Review of

From memory access to memory reports:

The role of conversion processes in shaping accuracy

PhD Thesis by Ewa Skopicz-Radkiewicz

Review completed by Prof Phil Higham

Introduction

The introductory chapter provides an overview of the three papers that constitute the main content of the thesis. The first paper is a review of several related literatures: the strategic regulation of memory accuracy, metamemory topics in autobiographical memory, and the source-monitoring framework. The second is an empirical study investigating metamemory processes with unanswerable questions. The third paper investigates conversion processes in the eyewitness misinformation paradigm. This introductory section also includes discussion of the content of each paper in various forms. For example, sections 1.1, 1.2, and 1.3 relate to the core ideas, background, and rationale for conducting the review/research in each paper. For example, the first paper, which is a literature review, focuses on manipulations and interventions aimed at improving resolution (the ability to discriminate between accurate and inaccurate responses). Section 2 provides overviews of each of the three papers. Section 3 discusses the theoretical and applied implications of each paper. Finally, there is a section on limitations and future directions.

Generally, I thought this introductory chapter was well-written and provided readers with the context necessary to integrate the papers into a coherent whole. Without it, there would be a danger that readers would not see the relationship between research on unanswerable questions (Paper 2) and the misinformation effect (Paper 3). That said, I thought this chapter could be organised better. The present structure is to provide core ideas, background, and rationale for conducting the review/research in each paper, then to provide overviews for each paper, and then to discuss theoretical and applied implications of each paper. This structure means that readers are bouncing around from one paper to the next and back to the first paper again. It would be preferable to include all discussion of Paper 1, then move to Paper 2 and include all discussion, then move to a section that includes all discussion of Paper 3.

Specific Comments

p. 16, first paragraph: “fromat” should be “format”. Also, I’m not convinced by the argument that recall is more ecologically valid than other formats. For example, witnesses required to identify a suspect from a line-up is using recognition, not recall, and that certainly ecologically valid.

p. 17, “…distinct processes which not always overlap”. Insert “do” between “which” and “not”.

p. 20: “…the misinformation effect can increase in double”. Awkward sentence structure. Reword.

p. 24, bottom: Run-on sentence. Break down into two or more sentences.

p. 26: "…both the processing misinformation…" Insert "of" between "processing" and "misinformation". Also, participants' initial assumptions about post-event misinformation being accurate are discussed. Has anyone ever asked participants what they assume? Psychology students, who are the typical participants in these studies, know that experimenters use deception. Perhaps they assume that experimenters are trying to trick them?

p. 26: "In sum, both witnesses and participants might have good reasons to find it difficult to conclude and communicate that a question is unanswerable.": I think this depends on the question. If a witness is asked about what the robber looked like when they are fully aware that they were nowhere near the scene of the robbery when it took place, then they would have no problem saying "I don't know".

p. 27: "…present project, in which it is not different from…". Try instead "…present project, which is not different from…".

Paper 1 – Literature Review

The first paper of the thesis is not an empirical paper, but a literature review that integrates different subfields of metacognitive research. Specifically, Koriat and Goldsmith's (1996) framework for the strategic regulation of memory accuracy is discussed alongside autobiographical memory research on believed memories and Marcia Johnson and colleagues' source-monitoring framework. The central focus here is on finding manipulations or interventions that improve metacognitive resolution. Such manipulations have been reported in the context of autobiographical memory, (e.g., Scoboria et al.'s, 2014, *brief retrieval training* and Niedzwienska's, 2004, classroom intervention) but less so in within the accuracy regulation framework. Indeed, many of the manipulations employed in the accuracy framework which affect input-bound quantity and output bound accuracy do so by affecting the report criterion, not resolution per se. For example, increasing the incentives for accuracy can improve output-bound accuracy, but they do so by reducing the quantity of information offered. If, instead, increases in accuracy came about because of improvements to resolution, then there need not be a cost in term of reduced quantity. Indeed, with better resolution, it is possible to improve both accuracy and quantity in tandem.

The review is courageous in that the authors are attempting to integrate two (or even three) literatures that normally do not cite each other much. I also agree that manipulations that affect criterion setting (control) are much easier to find than ones that affect resolution (monitoring). However, manipulations that affect resolution do exist. For example, comparing standard trivia questions with deceptive trivia questions will yield differences in resolution. (Deceptive trivia questions are questions that yield high-confident, but wrong answers; e.g., many people believe that the capital of Australia is Sydney when it is actually Canberra). Similarly, adding items to a memory test that yield confident correct rejections (e.g., including the participant's name as a foil on a recognition test) will improve resolution. Higham (2002; M&C; see also Higham & Tam, 2005, JML) also showed that context reinstatement in the classic encoding specificity cued-recall paradigm greatly affected resolution. However, I agree with the authors that training interventions that successfully improve resolution via enhanced retrieval and/or monitoring strategies are few and far between.

Although the review is ambitious with integrating somewhat separate literatures, I was disappointed that the only approach to accuracy regulation that was discussed was Koriat and Goldsmith's (1996) framework. However, this is only one approach, and some of the assumptions that are made in that framework are questionable. For example, there seems to be an assumption that people have a malleable report criterion that is sensitive to context and incentives, but that confidence ratings directly reflect underlying states of subjective confidence and are fixed. This assumption is seen in the way that the report criterion is computed in Koriat and Goldsmith's framework. Specifically, fit ratios are used to determine the best fit for the report criterion (denoted P_{rc}) with respect to confidence. However, if confidence ratings are also malleable and subject to context and incentives, then this computational procedure will not work (e.g., if the report criterion shifts to a more stringent position and the confidence criteria shift along with it, then P_{rc} will not change). This assumption seems to underpin the discussion on p. 200 (manuscript page number) of Paper 1 where research by Rechdan et al. (2018) is reported. Here, the authors note that a social feedback manipulation had an effect on control, as measured by grain-size volunteering, but not on resolution, as measured by confidence ratings. This is consistent with the Koriat and Goldsmith (1996) framework where resolution is measured with respect to confidence (usually with the Goodman-Kruskal gamma co-efficient) because of the assumption that confidence ratings are fixed and therefore provide a stable gold standard which is unaffected by control mechanisms.

However, signal detection theory provides another approach to understanding accuracy resolution that does not rely on these (strong) assumptions. In this framework, confidence ratings are subject to control mechanisms just as the report criterion or grain-size criterion is. Specifically, participants are assumed to assign confidence ratings by adopting multiple criteria. For example, to assign 50 on a 100-point scale, people must have enough subjective evidence to assign 50, but not so much that they would assign 60. Thus, in this framework, both the report criterion and the confidence criteria are subject to the influence of context and incentives (i.e., control mechanisms). In various areas of psychology, this assumption has been shown to have some credence using ROC analysis. For example, in the context of research on misinformation in social media, gamified inoculation interventions cause the points on the ROC to shift to more conservative positions. In other words, people become more sceptical of all news after playing these games, and this is shown by them adopting more conservative positions for their confidence criteria. Similarly, research in recognition memory has shown that people's confidence criteria can shift around depending on context and incentives. Also, Higham (2007; JEP:G) showed that a similar type of shifting can occur in metacognitive experiments when incentives are manipulated. Thus, I think the review would be stronger if this line of research was discussed.

Specific Comments

p. 194, column 2: "Therefore, if resolution was perfect, no quantity-accuracy trade-off would be observable." This is not quite right. It depends on the placement of the report criterion. If it is set too conservatively, some accurate statements will be withheld even though people can perfectly discriminate between their own correct and incorrect candidate answers. A case like this might occur in court if a lawyer won't permit a witness to report correct information she possesses (e.g., "just answer yes or no please").

Paper 2 – "It Was Not Mentioned": Improving Responses to Unanswerable Questions Using Retrieval Instructions

The authors report two experiments investigating how people respond to unanswerable questions (e.g., a question about the appearance of an item in a witnessed event that was not shown). In Experiment 1, participants viewed a video of a burglary and then were asked questions about it, some of which were answerable and some of which were not. There were three conditions: the BRT (Brief Retrieval Training) condition, where participants were provided with training designed to improve responding to unanswerable questions previously investigated by Scoboria et al. (2013) that included Review, Retrieve, References, Reflect, Reply prompts; (2) Criterion, which encouraged participants to adopt a more conservative report criterion; and (3) Control, which had no training. The results revealed that the BRT group performed the best on unanswerable questions, with the Criterion group responding conservatively. The results suggested that participants' task representations were improved by BRT training; any benefits were not just attributable to conservative responding. In Experiment 2, a 2 X 2 designed was used with the first factor varying the presence of BRT training and the second varying the presence of a "not shown" response option. The latter factor improved responding to unanswerable questions, presumably because of a better task representation, but the BRT training improved performance even more, suggesting the manipulations were not entirely redundant with each other. The results highlight the importance of conversion processes in memory tasks.

In general, I liked this paper. I think the experiments are clever and yielded interesting results. The discussion is insightful drawing on subtle but important distinctions (e.g., the meaning of "don't know" responses in the context of unanswerable questions). Also, the message is an important one – that we should consider the problem-solving elements of memory tasks, not just memory accessibility.

I have a few comments as follows:

p. 2: The authors state "If monitoring resolution was at the level of chance, the strategic control would not increase accuracy, but it could decrease quantity. If we could improve monitoring resolution, an exercise of strategic control would have smaller effects on quantity." Generally, this is true, but it depends on the start and end point of the report criterion as control is exercised. Consider the following thought experiment. A participant is answering a 100-item test and forced-report quantity is 50% (i.e., 50% of the answers are correct if all questions are answered). Suppose now that resolution is at chance and that the report criterion is shifted from a relatively moderate setting, such that 50% of the correct (and incorrect) answers are reported, to a highly conservative one such that no answers are reported. Quantity would decrease from 25% (50% of 50%) to 0%. Now suppose resolution is excellent and that, again, the report criterion is shifted from a relatively moderate setting, such that 50% of the correct (and no incorrect) answers are reported, to a highly conservative one such that no answers are reported. Quantity would again decrease from 25% to 0%, despite the massive difference in resolution.

p. 4: The authors report using a Cohen's *d* = .83 for the power analysis in Experiment 1. This value represents a very large effect size. To put it into context, Cohen recommended treating *d* = .80 as a "large" effect size, so the value chosen for the power analysis exceeds the "large" threshold. I understand that the value was taken from Scoboria et al. (2013), but it is not clear which effect was chosen or why. The worry here is that Experiment 1 will be underpowered. I also cannot replicate the results of the power analysis. Using G*Power, I computed the sample for a one-way, between-subjects ANOVA with three group, power = .95, alpha = .05, and effect size (Cohen's *f*) = .415 (Cohen's *f* is half Cohen's *d* in a two-group design). The result was 93 participants. This is close to the number of participants the power analysis for Experiment 1 suggested, but not the same. More to the point, if a default *medium* effect size is used instead in the same analysis (Cohens *f* = .25), then 252 participants are needed, an almost threefold increase in sample size. One of the problems with choosing very large effect sizes for a priori power analyses is regression to the mean. Extreme scores are likely to become less extreme when remeasured. Hence, it is safer to choose more moderate values to avoid being under powered.

p. 6, Experiment 1: "If, despite the instruction, participants persisted in responding DK, the answer was coded as an error and the confidence rating was coded as a 0." Did you test whether the results were any different if those trials were excluded rather than having them "filled in" by the experimenter?

p. 6: The analysis of metacognitive resolution in Experiment 1 was unusual. I've not seen that anywhere in the literature before. Usually, researchers compute a within-subjects correlation for each participant between accuracy and confidence. While Goodman-Kruskal's gamma with concordant/discordant observations is commonly used for this purpose, there are better ways to estimate gamma such as using ROC curves (see Higham & Higham, 2019; https://doi.org/10.3758/s13428-018-1125-5). There are no citations in this section, so I'm assuming that your analysis is unique. What guarantees do you have that this measure is performing properly and providing the information about resolution that you want? Did you compare the results of your analysis to more conventional statistics or examine the relationship between your new measure and other measures? What was the rationale for using a unique measure?

p. 8: Related to the previous point, I noticed that the criterion group produced a stricter criterion than the control group in that less incorrect information was provided. However, there was no reduction in the amount of correct information provided. In other words, there was no confidence-accuracy trade-off. This suggests that resolution was very high. Later on the same page it is noted that the BRM procedure did not improve resolution. Could that be because resolution was at ceiling, as suggested by the data in the criterion group?

pp. 9-10: There is some discussion about the normative suitability of rejecting unanswerable questions versus answering them incorrectly. That is, it may break social norms to reject answers as unanswerable too often just as saying DK too often does. I think this depends somewhat on how clearly unanswerable the questions are. For example, if I'm interviewed by the police about a crime that took place between 9:30pm and 10:00pm and know for certain that I was at a movie that didn't end until 10pm, then it is informative to tell the officer that I can't answer the question because I wasn't there. That is more informative (and accurate) that making something up. Alternatively, if the situation is a bit more ambiguous, such as whether I could see a certain object in a room that I know I was present in, then I might offer an answer as I cannot be sure that the question is unanswerable. This analysis suggests that people's willingness to reject a question might depend on how well they can discriminate between answerable and unanswerable questions. Have you or anyone else ever looked at this?

The main conclusion from this Experiment 2 is that a "reject" option is enough to improve performance to unanswerable questions, but that adding BRT as well boosts performance even further. The results are interpreted to mean that "BRT improved performance above and beyond the improvement conferred by the presence of the reject option, which means that not all of BRT efficacy can be attributed to the awareness of unanswerable questions." However, another interpretation is that the "reject" option did not achieve full awareness of unanswerable questions and that adding BRT boosted that awareness further. Participants may have habituated to the options after answering a few and their awareness may have dwindled. However, if BRT was added, then any effects of habituation might have been mitigated. This interpretation is quite different from the one provided in Paper 2. It may still be that full awareness is the only mechanisms at work here, but that neither BRT nor the reject option on its own is enough to ensure full awareness. (And, indeed, adding a third factor may boost awareness even further.)

Experiment 2, Results: Resolution was not measured in Experiment 2, presumably because there was no second round through the questions where participants provided confidence ratings. However, it is possible to measure resolution with respect to the report criterion. Using signal detection theory, for example, one can compute a metacognitive hit rate (i.e., the proportion of correct answers that are reported) and a false alarm rate (i.e., the proportion of incorrect answers that are reported), and discrimination (resolution) can be computed from those rates (e.g., see Higham, 2002; Higham & Tam, 2005). This marks another advantage of the signal detection approach to accuracy regulation compared to Koriat and Goldsmith's (1996) framework. That is, your hypothesis about BRT potentially improving resolution could be tested in Experiment 2 as well as Experiment 1 by adopting the signal detection framework and using the report criterion instead of confidence ratings to compute resolution.

Paper 3 - Influence of experimentally manipulated misinformation availability on discrepancy detection and double misinformation processing

The authors ran a single experiment on eyewitness suggestibility. Two variables were manipulated: misinformation availability (high vs. low) and single vs. double misinformation. Participants first watched a video, and after completing a distracter task, they read two narratives. In the single condition, a given video detail was subject to misinformation once (in one of the narratives), whereas in the double condition, two different misinformation details were mentioned in the narratives, one in each. Additionally, some video details were not subject to misinformation at all. High versus low misinformation availability was manipulated by the manner in which the narratives were presented. In the high condition, participants could devote full attention to encoding the narrative, both narratives were visible to facilitate comparisons, and they were encouraged to highlight any discrepancies. In the low condition, participants' attention was divided, and they had limited time to encode the narratives. The authors expected that double misinformation would produce a larger misinformation effect (and lower accuracy) in the low availability condition, but the opposite in the high-availability condition (because of discrepancy detection leading participants to not trust the narratives). However, the predictions were not realised in the data.

Overall, I think this is interesting research. Discrepancy detection and its role in memory and misinformation effects has been investigated before. However, this research is unique in that its focus is on discrepancy between two pieces of misinformation pertaining to the same detail. That said, I think this paper is the weakest of the three. Perhaps that is because, unlike Papers 1 and 2, it has not yet been subject to revisions based on journal reviews. In my view, there are multiple problems ranging from unclear methods to potential confounding variables. I detail these problems below.

1. The first problem as I see it is that the manipulation of misinformation availability doesn't just affect the availability of the misinformation, it also affects the availability of true information about the video. Most of the details in the narratives are accurate. Thus, providing participants with the opportunity to study the narratives uninterrupted means that they are learning about accurate video details in greater depth. This is bound to affect accuracy should the memory test query anything to do with these accurate narrative details (or details related to them). A second related problem is that deep learning of the misinformation is also likely to encourage discrepancy detections between the video details and the misinformation details in the narratives. The authors claim that they are interested primarily in the detection of discrepancies between the two misinformation details in the double condition. However, the impact of video/narrative discrepancy detection can't just be ignored altogether, particularly if the manipulations are likely to affect it.

2. The Method section is missing some important details. For example, readers are told that all participants read two narratives following viewing of the video. However, there are three types of items: double misinformation items, single misinformation items, and control items. Exactly how did these details appear in the narratives? Were control items present in both narratives in the double conditions? In the single condition, were all the control items in the narrative without misinformation? Or just what? Also, what was done to counterbalance the items within the narratives to eliminate item selection effects? Presumably, critical items served equally often in the control, single misinformation, and double misinformation conditions? How was this achieved? In the single misinformation condition, did the misleading detail occur in the first narrative, the second narrative, or some combination of the two? As far as I can tell, none of this important information is included in the method section.

3. In addition to the missing information from the method section, the literature review is incomplete. Research on discrepancy detection has been conducted for decades, I believe beginning with Tousignant et al. (1986; http://dx.doi.org/10.3758/BF03202511) . More recently, Higham et al. (2017; http://dx.doi.org/10.1037/xap0000140), using a procedure for determining which discrepancies were detected that was very similar to the one used here, found that discrepancy detection had a profound effect on performance. On the other hand, Neil et al. (2021; https://doi.org/10.1037/xge0001023) investigated discrepancy detection in the concurrent misinformation paradigm (where misleading subtitles are presented at the same time as a video with distorted audio), and found that it played a role, but a smaller one.

Hypothesis 2 pertains to the effect of detecting discrepancies in the high availability condition; specifically, it states that performance will improve (less misinformation endorsement and more recall of video details) when double misinformation is presented. This is attributed to participants noticing discrepancies between the narratives, which changes the task representation and leads participants to distrust the sources of the narratives. However, discrepancy detection can bring about improved memory performance for reasons entirely separate from considerations of task representation and source credibility. For example, the *recursive remindings framework* (Hintzman, 2011; https://doi.org/10.1177/1745691611406924) suggests that the process of detecting change causes people to retrieve the original, pre-changed event and that this act of retrieval serves to strengthen memory for that event. Indeed, Jacoby et al. (2015; http://dx.doi.org/10.1037/xlm0000123) showed that change detection in the classical retroactive interference paradigm, where cued-recall performance is compared between an interference condition (A–B, A–D) and a control condition (A–B, C–D), can ironically lead to better memory for the original event (B) in the interference condition. Putnam et al. (2017; http://dx.doi.org/10.1177/09567 97616672268) and Higham et al. (2017, Experiment 2) found similar results in the classical misinformation paradigm, which is a special case of retroactive interference. Thus, improved performance with highly available misinformation (which facilitates discrepancy detection) in the double misinformation condition could occur purely for memory reasons, specifically, because of covert retrieval practice of the original event.

None of the papers mentioned above is cited, despite the clear relevance to the current research. Paper 3 should be edited accordingly.

3. In addition to the missing literature, the Introduction is hard to follow because it is seemingly reliant on readers having read Blank et al. (2022). Additionally, terms are used before they are described (e.g., discrepancy detection test), call outs to tables are out of order (Table 2 is called out before Table 1), page numbers and running head are missing, several headings are not in title case, and Table 2 is missing definitions of acronyms. Also, Figures 2 and 3 should be deleted because they are redundant with Table 4.

4. The two main dependent variables are “misinformation endorsement” and “accuracy”. Accuracy refers to recall of the original event detail. These are treated as though they are independent, but of course they are not. For any given question on the memory test, participants can only provide one answer. Hence, if they endorse the misinformation, they cannot also respond with the event detail. In other words, these variables are in a trade-off relationship; as the probability of one response increases, the probability of the other necessarily decreases. One way around this problem would be to convert the frequencies into probabilities (which is a good idea anyway as it obviates the need for readers to constantly be reminded how many critical items there are) and then compute how likely retrieving the video detail is out of the available opportunities. In other words,

Accuracy = P(video detail)/ (1 – P(misinformation endorsement)).

This kind of correction has been used in the remember/know literature for the same reason (e.g., Yonelinas & Jacoby, 1995). That is, rather than using the raw K probability as an estimate of familiarity in recognition memory experiments (because it necessarily low if there is a lot of recollection), familiarity is computed as,

Familiarity = P(K)/(1 – P(R))

At the very least, the trade-off relationship between these two dependent variables, and the implications for the results, should be discussed.

5. Paper 3 is missing a power analysis in the main text. If that was not done a priori, then a post hoc sensitivity analysis should be conducted to determine how much power the study has. If a power analysis was not conducted, the stopping rule pertaining to the sample size

should be described in the Participants section. Also, the Results section is quite long, and it felt at times that the data were being tortured (i.e., over-analysed). I think Paper 3 would benefit from some streamlining, particularly as the main manipulation – double vs. single misinformation – had little effect.

6. As there were a number of null results, the paper would benefit from reporting some Bayes factors. Was there enough power to make the null results meaningful? Or was the study underpowered such that null results cannot be trusted? Without a power or sensitivity analysis (see point 5 above), it is difficult to tell.

7. Toward the end of the paper, the authors comment on how "…even when fully detecting a discrepancy between the two contradictory details, of which only one can logically be true, participants frequently reported misinformation." However, depending on the state of participants' memory for the original event, this is not such an odd thing to do. For example, if participants had no memory for the original event, and then were presented two pieces of misinformation, they may believe one is correct and the other is wrong. This again highlights that is impossible to understand performance in this paradigm without also considering the status of the original event memory and whether discrepancies were detected between the original event and the content of the narratives. As noted in point 1 above, discrepancy detection of this sort is largely ignored until the end of Paper 3. In my view, it should be considered from the outset because it is impossible to understand participants' behaviour in this paradigm without taking it into account.

Conclusion

The reviewed doctoral dissertation aligns with the criteria outlined in Article 187 of the Act, specifically points 1 and 2. The candidate has demonstrated comprehensive theoretical knowledge in the field of psychology and the ability to independently conduct research. The solution presented within the dissertation itself is original.