of confidence are addressed. Finally, concluding remarks are made, the limitations of this thesis are addressed and future directions are suggested.

The making of successful second-order confidence judgments

In Study I the task used for trying to increase the realism of confidence was the Exclusion task, whereby participants were asked to exclude the 15 items they believed had the worst realism. This task can be seen as a generalization of Koriat and Goldsmith's memory model (1996), in which

confidence judgments are used to regulate which memories should be reported. However, in Study I the regulation was not of memory reports *per se* but was done in order to improve the realism in confidence. Contrary to our expectations, the results showed that it was only the participants in the recall condition in Experiment 2 who managed to increase their confidence realism, and that the recognition condition actually worsened their realism of confidence as a consequence of the Exclusion task. The differences between recognition and recall answers will be discussed further in the recognition and recall section of the general discussion.

A more successful method for increasing the realism was the Adjustment task used in Studies II, III and IV. This task includes both the *identification* of confidence judgments in need of adjustment (lacking realism) and the *adjustment* of the confidence level of these items in order to increase the realism of confidence. For the second-order judgment to be successful, both the identification and the adjustment parts of the task need to be successful.

Study II showed that participants were actually able to increase the realism in their confidence judgments, as measured with both the absolute bias and the calibration measures, by being able to adjust the confidence judgments they believed were unrealistic. These results are in line with those found for the making of successful second-order metacognitive judgments, such as judgments of learning (Dunlosky et al., 2005), confidence intervals (Cesarini et al., 2006) and global judgments (Miller & Geraci, 2011).

In Study III there was no significant improvement on the absolute bias measure. However, the participants did manage to increase the confidence for *correct items*, thus increasing the realism in their confidence judgments for these items. The results of Study II were partly replicated in Study IV, as the participants managed to increase the realism in their confidence judgments. However, this improvement was only found on the absolute bias measure. The lack of improvement in calibration measure could be due to a floor effect, since the participants were very realistic to begin with, as measured with this measure.

In general, the effect sizes of the improvements in realism in Studies II, III and IV were small. Similarly, research shows that in previous debiasing attempts in which people have tried to improve the realism of their confidence judgments by receiving feedback, this has not resulted in large improvements (e.g. Lichtenstein & Fischhoff, 1980). There are at least two reasons why the improvement in Studies II, III and IV was so modest. First, the participants did not change as many confidence judgments as they could

have. In addition, as indicated by the realism measures after the Adjustment task, there was still room for improvement. The formula used in Studies III and IV also indicated that the participants did not identify all items that deviated from perfect realism. Second, the realism levels were quite good in all studies after the first-order confidence judgments, thus making improvement more difficult. Therefore, in future studies investigating improvements of realism, tasks with higher initial bias levels might be chosen in order to increase the sensitivity of the investigations. It would also be interesting to investigate tasks for which there is an evident underconfidence effect (Baranski & Petrusic, 1999; Björkman et al., 1993; Stankov, 1998) in order to study whether participants are able to improve the realism in these tasks as well.

When it comes to the validity of the two regulation tasks (the Exclusion task and the Adjustment task) it could be argued that the Adjustment task is more valid from an ecological perspective. In real life it is more likely that we are given the chance to adjust our confidence judgments (Studies II, III and IV) rather than simply excluding them altogether (Study I). For example, a witness in court can be asked to evaluate – and consequently might adjust – a confidence judgment of a statement already made, if he or she feels this is necessary. However, with respect to an educational setting, the version of the regulation task used in Study I, the Exclusion task, may have some validity. When answering a test, a student might need to reevaluate and, if necessary, modify his or her confidence judgments regarding the answer alternatives he or she deems likely. However, it is possible that the student will decide not to answer the question at all or to provide a new answer as an effect of losing confidence in the original answer. On a similar note, a recent study has investigated how students can regulate their accuracy by excluding certain answers via a plurality option technique (Higham, 2013).

In all, the results of Studies II, III and IV provide further support for a meta-metacognitive ability. This ability can be divided into two parts, shown in Figure 1, namely identification and adjustment. The identification part can be seen as incorporating what Nelson and Narens (1990,1994) consider the monitoring processes on a meta level (the accuracy of a cognition, in this case a confidence judgment). That is, the identification concerns the observation and evaluation of the accuracy of the confidence judgment; that is, its realism. The adjustment part concerns the adjustment of the confidence judgment to better fit the accuracy level. This process may not differ from a first-order confidence judgment, and may therefore only be a metacognitive

judgment rather than a meta-metacognitive one. Thus, the adjustment part bears similarities with the control process in Nelson and Narens' model.

We do not know which specific processes underlie the regulation of confidence. The error models have shown that when people make several confidence judgments their realism improves, since the random error decreases (Johnson et al., 2001; Wallsten & Diederich, 2001). Speculatively, it could be that the increase in realism is an artifact of merely redoing the first confidence judgment. However, studies have shown that merely redoing a confidence judgment does not lead to better realism (Allwood, Granhag, and Johansson, 2003). On a similar note, researchers investigating the dialectical bootstrapping model (DBM) found that when participants were asked to merely redo their estimations they did not improve their judgments (Herzog & Hertwig, 2009).

It may be that some of the participants followed the procedure detailed in the DBM, reported by Herzog and Hertwig (2009). DBM draws on previous research showing that the average of the predictions of many individuals is better than a typical estimate in the same group. In the DBM, a participant first must answer a numerical estimation question, and then after some time provide a new estimation while considering that their first estimation is off the mark and thinking about reasons why this could be the case. Based on this the participants are to make a second estimate, and then the average of these estimations is used. However, it may seem somewhat unlikely to argue that most participants carried out the final part of the Bootstrapping model (averaging). Instead, their ability to regulate the realism in their confidence judgments may be explained by the fact that they, when asked to increase their realism, engaged in different associations compared to when they did the Confidence task. If so, two reasons for this may be: first, that the activation pattern in their long-term memory (LTM) was somewhat different when they attempted the Adjustment task compared to when they attempted the Confidence task; and second, that the two tasks as such were slightly different.

The DBM and our Adjustment task, used in Studies II, III and IV, are similar in that they require two responses that to some extent draw on somewhat different knowledge. However, they differ in two important ways. First, the Adjustment task includes identifying the confidence judgments to adjust rather than considering an answer to a knowledge task. Second, the DBM in the final step crucially involves averaging the two estimates.

However, further research is needed to investigate which processes underlie the regulation of confidence realism.

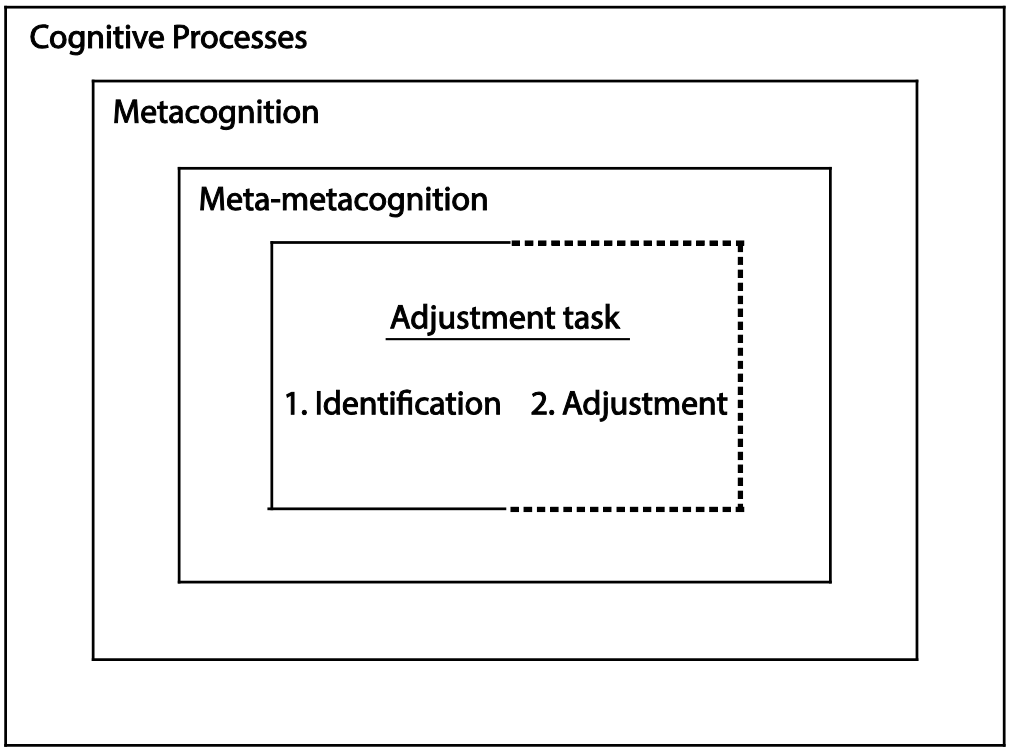

Figure 1. The conceptual relationship between the two parts of the Adjustment task, meta-metacognition and metacognition. Figure from Buratti, Allwood and Kleitman (2013).

Differences in regulating realism between episodic and semantic memory tasks

An interesting result in this thesis was the difference found in the regulation of realism between episodic and semantic memory tasks. For episodic tasks participants were better at increasing the realism for incorrect items (Studies II and IV), and for semantic memory tasks they were better at increasing the realism for correct items (Study III). Some researchers have theorized that there are differences when it comes to making confidence judgments for episodic and semantic memory answers, for example the availability of encoding possibilities, feedback and heuristics used (Perfect, Watson & Wagstaff, 1993, Perfect, 2004). Similar differences can be important during the regulation of confidence judgments.

One reason for the differences found in the regulation of confidence for episodic and semantic tasks is differences in the form of feedback received in

the context of the different memory tasks. In similarity with the confirmatory bias model (Koriat et al., 1980), it may be that confidence judgments for semantic memory information are associated with more confirmatory feedback than episodic memories. For semantic memory information, it is easier to establish whether an answer is correct or not since a person is expected to receive feedback confirming correct answers more often during their lifetime and thus has more possibilities to encode the correct answer compared with incorrect answers. For example, knowing that Paris is the capital of France is a fact that a person can receive confirmatory feedback on in many different contexts, e.g. while learning at school, when scheduling a holiday, etc. Thus, knowing whether a correct answer is correct is often easier in a general knowledge task than knowing whether an incorrect answer actually is incorrect, since there is a lack of confirmatory feedback in the latter case. Contrary to semantic memory tasks, for episodic memory tasks individuals do not have the same possibility to have received confirmatory feedback for single questions during their lifetime. Moreover, the multiple encoding possibilities may not exist in the same way as for semantic memory. For example, a person experiencing an event may only experience the salient features of the event once. For example, the person that they see during the event they may never have seen before. Thus, a lack of confirmatory feedback and of encoding possibilities in episodic memory might make it hard for people when they try to regulate the realism of their confidence judgments of answers to episodic memory questions. This may cause them to instead give more attention to answers that may be incorrect. Moreover, it may also be easier for people during the regulation task to be able to target what stands out and what seems unlikely; that is, answers they believe are incorrect.

Another reason why the incorrect answers were targeted more during episodic memory tasks could be that participants focused more on the processing fluency cue for episodic reports than for semantic memory reports. Since many people have beliefs regarding how well they perform on different semantic memory tasks, it may be more likely that they let these so-called information-based cues guide them in making confidence judgments for semantic reports. However, when there is a lack of belief regarding how well one performs during a task (which might be true to a larger degree for episodic memory), it is likely that these confidence judgments will be based on external cues such as processing fluency (Koriat et al., 2008). As Study II showed, processing fluency seemed to be a cue for choosing which items to regulate during the Adjustment task.

Regulating the realism when using recognition and recall questions

In Study I the recognition conditions did not manage to increase the realism of the participants' confidence judgments; in Experiment 1 the participants in the recognition condition even worsened their realism significantly, and the recall condition in Experiment 2 managed to increase their realism. Other studies in this thesis showed that people were able to increase their realism for recall questions in both episodic (Studies II and IV) and semantic memory contexts (Study III). Even though recognition questions were only used in Study I, one might be tempted to draw the conclusion that this type of question is not ideal when trying to improve the realism of confidence. However, in this thesis the regulation of realism when using recognition questions was only assessed by asking the participants to perform the Exclusion task. As the results from Study I showed, this task rendered very small improvements effects for the recall condition as well (only significant when non-parametric tests were used).Thus, it is difficult to conclude whether it is the recognition questions or the Exclusion task (or both) that caused the unsuccessful attempts at improving the realism in these conditions. Interestingly, the recognition conditions in Experiment 1 actually *decreased* the participants' level of realism of confidence. However, this finding was not replicated in Experiment 2. One reason for this could be that the confidence scales were different in Experiment 1 and Experiment 2. Since the start and end of the full-range confidence scale (i.e., ranging from 0%-100%) used in Experiment 1 contained more unrealistic responses, excluding confidence judgments from the middle of the scale ("Guessing") could lead to an increase in absolute bias, which is what happened in Experiment 1. Since the scale in Experiment 2 was half-range, going from 50% ("Guessing") to 100%, the exclusion of items from and near the 50% confidence class would not lead to the same significant decrease in realism. This in turn could explain why the decrease in level of realism in the recognition conditions was not replicated. However, contradicting this, an extra analysis (not presented in the results section above), not including the confidence judgments below 50% in the recognition conditions in Experiment 1, showed that the participants still decreased the level of realism as measured with the absolute bias measure. Presently, there is no clear explanation for why the decrease in the recognition conditions in Experiment 1 was not replicated in Experiment 2.

Cues for increasing the realism of confidence

Confidence judgments are often based on different types of cues (Koriat et al., 2008). In Study II we investigated two such different types, namely processing fluency and phenomenological memory quality (Tulving, 1985). When investigating processing fluency as a cue for making confidence judgments in Study II, a high positive correlation between the reported processing fluency and the level of the confidence judgments was found. This correlation could indicate that processing fluency is used as a cue for assessing the correctness in the memory report. When a memory is easily retrieved, people are more confident in its correctness. The results also showed that items chosen to be modified had a lower processing fluency score than items not chosen to be modified. This indicates that processing fluency may have been used as a cue for identifying the items in need of increased realism since the participants chose to modify confidence judgments with lower processing fluency scores. However, it should be said that – based on the results from Study II – it is not possible to say for certain that it was processing fluency that affected confidence judgments and not the other way around, that it was the confidence judgments that affected the subjective feeling of ease. However, there is evidence from other experimental studies supporting the possibility that it is processing fluency that affects confidence judgments (Kelley & Lindsay, 1993; Koriat, 1993). It should be noted that even if the participants in the Fluency condition in Study II chose the confidence judgments with the lowest processing fluency score, this may not have had the expected effect, since they did not manage to increase the realism of their confidence judgments in this particular condition. This could be due partly to a floor effect, since the bias and calibration level was lower for this condition than for the control condition. It could be that the processing fluency rating aided participants in making more realistic first-order judgments in the Confidence task. These issues should be further investigated in future research.

The results of Study II also showed that the confidence judgments for the items associated with the memory quality “Know” were chosen to be modified more often than those associated with the other memory qualities. In accordance with previous findings, the realism was the worst for “Know” answers (Seemungal and Stevenage, 2002). The reason the confidence judgments for “Know” responses were chosen to be modified more often could be that the phenomenological memory quality was used as an implicit or explicit experience-based cue (Koriat et al., 2008). Regardless of whether it was an implicit or explicit cue, it was still a valid cue for increasing the

realism of confidence. However, using “Know” as a cue did not have the expected effect since the participants in this condition did not manage to increase the realism of their confidence judgments. But the lack of such of effect could be due to the same floor effect that might have affected the Fluency condition, since the Remember/Know condition in Study II was also found to be highly realistic even during the Confidence task. Again, these issues should be further investigated in future research.

Yet another possible cue for increasing the realism of confidence emerged during Study I. The results indicated that the majority of confidence judgments chosen to be excluded came from the confidence class labeled “Guessing” in the confidence scale. During Experiment 1 the “Guessing” label was applied to the 50% confidence class in both the recognition and recall conditions, and during the Exclusion task more items from this confidence class were excluded. However, when the “Guessing” label was changed from the 50% confidence class in Experiment 1 to the 0% confidence class in Experiment 2 for the recall condition, this led to a change in the exclusion pattern. Now, more items from the 0% confidence class were selected than from the 50% confidence class in the recall condition. This result may indicate that the participants used the “Guessing” label as a superficial cue for doubting an item’s realism. In Experiment 1 the “Guessing” confidence class was not one of the most biased classes; however, in Experiment 2 when the label was changed in the recall condition and applied to the 0% confidence class this was in fact one of the more biased confidence classes. Thus this superficial cue, choosing items from the confidence class labeled “Guessing”, might partly explain why there was an increase in realism for the recall condition in Study I Experiment 2 but not for any of the recognition conditions.

Finally, it should be noted that the realism regulation perspective may offer new ideas for investigating not only which cues are the basis for second-order confidence judgments but also which cues are the basis for first-order judgments.

The effect of individual differences on the ability to increase the realism of confidence

In this thesis, the relationship between various individual difference aspects, such as cognitive ability (Study I), personality and cognitive styles (Study III), and the regulation of confidence realism, was investigated. In Study I no relationship was found between cognitive ability (as measured in the digit span task) and the ability to increase the realism of confidence.

However, although no correlation was found between the digit span task and the ability to increase the realism of confidence, there may still exist some relationship between cognitive ability and the ability to increase the realism of confidence. It may be that the digit span task did not capture aspects of cognitive ability that are important when trying to increase the realism of confidence. One such aspect may concern the executive function, such as the ability to update information or the sensitivity to different retrieval cues.

In Study III, only a weak link was found between personality and cognitive style variables on the one hand and first- and second-order confidence judgments on the other. The Openness factor was found to predict the confidence level in both the Confidence task and the Adjustment task. Most interesting was that this factor predicted confidence for incorrect items but not for correct items. The reason for this result could be that people high in Openness are less likely to doubt their ability and intellectual competence, making them less likely to doubt the correctness of their answer.

The Extraversion/Narcissism factor also predicted confidence in first-order judgments for correct items. This result is in line with previous results indicating that people high in extraversion (Pallier et al., 2002 Schaefer et al., 2004) and narcissism (Campbell et al., 2004) tend to show overconfidence. However, in line with our general results concerning the weak association between personality and metacognitive realism, Dahl et al. (2010) only found very low associations between personality and first-order confidence judgments.

Furthermore, Study III found no relationship between participants' ability to identify items in need of adjustment and the personality and cognitive style factors. One of the reasons for this result could be that there were two floor effects. First, there was little room for improvement in the Adjustment task due to the very realistic confidence level after the Confidence task. Second, the number of adjustments was quite low.

The effect of trying to enhance the improvements in realism

Since the improvements in realism in Studies II and III were small and further analysis showed that there was room for more improvements after the Adjustment task, Study IV was designed to investigate whether these improvements could be enhanced.

This was investigated partly by giving participants advice, such as to pay attention to whether they believed an item was correct or incorrect and to increase or decrease their confidence accordingly (correct/incorrect

condition), or by telling them to focus on the phenomenological memory quality, in which a relationship between lower levels of realism and the quality "Know" had been observed. Interestingly, neither of these conditions showed improved realism more than the control condition. Thus, participants' tendency to focus more on incorrect than correct items during the regulation of episodic memory reports was not balanced out by the advice. A further analysis confirmed this since there was only a significant decrease in confidence for incorrect items during the Adjustment task also in this condition. Neither did the Remember/Know condition become more realistic after the Adjustment task, which is surprising since the results of Study II showed that the group who made Remember/Know ratings were more realistic in the Confidence task than the control group. However, both of these judgments made in Study IV regarding the answers' correctness vs. incorrectness, as well as their phenomenological memory quality, were delayed judgments made in the Adjustment task on answers given during the Confidence task. This difference might help explain why the Remember/Know condition was so unsuccessful in Study IV compared to the expectations derived from Study II.

The final attempt at trying to help the participants improve their realism by "trying more" was also unsuccessful. It even had the effect that the participants significantly lowered their realism, as measured with the absolute bias measure. There may be a number of reasons for this result. First, it is possible that it was the result of a floor effect, since the participants managed to increase their realism quite well during the Adjustment task, leaving very little room for improvement in the Extra Adjustment task. Secondly, the fact that the participants were asked to try more in the Extra Adjustment task might have made them feel pressured to make more adjustments even though they might have felt satisfied with their realism after the Adjustment task. Similar to these results, Lichtenstein and Fischhoff (1980) found that, when they tried to train participants to be more realistic in their confidence judgments, all the improvement in realism came between the first and second of eleven training sessions. Thus, it could be that there is a ceiling for how good participants can get when it comes to improving their realism. However, Lichtenstein and Fischhoff's study differs from Study IV in many aspects. Most importantly, they had a training session with their participants after each session with personalized feedback. In the current study, all participants were given an elaborate, but general, introduction to the concept *realism of confidence*.

It is therefore questionable whether “try more” instructions are relevant in helping people further improve their realism. In any case, this is likely to depend on how successful they are in their first regulation attempt. Thus, it may be difficult for people not only to identify a lack of realism in their confidence judgments, but also to identify when they have reached ideal realism in their confidence judgments.

Concluding remarks

In this thesis, the making of successful second-order confidence judgments of episodic and semantic memory was investigated. Studies I, II, III, and IV support the notion that people have a second-order metacognitive ability when it comes to regulating the realism of confidence judgments. The making of successful second-order confidence judgments differs from previous attempts of debiasing in very important ways. First, previous approaches to debiasing have focused their attempts on explaining why people are overconfident in their first-order judgments, and the improvement attempts have therefore concerned helping people make realistic first-order confidence judgments for different types of performances (e.g. Erev et al., 1994; Gigerenzer et al., 1991; Juslin, 1993, 1994 ; Griffin & Tversky, 1992, 2002). In this thesis, however, the intention was to investigate whether people have the ability to go back and adjust first-order judgments retrospectively by making second-order judgments.

Since previous research suggests that there are multiple reasons for overconfidence, for example that it depends on error (e.g. Erev et al., 1994), the representativeness of the sample (e.g. Gigerenzer et al., 1991) and individual differences (e.g. Campbell et al., 2004), among other things, it may be that the success of second-order judgments depends partly on the most prominent reasons for overconfidence in the particular context studied. For example, it might be easier for people to realize that the question they have answered is tricky and thereby adjust their confidence level accordingly during the Adjustment task than it is for people to consider how random error might affect their confidence judgments.

Limitations

One limitation of the present research is the high level of realism the participants showed already after the Confidence task (Studies I, II, III and IV), which may have left little room for the effects of attempts to improve the realism to be fully identifiable. Thus, the room for improvement in realism

was limited for the participants. It was only in the recall condition in Study I Experiment I that the levels of bias and calibration were unrealistic enough to be at a satisfactory level for making second-order judgments. There could be many reasons why the level of realism was so good in these studies. One not unlikely reason is that the introduction to the concept *realism of confidence* given to all participants before the Confidence task in the four studies might have changed the way they used the confidence scale to rate their answers. Although several studies have shown that warning participants does not make them less overconfident (Fischhoff, 1982; Gigerenzer et al., 1991; Hedborg, 1996), the introduction to the concept could still have affected how the participants here went about making confidence judgments. To further investigate people's ability to improve the realism in their confidence judgments, it could therefore be of value to use tasks and items for which people give less realistic confidence judgments to their answers. Such tasks may not be representative tasks, if considered in relation to the space of all possible tasks and their natural frequencies in the world; although this may be difficult to judge. Moreover, not all challenges and tasks people face in the world are representative, and sometimes they are challenged to deal with tricky tasks (Smith, Shantau, & Johnson, 2004).

In Study I different scales were used for the two experiments. This limits the interpretations of the findings in Study I. These scales also differed from those used in Studies II, III and IV. Therefore, the possibility of comparing Study I with the other studies is somewhat limited, and differences in results should be interpreted with caution.

Also, the term *recognition* in Study I is used in a more indirect and metaphorical sense, as the memory of the film clip was visually encoded and the response alternatives were described verbally. Thus, the participants had not seen the verbal descriptions in the response alternatives before answering the questions, and these descriptions can therefore not be recognized in a strict sense.

Future directions

Besides investigating tasks with a higher degree of bias, as noted above it would also be interesting to investigate tasks that have been shown to produce underconfidence effects. It would be especially interesting to observe whether people are able to regulate the realism of confidence for tasks displaying underconfidence effects using the Adjustment task method.

Moreover, future research may fruitfully further explore the difference in realism regulation between episodic and semantic memory – for example,

differences with respect to the use of different types of cues for making confidence judgments and when regulating them. It is possible that the making of confidence judgments in episodic memory context is governed more by so-called experience-based cues such as processing fluency, while information-based cues, such as people's beliefs about their recall ability, are more important in semantic memory contexts (Koriat et al., 2008). Although individual differences such as personality and cognitive styles, in similarity to other research, only showed a weak link to the making of first- and second-order judgments, it would be interesting to see whether personality and cognitive style have any predictive value concerning episodic memory tasks.

Further studies could also investigate whether children of different age groups have the same ability to regulate the realism of confidence as adults do, and when this kind of meta-metacognitive ability develops during one's lifespan.

Finally, it would also be of interest to examine whether the same methodology, especially the Adjustment task, could be used to investigate other types of meta-metacognitive judgments with respect to the extent to which they can be successfully regulated by people. Two examples, possibly with some promise for educational contexts, concern the regulation of feelings of knowing and judgments of learning.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*, 30-60. doi:10.1037/0033-2909.131.1.30

Allwood, C. M. (2010). Eyewitness confidence. In P. A. Granhag (Ed.), *Forensic psychology in context: Nordic and international approaches* (pp. 281-303). Devon, United Kingdom: Willan Publishing.

Allwood, C. M., & Granhag, P. A. (1996). The effects of arguments on realism in confidence judgements. *Acta Psychologica, 91*(2), 99-119. doi:10.1016/0001-6918(94)00057-3

Allwood, C. M., & Granhag, P. A. (1999). Feelings of confidence and the realism of confidence judgments in everyday life. In P. Juslin & H. Montgomery (Eds.), *Judgment and decision making: Neo-brunswikian and process-tracing approaches* (pp. 123-146). Mahwah, NJ: Lawrence Erlbaum Associates.

Allwood, C.M., Granhag, P.A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgments: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17*, 545-561. doi:10.1002/acp.888

Allwood, C. M., Innes-Ker, Å H., Holmgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses to free recall and focused questions. *Psychology, Crime & Law, 14*(6), 529-547. doi:10.1080/10683160801961231

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219-235. doi:10.1177/1088868309341564

Arkes, H. R. (1991). Costs and benefits of judgment errors - Implications for debiasing. *Psychological Bulletin, 110*(3), 486-498.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. Organizational Behavior and Human Decision Processes, 39(1), 133-144. Retrieved from http://search.proquest.com/docview/617255716?accountid=11162

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception and Psychophysics, 61*(7), 1369-1383. doi:10.3758/BF03206187

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting, 8*(4), 559-573.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception and Psychophysics, 54*(1), 75-81. doi:10.3758/BF03206939

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1-3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*(5), 938-956. doi:10.1037/0022-3514.92.5.938

Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First- and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and Learning, 8*(1), 79-102. doi: 10.1007/s11409-013-9096-5

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making, 17*(4), 297-311. doi:10.1002/bdm.475

Cesarini, D., Sandewall, Ö., & Johannesson, M. (2006). Confidence interval estimation task and the economics of overconfidence. *Journal of Economic Behavior & Organization, 61*, 453-470. doi: 10.1016/j.jebo.2004.10.010

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131. doi: 10.1037/0022-3514.42.1.116

Clance, P. R. (1985). *The Imposter Phenomenon: Overcoming the fear that haunts your success.* Atlanta: Peachtree.

Costa, P. T., & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO personality inventory. *Journal of Personality and Social Psychology, 54*(5), 853-863. doi:10.1037/0022-3514.54.5.853

Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology, 80*(2), 131-142. doi: 10.1037/0022-0663.80.2.131

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12*(1), 41-55. doi:10.1007/bf01064273

Dahl, M., Allwood, C. M., Rennemark, M., & Hagberg, B. (2010). The relation between personality and the realism in confidence judgements

in older adults. *European Journal of Ageing, 7*(4), 283-291. doi:10.1007/s10433-010-0164-2

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *Journal of General Psychology, 132*(4), 335-346. doi:10.3200/GENP.132.4.335-346

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous overconfidence and underconfidence - the role of error in judgment processes. *Psychological Review, 101*(3), 519-527. doi:10.1037/0033-295X.101.3.519

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422-444). NY: Cambridge University Press.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1*(2), 155-172. doi:10.1002/for.3980010203.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906-911. doi: 10.1037/0003-066X.34.10.906

Flavell, J. (1976). Metacognitive aspects of problem-solving. In L. Resnick (Ed.), *The Nature of Intelligence*. Hillsdale, NJ: Erlbaum Associates.

Flavell, J. H., Miller, S. A., & Miller, P. H. (1993). *Cognitive development* (3 ed.). Englewood Cliffs, N.J.: Prentice-Hall.

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

Goldberg, L.R. (1992). The development of markers for the Big Five structure. *Psychological Assessment, 4,* 26–42. doi: 10.1037/1040-3590.4.1.26

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making.* (pp. 177-198). Malden: Blackwell Publishing.

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

Griffin, D., & Tversky, A. (2002). The weighing of evidence and the determinants of confidence. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristic and Biases: The psychology of intuitive judgment* (pp. 230-249). Cambridge: Cambridge University press.

Hedborg, A. (1996). *Studies of framing, judgment and choice* (Doctoral dissertation). The Economic Research Institute. Stockholm School of Economics. Stockholm.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231-237. doi:10.1111/j.1467-9280.2009.02271.x

Higham, P.A. (2013) Regulating accuracy on university tests with the plurality option. *Learning and Instruction*. 24, 26-36.

Johnson, T. R., Budescu, D. V., & Wallsten, T. S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making, 14*(2), 123-140. doi: 10.1002/bdm.369

Jonsson, A.-C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences, 34*(4), 559-574. doi:10.1016/s0191-8869(02)00028-4

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence of one's general knowledge. *European Journal of Cognitive Psychology, 5*(1), 55-71. doi:10.1080/09541449308406514

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*(2), 226-246. doi:10.1006/obhd.1994.1013

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104*(2), 344-366. doi:10.1037/0033-295x.104.2.344

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making. Special Issue: Stochastic and cognitive models of confidence, 10*(3), 189-209. doi:10.1002/(SICI)1099-0771(199709)10:3<189::AID-BDM258>3.0.CO;2-4

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology - Learning Memory and Cognition, 22*(5), 1304-1316. doi:10.1037/0278-7393.22.5.1304

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1038-1052. doi:10.1037/0278-7393.25.4.1038

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107*(2), 384-396. doi:10.1037/0033-295x.107.2.384

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1-24. doi:10.1006/jmla.1993.1001

Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making, 10*(3), 269-278. doi:10.1002/(SICI)1099-0771(199709)10:3<269::AID-BDM281>3.0.CO;2-L

Kleitman, S. (2008). *Metacognition in the rationality debate. Self-confidence and its calibration*. Germany: VDM Verlag Dr. Mueller.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*(3), 321-341. doi:10.1002/acp.705

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences Journal, 17*(2), 161–173.

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

Koriat, A. (2002). Metacognition research: an interim report. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition*. Cambridge: Cambridge University Press.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490–517. doi:10.1037/0033-295X.103.3.490

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*(4), 405. doi:10.1006/jecp.2000.2612

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107-118. doi:10.1037/0278-7393.6.2.107

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky, & R. A. Bjork (Eds.) *A handbook of memory and metamemory* (pp 117-136). Mahwah, NJ: Lawrence Erlbaum Associates.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9*(5), 178-181. doi: 10.1111/1467-8721.00088

Lai, E. R. (2011). Metacognition: A literature review. Pearson Research Report Retrieved from: http://www.pearsonassessments.com/hai/images/tmrs/Metacognition_Literature_Review_Final.pdf

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know. *Organizational Behavior & Human Performance, 20*(2), 159-183. doi:10.1016/0030-5073(77)90001-0

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior & Human Performance, 26*(2), 149-171. doi:10.1016/0030-5073(80)90052-5

Lichtenstein, S., Fischhoff, B., & Phillips, L., D. (1982). Calibration of probabilities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: Cambridge University Press.

Lindsay, R. C., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*(1), 79-89. doi:10.1037/0021-9010.66.1.79

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodmann-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509-527. doi:10.1037/a0014876

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probability: Theories and models 1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability*. Oxford, England: John Wiley & Sons.

Merkle, E. C. (2009). The disutility of the hard-easy effect in choice confidence. *Psychonomic Bulletin & Review, 16*(1), 204-213. doi:10.3758/pbr.16.1.204

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 502-506. doi:10.1037/a0021802

Mirels, H. L., Greblo, P., & Dean, J. B. (2002). Judgmental self-doubt: Beliefs about one's judgmental prowess. *Personality and Individual Differences, 33*(5), 741-758. doi:10.1016/S0191-8869(01)00189-1

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502-517. doi:10.1037/0033-295X.115.2.502

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*, 595-600.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109-133. doi:10.1037/0033-2909.95.1.109

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10*(3), 257-260. doi:10.1002/(SICI)1099-0720(199606)10:3<257::AID-ACP400>3.0.CO;2-9

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect.". *Psychological Science, 2*(4), 267-270. doi:10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 1–45). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working Memory and Intelligence--Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 61-65. doi:10.1037/0033-2909.131.1.61

Oleson, K. C., Poehlmann, K. M., Yost, J. H., Lynch, M. E., & Arkin, R. M. (2000). Subjective overachievement: Individual differences in self-doubt and concern with performance. *Journal of Personality, 68*(3), 491-524. doi:10.1111/1467-6494.00104

Pallier, G., Wilkinson, R., Danthir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology, 129*(3), 257-299. doi:10.1080/00221300209602099

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 18*(2), 157-168. doi: 10.1002/acp.952

Perfect, T. J., Watson, E. L., & Wagstaff, G. F. (1993). Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology, 78*(1), 144-147. doi:10.1037/0021-9010.78.1.144

Raskin, R., & Terry, H. (1988). A principal components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology, 54,* 890–902. doi: 10.1037/0022-3514.54.5.890

Riding, R., & Cheema, I. (1991). Cognitive styles: An overview and integration. *Educational Psychology, 11*(3-4), 193-215. doi:10.1080/0144341910110301

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*(3), 416-425. doi:10.1037/0021-9010.82.3.416

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology-Applied, 6*(3), 207-221. doi:10.1037//1076-898x.6.3.207

Sadowski, C. J., & Cogburn, H.E. (1997). Need for cognition in the Big-Five factor structure. *Journal of Psychology: Interdisciplinary and Applied, 131*(3), 307-312.

Schaefer, P. S., Williams, C. C., Goodie, A. S., & Campbell, W. K. (2004). Overconfidence and the Big Five. *Journal of Research in Personality, 38*(5), 473-480. doi:10.1016/j.jrp.2003.09.010

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415-429). New York: Routledge.

Schraw, G., Crippen, K., & Hartley, K. (2006). Promoting self-regulation in science education: metacognition as part of a broader perspective on learning. *Research in Science Education, 36,* 111-139. doi: 10.1007/s11165-005-3917-8

Seemungal, F. V., & Stevenage, S. V. (2002). Using state of awareness judgements to improve eyewitness confidence-accuracy judgements. In P. Chambres, M. Izaute & P.-J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 219-231). Dordrecht, Netherlands: Kluwer Academic Publishers.

Sieck, W. R., Merkle, E. C., & Van Zandt, T. (2007). Option fixation: A cognitive contributor to overconfidence. *Organizational Behavior and Human Decision Processes, 103*(1), 68-83. doi:10.1016/j.obhdp.2006.11.001

Smith, K., Shanteau, J., Johnson, P. (2004). Introduction: What does it mean to be competent? In K. Smith, J. Shanteau & P. Johnson (Eds.), *Psychological investigations of competence in decision making* (pp. 1-4). Cambridge: Cambridge University Press.

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2008). Measuring memory and metamemory: Theoretical and statistical problems with assessing learning (in general) and using gamma (in particular) to do so. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 95-114). New York: Psychology Press.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315-327. doi:10.1037/0033-2909.118.3.315

Stankov, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tasks. *Learning and Individual Differences, 10*(1), 29-50. doi:10.1016/s1041-6080(99)80141-1

Stankov, L., Lee, J., & Paek, I. (2009). Realism of confidence judgments. *European Journal of Psychological Assessment, 25*(2), 123-130. doi: http://dx.doi.org/10.1027/1015-5759.25.2.123

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83*(2), 282-309. doi:10.1006/obhd.2000.2910

Tanaka, J. S., Panter, A. T., & Winborne, W. C. (1988). Dimensions of the need for cognition: Subscales and gender differences. *Multivariate Behavioral Research, 23*(1), 35-50. doi:10.1207/s15327906mbr2301_2

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26*(1), 1-12.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502-517. doi: http://dx.doi.org/10.1037/0033-295X.115.2.502

Wallsten, T.S.,& Diederich, A (2001). Understanding Pooled Subjective Probability Estimates. *Mathematical Social Science, 41,* 1-18. doi: 10.1016/S0165-4896(00)00053-6

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *Journal of Educational Research, 84*(1), 30-43.

Want, J., & Kleitman, S. (2006). Imposter phenomenon and self-handicapping: Links with parenting styles and self-confidence. *Personality and Individual Differences, 40*(5), 961–971. doi:10.1016/j.paid.2005.10.005.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6), 1049-1062. doi:10.1037/0022-3514.67.6.1049

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66*(6), 688-696. doi:10.1037/0021-9010.66.6.688

Wells, G.L., Small, M., Penrod, S., Malpass, R.S., Fulero, S.M., & Brimacombe, C.A.E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22,* 603–647.

Wolfe, R. N., & Grosch, J W. (1990). Personality correlates of confidence in one's decisions. *Journal of Personality, 58,* 515-534. doi:10.1111/j.1467-6494.1990.tb00241.x

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior & Human Performance, 30*(1), 132-156. doi:10.1016/0030-5073(82)90237-9

Yates, J. F. (1988). Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition. *Organizational Behavior and Human Decision Processes, 41*(3), 281-299. doi:10.1016/0749-5978(88)90031-3

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. A. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 381-410). Oxford: John Wiley and Sons.

Appendix

I. Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive Processing, 13*(3), 243-253. doi: 10.1007 /s10339-012-0440-5

II. Buratti, S., & Allwood, C. M. (2012). Improved realism of confidence for an episodic memory event. *Judgment and Decision Making*, 7(5), 590-601. Retrieved from http://journal.sjdm.org /11/111121a/jdm111121a.html

III. Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First- and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and Learning, 8*(1), 79-102. doi: 10.1007/s11409-013-9096-5

IV. Buratti, S., & Allwood, C. M. (2013). *The effects of advice and "try more" instructions on improving realism of confidence*. Manuscript submitted for publication.