UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Psychology

The Underconfidence-With-Practice Effect:

Mechanisms and Boundaries

by

Katarzyna Zawadzka

Thesis for the degree of Doctor of Philosophy

September 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Psychology

Thesis for the degree of Doctor of Philosophy

THE UNDERCONFIDENCE WITH PRACTICE EFFECT: MECHANISMS AND BOUNDARIES

by Katarzyna Zawadzka

The present thesis examined the underconfidence-with-practice (UWP) effect - a common finding in research concerned with judgements of learning (JOLs), which are predictions of future memory performance. In cued-recall tasks consisting of multiple study-test cycles, immediate JOLs underestimate memory performance on cycle 2 and beyond, revealing the UWP effect.

There has been no consensus as to the origins of the UWP effect, with some accounts interpreting UWP as a manifestation of psychological underconfidence, and others as an artefact of using a 0-100% rating scale to elicit JOLs. The present research aimed at solving this conundrum by investigating the interpretation of 0-100% JOLs in the UWP paradigm. Three possible interpretations have been proposed. From those, the *probability* interpretation was consistent with the psychological underconfidence account of the UWP effect, while the *distorted rating* and *ranking* interpretations assumed that UWP is an artefactual pattern.

Across seven experiments, three different methods have been used to distinguish between the three interpretations of 0-100% JOLs. All methods favoured the ranking interpretation, which cannot be accommodated by the accounts proposing that the UWP pattern reflects psychological underconfidence. The ranking interpretation of 0-100% JOLs proposes that, in the multi-cycle task, participants use JOLs to rank order items in terms of their *evidence for* rather than *probability of* future recall. If JOLs are not probability assessments, then the correspondence between the mean of JOLs and recall performance cannot be meaningfully assessed. In this way, the UWP effect, which is inferred from this correspondence, becomes an artefact of misinterpreting the rating scale. A novel, *recalibration* account is proposed to account for the UWP data.

Table of Contents

1. Underconfidence with practice 13

2. Possible mechanisms of the UWP effect 21

2.1. The UWP effect as a manifestation of psychological underconfidence 21

2.1.1. Cue-utilisation approach 22

2.1.2. Mnemonic debiasing account 28

2.1.3. Memory-for-past-test heuristic 31

2.1.4. Stability bias 38

2.1.5. Summary and conclusions 40

2.2. The UWP effect as an artefact 42

2.2.1. Anchoring-and-adjustment hypothesis 42

2.2.2. Reduced variability for the highest JOLs 47

2.2.3. Summary and conclusions 48

3. Present research 51

3.1. Rationale 51

3.2. Experimental overview 56

3.2.1. Paper 1 56

3.2.2. Paper 2 58

3.2.3. Paper 3 60

4. Authorship 63

5. Paper 1 65

Introduction 65

Experiment 1 67

Experiment 2 76

General Discussion 79

6. Paper 2 85

Introduction 85

Experiment 1 90

Experiment 2 96

Experiment 3 102

General Discussion 111

7. Paper 3 117

Introduction 117

Experiment 1 130

Experiment 2 141

General Discussion 147

8. Conclusions 157

8.1. The interpretation of 0-100% JOLs 157

8.2. Recalibration in the UWP paradigm 164

8.3. Summary 169

9. Appendices 171

Appendix A 171

Appendix B 172

10. References 175

List of tables

Table 5.1. JOLs, bets, recall performance and resolution in Experiments 1 and 2 70

Table 5.2. JOLs, bets and recall performance as a function of cycle-3 recall pattern in Experiments 1 and 2 73

Table 6.1. JOLs, recall performance and resolution in Experiments 1-3 93

Table 6.2. JOLs and recall performance as a function of rating in Experiment 2 99

Table 6.3. JOLs and recall performance as a function of rating in Experiment 3 106

Table 6.4. JOLs and recall performance as a function of confidence rating in Experiment 3 110

Table 7.1. JOLs and recall performance as a function of pair type in Experiment 1 133

Table 7.2. Resolution as a function of group in Experiments 1 and 2 134

Table 7.3. Mean c_1 scores as a function of group in Experiments 1 and 2 139

Table 7.4. JOLs and recall performance as a function of pair type in Experiment 2 144

List of figures

Figure 1.1. Immediate and delayed JOLs: A comparison 17

Figure 3.1. A signal-detection model of responding in the JOL task 60

Figure 5.1. JOLs, bets and recall performance as a function of number of successful recall attempts in Experiments 1 and 2 71

Figure 6.1. JOLs and recall performance as a function of rating in Experiment 1 92

Figure 6.2. JOLs and recall performance as a function of rating in Experiment 2 101

Figure 6.3. JOLs and recall performance as a function of context rating in Experiment 3 108

Figure 7.1. A signal-detection representation of the JOL task 122

Figure 7.2. A metacognitive ROC curve 124

Figure 7.3. Predictions of the bias account 128

Figure 7.4. ROCs for the experimental and control groups in Experiments 1 and 2 137

Figure 7.5. A graphical representation of the UWP effect 149

Figure 8.1. A signal-detection representation of anchoring 161

Declaration of Autorship

I, Katarzyna Zawadzka,

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

The underconfidence-with practice effect: Mechanisms and boundaries

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as:

   Paper 1 is currently under revision; Papers 2 and 3 are currently not submitted, but are likely to be in the future.

Signed:

Date:

Acknowledgements

I would like to express deep gratitude to:

1) my supervisor, Phil Higham, for his guidance, support, and patience;
2) Greg Neil, for teaching me how to program in LiveCode;
3) Dylan Jones and Ed Wilding, thanks to whom I was able to spend the final year of my studies at Cardiff University;
4) Maciej Hanczakowski, who came up with the idea of researching the UWP effect when we were looking for something to do during one Easter break in Poland.

1. Underconfidence with practice

Judgements of learning (JOLs) are predictions of future memory performance elicited at or after the time of study, but before a test. In a typical procedure used for investigating JOLs, participants are presented with a list of words or word pairs for study. For each item on the study list, participants are asked to predict their performance at a future test (so-called *item-by-item* JOLs). These predictions can be elicited on a scale, such as 0-100% or 1-to-*n*, or they can be made in a binary format, such as yes/no. After the study phase, a test follows. At test, participants are asked to retrieve the items presented during the study phase. If single words were presented for study, the task is commonly one of free recall (e.g., Mazzoni & Nelson, 1995; Susser, Mulligan, & Besker, 2013). If participants were presented with word pairs, at test they are given a cued-recall task: they are supposed to retrieve the second word from the pair - a *target* - when presented with the first word - a *cue* (e.g., Dougherty, Scheck, Nelson, & Narens, 2005; Koriat, 1997; Pyc & Rawson, 2012).

From the results of a JOL task, two measures of correspondence between JOLs and memory performance can be derived. *Resolution* is the correspondence between these measures on an item-by-item level. It refers to the ability of JOLs to distinguish between subsequently recalled and unrecalled items. The better the judgements are at distinguishing between these items, the higher the resolution. In the case of perfect resolution, all recalled items are assigned higher JOLs than the unrecalled items. *Calibration* refers to the correspondence between recall performance and the mean of JOLs. There are three possible outcomes. If JOLs and recall performance are approximately equal, good calibration (or realism) can be inferred. If the two measures do not match, calibration is said to be impaired. JOLs exceeding recall performance are interpreted as overconfidence, while the reverse pattern is described as underconfidence. Calibration can be assessed for the whole study list by comparing the mean of all JOLs assigned to the studied items to overall

recall performance. Alternatively, calibration can be assessed on different levels of confidence. In this case, for each confidence level (e.g., 0%, 10%, ... 100%) the percentage of correct responses is calculated, and these values are then compared. For example, if only 40% of items assigned a JOL of 50% were correctly recalled, that would be interpreted as underconfidence.

The underconfidence-with-practice (UWP) effect (cf. Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002) is a finding from research on JOLs that describes the impairment of calibration with repeated learning. In a typical experiment in the UWP paradigm, participants first study a list of paired associates and are asked to provide item-by-item JOLs for each of the pairs, assessing the likelihood that the target will be recalled on a subsequent test when the cue is presented. After the study phase, the test phase follows, during which participants are presented with cues taken from these pairs and required to recall the associated target. This study-test cycle is then repeated one to three times. Typical results from studies investigating the UWP effect show that on the first study-test cycle people's mean item-by-item JOL ratings either match or exceed their mean recall performance (revealing good calibration or overconfidence, respectively), whereas from the second cycle onwards their mean JOLs are lower than their actual performance - a pattern of results that is assumed to reflect people's underconfidence in their future performance. This impairment of calibration occurs in spite of an increase in the number of items that can be recalled and an improvement of resolution. Thus, as people learn more and become more adept at distinguishing on an item-by-item basis between the items that they will recall and those that they will not recall at test, their ability to track their recall levels with JOLs at a global, test level decreases.

The UWP effect is usually found in multi-cycle experiments, consisting of two or more study-test phases. It has been shown on each of the study-test cycles that follow cycle 1 in studies that employed two

cycles (e.g. Koriat, 1997, Experiment 1; Scheck & Nelson, 2005; Finn & Metcalfe, 2007, 2008), three cycles (e.g. Koriat, Ma'ayan, Sheffer, and Bjork, 2006; Karpicke, 2009, Experiment 3; Tauber & Rhodes, 2012), or four cycles (e.g. Koriat, 1997, Experiment 2; Koriat & Bjork, 2006). None of these studies showed that the magnitude of the UWP effect changes systematically from cycle to cycle. Some manipulations were also shown to produce the UWP effect when only one study-test cycle was employed: for example, Koriat (1997) demonstrated the UWP pattern in experiments in which the number of presentations of items or presentation times were manipulated within one list. In this case, however, the effect seems to be less robust, as other authors have failed to replicate it (Finn & Metcalfe, 2008).

The UWP effect is resistant to many experimental manipulations. It can be moderated, but not eliminated by offering incentives for correct recall. In the differential incentive condition of the unpublished study by Koriat, Ma'ayan, and Levy-Sadot (as cited in Koriat et al., 2002), some of the word pairs were assigned three bonus points for correct recall, whereas others were assigned one point. In the constant incentive condition of the same study, all words were assigned the same incentive (two points). No penalties were associated with incorrect recall or lack of recall. Incentives did not change between the three study-test cycles. The results showed that although higher incentives in the differential incentive condition led to longer study times, recall was not influenced by this manipulation and recall performance for items assigned low and high incentives was almost identical. Mean JOLs, on the other hand, were higher for items assigned higher incentives. The effects of feedback are similar to those of incentives. In Koriat's (1997) Experiment 2, which examined whether feedback would reduce the discrepancy between mean JOL and recall performance, half of the participants received feedback concerning the correctness of their responses immediately after each answer, whereas the other half did not. The results showed that, although

mean JOLs were lower than recall performance on cycles 2 and beyond in both groups showing the UWP effect, this effect was more pronounced in the no feedback group. Recall performance, on the contrary, was not influenced by the feedback manipulation.

The UWP effect is also resistant to changes in the basic experimental procedure which consists of alternating study and test phases. Finn and Metcalfe (2008) examined whether explicit JOLs made during the study phase are necessary for the UWP effect to emerge. This was motivated by an observation that JOLs elicited on cycle 1 are, on average, lower than those elicited on cycle 2. Memory for those low cycle-1 JOLs could later lead to a downward bias on cycle 2, reducing the magnitude of JOLs elicited on that cycle. Finn and Metcalfe therefore assumed that if cycle 1 JOLs contribute to the UWP effect, then not eliciting them would mitigate or eliminate the effect. Thus, cycle 1 of their procedure consisted of a study phase (S) and a test (T), while cycle 2 remained intact and consisted of a study/JOL (SJ) phase and a test (S-T-SJ-T). The results of this experiment were no different from results of their previous experiments which employed a full UWP procedure. Thus it seems that eliminating an overt JOL phase does not change the results.

Also, eliminating test phases does not eliminate the UWP effect. Karpicke (2009) employed repeated study-test cycles in a study investigating the testing effect (cf. Roediger & Karpicke, 2006). To examine the influence of repeated testing on memory performance, he created three experimental conditions which differed in the number of study/JOL and test phases. The first condition resembled the one commonly used in investigating the UWP effect and consisted of three consecutive study-test cycles with immediate JOLs in each study phase (SJ-T-SJ-T-SJ-T). In the second condition, the first test phase was replaced with an additional study/JOL phase (SJ-SJ-SJ-T-SJ-T), and in the third condition also the second test was replaced with a study/JOL phase (SJ-SJ-SJ-SJ-SJ-T). The results revealed the typical UWP pattern in the

SJ-T-SJ-T-SJ-T condition, with mean JOLs being higher than recall performance on cycle 1, and significantly lower on cycles 2 and 3. Of more interest, however, is that the UWP effect was present for all tests both in the SJ-SJ-SJ-T-SJ-T and SJ-SJ-SJ-SJ-SJ-T conditions. These results suggest that prior testing is not a necessary prerequisite for the UWP effect to occur.

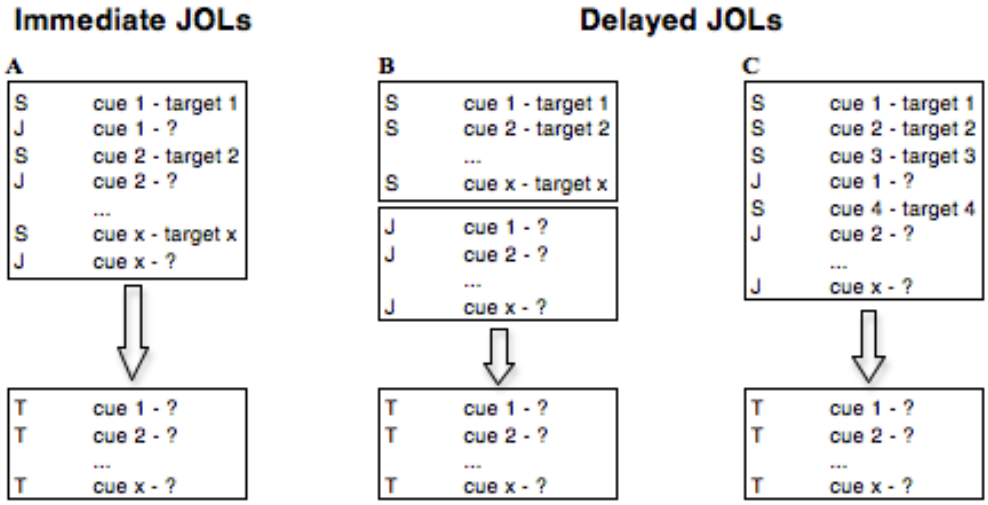

Figure 1.1. A comparison of procedures used to elicit immediate and delayed JOLs. S, J and T refer to study, JOL and test stages of the procedure, respectively.

Finally, there have been claims that even eliminating both the JOL phase and the test phase from cycle 1 does not eliminate the effect. In one of the conditions of Meeter and Nelson's (2003) experiment, no JOLs were elicited on cycle 1 and the first test was omitted (S-SJ-T). Their results for this condition showed that JOLs were 8% lower than recall performance on cycle 2. However, the authors did not report whether this effect was statistically significant.

The UWP effect is usually examined with immediate, rather than delayed JOLs. The difference between these two types of JOLs is presented in Figure 1.1. Whereas immediate JOLs are made immediately after an item had been presented (panel A), delayed JOLs are elicited

either after the whole study phase (panel B) or during the study phase, but the study and JOL stages for each pair are separated by some amount of time or some number of other items (panel C).

The main difference between immediate and delayed JOLs with respect to the UWP effect lies in the fact that whereas for immediate JOLs this pattern of results is found under most circumstances, for delayed JOLs the results are much less consistent. In some studies the UWP effect was present. For example, Scheck and Nelson (2005) found that for easy word pairs, the effect appeared on cycle 2. Similarly, in Serra and Dunlosky's (2005) Experiment 1 participants in the delayed condition were well calibrated on cycle 1, but their JOLs increased from cycle 1 to cycle 2 to a lesser extent than recall did. The difference between JOLs and recall was much smaller than in their immediate condition, though, and equalled 5.1%; the authors did not report whether this difference was statistically significant. Other studies have reported no differences between the measures or even a slight tendency of JOLs to exceed recall performance. In Scheck and Nelson's difficult condition, participants' JOLs were higher than recall on cycle 1, but were well calibrated on cycle two. The results of Finn and Metcalfe's (2007) two experiments showed that in the delayed condition JOLs exceeded recall performance on cycle 1; however, on cycle 2, JOLs and recall were of the same magnitude. In Koriat et al.'s (2006) Experiment 2, the mean of delayed JOLs exceeded recall performance on all cycles. JOLs were significantly higher than recall on cycle 1, and, in spite of a numerical reduction from cycle 1 to cycle 2, this difference still approached significance on cycles 2 and 3 ($p = .06$ and $p = .09$, respectively).

One possible reason for this inconsistency between immediate and delayed JOLs may be the way in which JOLs were elicited. When delayed JOLs were made during the study phase (see panel C of Figure 1.1), as in Scheck and Nelson's (2005) and Serra and Dunlosky's (2005) experiments, the UWP effect for easy word pairs was present. When a

separate JOL phase was implemented (see panel B of Figure 1.1), however, as in Finn and Metcalfe's (2007) study and Koriat et al.'s (2006) Experiment 2, no UWP occurred. It is viable that people base their delayed JOLs on different types of cues, depending on whether JOLs are elicited during or after the study phase; however, no systematic research on this topic has been conducted.

Another possible explanation for different results for the two JOL types will be explained in more detail in the section on Scheck and Nelson's (2005) anchoring-and-adjustment hypothesis. To preview, according to this explanation, the magnitude and direction of the difference between JOLs and recall performance depends on how much people remember from the study list. If the percentage of correct answers is low (less than 30%), a tendency for JOLs to exceed the level of recall can be found. If, however, people can answer half of the questions or more, this pattern of results is usually reversed. As the procedures aimed at eliciting delayed JOLs usually produce lower recall performance than when immediate JOLs are elicited even if presentation times are kept constant, it is possible that it is this low memory for targets that sometimes eliminates the effect.[1]

[1] The reason for which procedures during which delayed JOLs are elicited produce lower recall performance than when immediate JOLs are made can be seen in Figure 1.1. When delayed JOLs are elicited (panels B and C), a presentation of a pair is immediately followed either by another pair or a JOL stage for a different pair. In the immediate task (panel A), on the other hand, presentation of a pair is immediately followed by a JOL stage for the same pair, with a cue still displayed on the screen, and time for making a JOL is not limited. This gives participants more time to apply effective encoding strategies which, in turn, improve recall performance.

2. Possible mechanisms of the UWP effect

There has been a debate in the literature concerning the possible mechanisms causing or contributing to the magnitude of the UWP effect. Generally, these explanations can be divided into two groups differing at a very basic level. According to the advocates of the first group of mechanisms, the UWP effect is a manifestation of a psychological mechanism of underconfidence, whereas the proponents of the second approach claim that it is merely an artefact of using scales to elicit JOLs. The next sections will be aimed at presenting the basic assumptions underlying each of these approaches and theories of the UWP effect that have been developed.

2.1 The UWP effect as a manifestation of psychological underconfidence

This approach to UWP was chronologically first - hence the effect was named *underconfidence*-with-practice. According to this viewpoint, the mean of JOLs has a psychological meaning: when compared with recall performance, the mean JOL indicates the extent to which people are aware of the effectiveness of their learning. If a person's mean JOL exceeds the level of recall, it can be assumed that this person is overconfident - that is, she overestimates her ability to learn from cycle to cycle. Analogously, JOLs lower than recall suggest underconfidence. Finally, when JOLs and recall performance are approximately equal, this suggests that this person is well aware of her learning abilities.

This understanding of the UWP effect requires an assumption that is rarely formulated explicitly in the literature: that by making JOLs, participants judge the *probability* of future recall. In other words, the ultimate goal of a participant is to assign JOLs that would correctly predict the proportion of items recalled at a later test (a *frequentist approach* to probability). For any particular value of JOL, calibration is maximised when the proportion correct of all items for which this value was assigned

matches that value: for example, 10% of items assigned a 10% JOL are recalled. Recalling 15% of such items would suggests that a person was underconfident at this particular level. Therefore, each value on a percentage scale has a predefined, objective meaning. As a result if people were well calibrated at each level of JOLs, their mean JOLs from all study phase should match global recall performance. If, however, people were well calibrated only on some levels, but underconfident at others, global underconfidence would be found.

To date, four mechanisms of the UWP effect have been proposed that assume that the effect is a result of psychological underconfidence. The next sections will provide an overview and discussion of each of these mechanisms.

2.1.1 Cue-utilisation approach

The first psychological explanation of the UWP effect was suggested by Koriat (1997) and elaborated by Koriat et al. (2002). According to the cue-utilisation approach proposed by Koriat, there are three types of cues that influence JOLs: intrinsic, which pertain to the characteristics of studied items; extrinsic, which pertain to the characteristics of the learning experience; and mnemonic, which are based on learners' subjective experience with studying a particular item. Intrinsic and extrinsic cues can affect JOLs directly or indirectly. A direct influence of these two types of cues involves applying rules or theories about the influence of these cues on memory (Koriat, 1997); for example, a person may apply a rule that states that longer presentation times should improve learning. These cues can also affect JOLs by influencing mnemonic cues, which in turn influence JOLs. Koriat's initial explanation of the UWP pattern was based on two implications of his cue-utilisation framework. The first states that people discount the contribution of extrinsic factors to learning, concentrating more on the intrinsic factors. The second implication is that

as learning progresses, people's reliance on intrinsic cues decreases as mnemonic cues become more available.

The first of these propositions was directly tested by Koriat (1997) in four experiments. In his Experiment 1, participants were assigned to one of two conditions: in the first (Same condition), they were presented with the same list of paired associates in two study-test cycles, whereas in the second condition (Different condition), participants studied two different lists, and each of the lists was only presented and tested once. Except for this manipulation, the procedure was identical in both groups. The aim of this experiment was to investigate the extent to which JOLs and recall performance would be influenced by an extrinsic factor of list repetition. The results for calibration showed that whereas in both groups participants were relatively well calibrated on cycle 1, on cycle 2 the mean of their JOLs was significantly lower than their recall performance, showing the UWP pattern. The effect was, however, stronger in the Same than in the Different group and the data provided by Koriat in his paper does not answer the question whether the difference on cycle 2 in the Different group was significant. List repetition also produced the UWP pattern in Experiment 2, although the effect was moderated by feedback concerning correctness of the answers. These results suggest, according to Koriat, that the extrinsic factor of list repetition affects JOLs and recall to a different extent: JOLs increase less from cycle to cycle than recall does.

Koriat's (1997) Experiment 3 tested the prediction that another extrinsic factor - the number of repetitions within one list - would also influence recall to a larger extent than JOLs even when there was only one study-test cycle. To test this hypothesis, some of the words within a list were repeated to investigate whether the same pattern that exists for list repetition would emerge. To this end, word pairs were presented one, two or three times within one study list. The results showed that although for items that were presented once mean JOLs matched recall performance,

suggesting good calibration, for items that were presented two or three times the UWP effect was revealed.

Finally, in Experiment 4 presentation times were varied within one list. Participants underwent only one study-test cycle and each pair of words from the to-be-learned list was presented only once for 2, 4 or 8 s. As it could be expected that longer presentation times would lead to better learning, the question was whether this manipulation would also exert stronger influence on recall than on JOLs. The results confirmed this prediction: whereas for the shortest presentation times mean JOLs and recall performance were almost equal, longer presentation times increased JOLs to a smaller extent than recall, revealing the UWP pattern.

Overall, the results of these four experiments suggested that JOLs are influenced by extrinsic factors such as list repetition, pair repetition or presentation times to a smaller extent that recall is, thus supporting the first part of Koriat's (1997) prediction - that people discount the information provided by extrinsic cues when making JOLs. The second part of this prediction - that people rely on intrinsic cues - was tested in Koriat's study by manipulating item difficulty within each study list. In Experiment 2, "easy" pairs were defined as those for which the cue elicited the target with a probability exceeding .05, while for the "difficult" pairs there was no association whatsoever between the cue and the target. The difficulty of pairs used in Experiments 1, 3 and 4 was assessed in a separate study by a different group of participants who assessed memorability of each pair on a scale from 0 to 100%; "easy" targets were defined as those for which memorability was higher than median, whereas for "difficult" items memorability was lower than median.

The results were inconsistent: whereas in Experiment 1 the difference between easy and difficult pairs was higher for JOLs than for recall, revealing the UWP pattern for the latter type of pairs, in Experiment 4 this pattern was reversed, and in Experiments 2 and 3 pair difficulty influenced JOLs and recall to the same extent. Koriat's (1997) results

suggested, thus, that the intrinsic factor of pair difficulty does not contribute to the UWP effect. It is, however, also possible that this inconsistency could be explained by Scheck and Nelson's (2005) anchoring-and-adjustment hypothesis. This explanation of the UWP effect will be covered in more detail later in this chapter; to preview, according to Scheck and Nelson at the beginning of a study cycle participants set an anchor for their JOLs, usually between 30% and 50%, and their subsequent judgements are pulled towards this anchor. When recall performance is above the anchor set for JOLs, this distortion of judgements produces the UWP pattern. This was the case for both difficult and easy pairs in Koriat's Experiment 1, for which recall performance was 56% and 84%, respectively.[2] Thus the hypothesis that the intrinsic factor of pair difficulty influences JOLs should not be ruled out without further research.

The other implication of the cue-utilisation approach was not tested directly by either Koriat (1997) or Koriat et al. (2002) who only speculated about the ways in which using mnemonic cues could impair calibration. Koriat et al. compared the pattern of results found in the experiments investigating the UWP effect to the one present in Runeson, Juslin, and Olsson's (2000) study on perception of dynamic scenes. Runeson et al. noted that when their participants started learning to perform a complex visual task, they used simpler, inferential cues; with experience, however, they switched to using sensory-based heuristics.[3] What is important is that

[2] For Experiments 2-4, no means for recall performance were presented for easy and difficult word pairs. The means presented here for Experiment 1 are approximations taken from Koriat's (1997) Figure 2.

[3] In the study of Runeson et al. (2000), participants observed collisions of objects and their task was to decide which of the colliding objects was heavier. The example of a switch from inferential to perceptual cues is best described by the experience of one participant, who at the beginning "had struggled ambitiously to master the task by attending to the motions and how they changed, trying various ways to infer the relative

this shift in strategies used to perform the task was accompanied by a shift from either overconfidence or good calibration towards underconfidence. Koriat et al. found this result analogous to the pattern present in their JOL tasks in which participants switched from inferential to mnemonic, heuristic-based cues. However, there were no data to support the assumption that this change in the type of cues that participants use actually produces the UWP pattern.

Serra and Dunlosky (2005) tested a hypothesis that the UWP effect may be caused by giving unduly low JOLs to items which are retrieved successfully, but with some difficulties. The idea was based on the results of Benjamin, Bjork, and Schwartz (1998) who showed that retrieval fluency - a mnemonic cue, according to Koriat's (1997) cue-utilisation framework - can serve as a basis for JOLs. They found that as response latency on cycle 1 (their measure of retrieval fluency) increases, the possibility of correct retrieval also increases, but cycle-2 JOLs decrease. If this factor were to cause UWP in a multi-cycle study-and-test procedure, the UWP effect should be more pronounced for items with cycle-1 long retrieval latencies and JOLs should be negatively correlated with the speed of retrieval at cycle-1 test.

To test these predictions, Serra and Dunlosky recorded retrieval latencies in a two-cycle procedure. In Experiment 1, they compared gamma correlations between JOLs (immediate or delayed) and cycle 1 retrieval latencies. They found that although mean gammas were no different from 0 for delayed JOLs, they were different - although not to a great extent - for immediate JOLs. This result suggested that immediate JOLs may at least partially be based on cycle 1 retrieval fluency. However, the authors failed to find a link between retrieval latencies and the magnitude of the UWP effect: the difference between the mean of

---

mass from physical principles". As the task progressed, this person "felt there was no use trying any more and started to just look and respond, to get it over with" (p. 547).

immediate JOLs and recall performance did not change with increasing retrieval latencies.

In Experiment 2, Serra and Dunlosky (2005) added an overt pre-JOL recall phase to the delayed condition, assuming that if it is the outcome of this additional retrieval attempt made during the JOL stage of the procedure that influences delayed JOLs, then it is at this stage that fluency of retrieval could matter. To test this hypothesis, they recorded retrieval fluency for those pre-JOL retrieval attempts. This time, the mean gamma correlation between retrieval fluency and delayed JOLs was significantly different than 0. However, again there was no difference in the magnitude of the UWP effect when retrieval latencies were taken into account. To test whether the results of the first two experiments were contaminated by mixing delayed and immediate judgements within one study list, in Experiment 3 judgement type (immediate or delayed) was manipulated between participants. As in Experiment 2, pre-JOL attempts were recorded in the delayed condition. Nevertheless, the pattern of results was consistent with the one found in previous experiments, showing weak gamma correlations between both immediate and delayed JOLs and retrieval latencies, but no direct relationship between retrieval latencies and the UWP effect for either delayed or immediate JOLs. Therefore the results of Serra and Dunlosky's experiments suggest that even if retrieval fluency can affect the magnitude of JOLs in a multi-cycle procedure, its effects are probably negligible.

In general, only one of two predictions stemming from Koriat's (1997) cue-utilisation approach garnered some evidence as a possible contributor to the UWP effect. Koriat's data suggest that undervaluing extrinsic cues may at least partially explain this pattern of results. On the other hand, the evidence for the assumption that intrinsic cues influence JOLs and recall to the same extent is mixed, and currently there is little empirical support for the idea that it is a general switch from intrinsic to mnemonic cues that produces the UWP pattern of results. The results of Serra and Dunlosky

(2005) concerning retrieval fluency, and the mixed results for delayed JOLs - thought to be influenced mostly by the outcome of a covert retrieval attempt (cf. Nelson & Dunlosky, 1991), which is a mnemonic cue (Koriat, 1997) - also challenge the assumption that mnemonic cues in general lead to the UWP pattern. However, it is still possible that a specific type of mnemonic cue may contribute to the UWP effect. Such a cue will later be discussed in section 2.1.3 of this chapter which describes the memory-for-past-test account of the UWP effect.

2.1.2 Mnemonic debiasing account

Koriat et al. (2006) proposed that the UWP pattern can be partially explained by the foresight bias (Koriat & Bjork, 2005, 2006) - an illusion of competence caused by making judgements either in the presence of the to-be-remembered material or immediately after its presentation. To examine this type of bias, Koriat and his colleagues used a priori and a posteriori associated pairs. The difference between these two types of pairs lies in the direction of the relationship between the words consisting the pair: whereas a priori associations occur when the probability of the cue bringing the target to mind is high, a posteriori associations are evident only when both words from a pair are presented simultaneously. One example of such a posteriori pairs that the authors used in their experiments are backward-associated pairs.[4] Koriat and Bjork (2005) showed that when pairs were backward associated, participants had a strong illusion of competence which inflated their JOLs so that they

[4] Consider a backward-associated a posteriori pair *cheese-cheddar*. Although the association between the words seems to be strong, the word *cheddar* is not among the common associates of the word *cheese* (according to Nelson, McEvoy and Schreiber's, 1998, norms, the likelihood of eliciting *cheddar* when cued with *cheese* is only 5%). Were the order of the words reversed, however, the likelihood of eliciting *cheese* when cued with *cheddar* would be high (92%), making *cheddar-cheese* an a priori (and forward-associated) pair.

became higher than recall performance. For forward-associated pairs, on the other hand, participants showed good calibration. This effect stemmed from the fact that for both types of pairs participants' JOLs did not differ, whereas recall was lower for pairs associated backwards. Similarly, weakly associated a priori pairs produced an illusion of competence, making mean JOLs higher than recall performance.

In a follow-up study, Koriat and Bjork (2006) analysed the influence of such illusions of competence on JOLs when the experimental procedure consisted of more than one study-test cycle. They assumed that the foresight bias, which would inflate cycle 1 JOLs, should be alleviated on cycle 2 and beyond because of mnemonic debiasing - moderating the initial overconfidence by adjusting the inflated JOLs downward. According to this explanation, this should happen because participants' reliance on mnemonic cues increases after the first study-test cycle. If this was the case, no UWP should be found in conditions in which the difference between cycle 1 JOLs and recall was the highest, and UWP should occur in conditions where this difference was low or nonexistent. In their Experiment 1 they showed that on cycle 1, JOLs for backward-related and unrelated pairs exceeded their recall performance, while for forward-related pairs the two measures were no different from each other. From cycle 2 onward, on the other hand, participants' JOLs were lower than recall for all types of pairs. These results were further replicated by Koriat et al. (2006). Taken together, these data suggest that the initial illusion of competence does not eliminate the UWP effect on subsequent study-test cycles.

Koriat et al. (2006) directly tested another prediction stemming from the mnemonic debiasing hypothesis that it is the presence of the to-be-remembered material that contributes to the UWP effect. To this end, they compared immediate and delayed JOLs in one experimental design. They assumed that the illusions of competence should be more pronounced for the former type of JOLs, which are made immediately after the

presentation of a cue-target pair, than for the latter, which are usually made when only the cue is present. Experiment 2, consisting of three study-test cycles, revealed the UWP pattern on cycles 2 and 3 in the immediate condition and good calibration in the delayed condition. In Experiment 3 only one aspect of the procedure was changed: JOLs were made in the presence of the cue-target pair rather than cue alone in order to induce foresight bias on cycle 1 in the delayed condition. It was thus expected that the UWP pattern would emerge in both conditions. The results confirmed this prediction: the UWP effect was present on cycles 2 and 3 both in the immediate and in the delayed condition.

Koriat et al. (2006) interpreted their results as supporting the mnemonic debiasing account. However, they noted that it cannot be the sole cause of the UWP effect: as mnemonic debiasing is supposed to reduce the initial illusions of competence, it may be responsible for the change in the magnitude of JOLs from cycle 1 to cycle 2, but there is no reason to suspect that this mechanism would also influence cycle 3 JOLs, as on cycle 2 and beyond people underestimate rather than overestimate their performance. Another problem is the composition of the study list used by Koriat and his colleagues (Koriat and Bjork, 2005, 2006; Koriat et al., 2006) in which different pair types were used to induce illusions of competence. It is unknown to what extent, if any, such illusions can occur when a study list consists of unrelated word pairs, which is the case in most studies investigating the UWP effect. It is conceivable that even when pairs are unrelated, participants may overestimate the probability of cycle 1 recall because of factors such as lack of experience with the experimental task; this would be an example of an illusion of competence caused by a factor different than foresight bias. In this case, it would be possible that similar debiasing would occur between cycles 1 and 2, producing the UWP effect. Again, however, there would be little need for this process to take place between cycles 2 and 3. It can be thus concluded that although there is some support for the mnemonic debiasing

explanation of the UWP effect, this account is unable to explain most of the results found in the literature.

2.1.3 Memory-for-past-test heuristic

Another possible mechanism contributing to the UWP effect has been put forward by Finn and Metcalfe (2007, 2008). Their starting point was that one of very few manipulations that was found to prevent the UWP effect from emerging was delaying JOLs. For this reason the authors assumed that different heuristics must be used when people make immediate and delayed JOLs. It is usually assumed that when people make delayed JOLs, they base them on a retrieval attempt made during the JOL stage of the procedure (cf. Nelson & Dunlosky, 1991; Metcalfe & Finn, 2008). This information is, however, unavailable when immediate JOLs are made: the item is still in working memory so there is no opportunity to retrieve a target from long term memory while making such judgements. Nevertheless, Finn and Metcalfe (2007) proposed that people also base immediate JOLs on their recall performance, albeit a different type: on cycle 2 and beyond they have access to information about the effectiveness of their recall attempt at the preceding test. This heuristic, dubbed memory-for-past-test (MPT), could then explain why people underestimate their recall performance - according to Finn and Metcalfe's explanation, people pay too little attention to the fact that JOLs are made during an additional study phase which should lead to an increase in the degree of mastery of to-be-learned material. This heuristic would be used for immediate, but not delayed JOLs, as for the latter type of judgements more diagnostic information about the current, covert retrieval attempt is available. The MPT hypothesis predicts a specific pattern of results for immediate JOLs. According to this hypothesis, items recalled on the preceding test should be assigned high JOLs, whereas items that were not recalled should be given low JOLs.

To test their hypothesis, Finn and Metcalfe (2007) compared immediate and delayed JOLs in a two-cycle design. Although UWP is a calibration effect that can only be found at the list level, to see whether a JOL for a word pair can be a result of a past retrieval attempt of a target from this pair an analysis at the single-item level was necessary. To this end, the authors used multiple regression to evaluate the contribution of past and future test performance on cycle 2 immediate and delayed JOLs. For immediate JOLs, standardised beta coefficients were higher for cycle 1 test than for cycle 2 test, suggesting that past test performance can better predict cycle 2 JOLs than future test performance. For delayed JOLs, on the other hand, this pattern was reversed. These results are consistent with the MPT hypothesis, according to which past test should predict well cycle 2 immediate, but not delayed JOLs.

In a subsequent study, Finn and Metcalfe (2008) provided further experimental evidence that participants employ the MPT heuristic when making immediate JOLs. They assumed that if people based their JOLs on past test performance, manipulating cycle 1 encoding while keeping cycle 2 recall constant should provide a test of this explanation. To this end they varied the number of presentations of each pair on cycles 1 and 2 in Experiment 1 and presentation times in Experiment 2. In Experiment 1, word pairs were presented 6 times in total: half were presented 5 times on cycle 1 and 1 time on cycle 2 (5-1 condition), whereas the reverse was true for the other half (1-5 condition). In Experiment 2, word pairs were presented in two cycles for 9 s in total, either for 1 s on cycle 1 and 8 s on cycle 2 (1-8 condition) or for 8 s on cycle 1 and 1 s on cycle 2 (8-1 condition).

The reason for these manipulations was to examine the influence of cycle 1 test performance on cycle 2 JOLs. The results of both experiments confirmed Finn and Metcalfe's predictions. In the conditions in which cycle 1 encoding was better (5-1 in Experiment 1 and 8-1 in Experiment 2) participants' cycle 2 JOLs were higher than in the remaining conditions (1-

5 and 1-8, respectively) even though at the time they were elicited the total number of presentations or total presentation times were equal for all pairs. The results suggest that the larger the increase in recall performance from cycle to cycle, the less able people are to track their recall with their JOLs and thus the greater the discrepancy between JOLs and recall becomes. However, JOLs for items not recalled at test 1 and recalled at test 2 were significantly higher than those made for items that were not recalled either at test 1 or at test 2. This result should not emerge if people relied solely on the MPT heuristic: in this case test 1 recall performance was identical for both classes of items, so no differences in cycle 2 JOLs should be expected.

To test one of the implications stemming from the MPT hypothesis - that people need to have access to the effectiveness of their previous recall attempts - Finn and Metcalfe (2008) asked their participants to judge their past test performance instead of making cycle 2 JOLs. As predicted, people were accurate in assessing their previous recall: in 94% of the cases participants correctly identified recalled items as such, and false alarms were at floor level. It thus seems that participants can potentially access information that is necessary for them to use the MPT heuristic.

Finn and Metcalfe's (2008) further experiments examined alternative explanations of the results of their Experiments 1 and 2. In Experiment 4, in which the same 1-1, 1-5, 5-1 and 5-5 conditions as in Experiment 1 were created, participants were asked to make encoding fluency judgements instead of JOLs. The authors hypothesised that a greater number of presentations could lead to more fluent encoding, and this in turn could influence cycle 2 JOLs. In order to use this information, however, people would have to be able to access it on cycle 2. To test whether this information is available to participants during the second study phase, Finn and Metcalfe eliminated cycle 1 JOL and test phases and instructed their participants to judge the fluency of cycle 1 encoding immediately after the presentation of each pair on cycle 2. These

judgements of past encoding fluency were made on a 0-100% scale. The results revealed no differences between the 5-1 and 1-5 conditions with regard to rated cycle 1 encoding fluency, suggesting that participants were unable to notice any differences in past encoding. However, the differences between 1-1 and 1-5 conditions were significant, as were the differences between 5-1 and 5-5 conditions. Note that these pairs of conditions differed only with respect to *current* encoding, while past encoding, which was supposed to be assessed by participants, was kept constant. Together, the results suggest that information regarding encoding fluency on cycle 1 is not easily accessible to participants on cycle 2. It is therefore unlikely that memory for past encoding fluency could be responsible for the UWP effect in the first two experiments.

Experiments 5a and 5b investigated another possible hypothesis - that people base their cycle 2 JOLs on their cycle 1 JOLs. It seemed possible that the effect could be caused by cycle 1 JOLs for items presented once being lower than JOLs for items presented 5 times. Therefore, participants in Experiment 5a were asked to recall their past JOLs instead of giving prospective JOLs on cycle 2. The results revealed a pattern of results similar to the one found in Experiments 1 and 2, with past JOLs being lower than recall levels both in the 1-5 condition and in the 5-1 condition. Therefore it seems that using memory for past JOLs could produce the UWP pattern. However, the difference between past JOLs and recall did not differ significantly between the conditions, contrary to what was found in the previous experiments. Also the improvement in resolution from cycle 1 to cycle 2 was not as pronounced as in Experiments 1 and 2.

All in all, these results suggest that participants have access to their past JOLs and although it is not impossible that they make use of this information when making cycle 2 JOLs, it is unlikely that memory for past JOLs is the sole cause of the UWP effect. To find further evidence that memory for past JOLs is not enough to produce the UWP pattern, in

Experiment 5b the authors eliminated cycle 1 JOLs. If cycle 2 JOLs were indeed based on cycle 1 JOLs, this should have eliminated the UWP effect. Thus, the experimental procedure consisted only of cycle 1 study phase followed by a test and cycle 2 study/JOL phase followed by a test (S-T-SJ-T). Except for that, the design was identical as in Finn and Metcalfe's Experiment 1, with 5-1 and 1-5 conditions differing in the number of pair repetitions on cycles 1 and 2. Also the results were no different: the UWP pattern was found in the 1-5, but not in the 5-1 condition. Finn and Metcalfe thus concluded that participants did not base their cycle 2 JOLs on their previous judgements. However, it has to be noted that it cannot be known whether this manipulation managed to eliminate cycle 1 JOLs. It is possible that in spite of eliminating *explicit* JOLs, participants still make *implicit* judgements during the study phase, which in turn contribute to the difference between the levels of JOLs and recall. Although it is not likely that such implicit JOLs would be able to be the sole cause the UWP effect in this case, they may nevertheless affect its magnitude.

Tauber and Rhodes (2012) proposed and tested two possible bases of the MPT heuristic. Their direct-memory hypothesis suggested that participants explicitly retrieve information about previous recall and use this information to make JOLs. This explanation was consistent with the one suggested by Finn and Metcalfe (2007, 2008). On the other hand, the indirect-memory hypothesis suggested that MPT influences other processes, such as processing fluency, which in turn influence JOLs.

To test these hypotheses, Tauber and Rhodes (2012) compared performance of older and younger adults in a three-cycle procedure. As older adults' episodic memory is worse than in younger adults, if the MPT heuristic required direct access to outcomes of previous retrieval attempts, worse performance should be observed in the older age group. If, however, MPT influenced JOLs indirectly, no differences should be

observed between the age groups, as implicit memory measures are usually equivalent for older and younger adults.

There were three experimental conditions: one consisting of older participants and two of young adults. For one of the young adults conditions, the procedure was identical to that in the older adults condition (the young-same condition). In the other condition, presentation times were shortened to make younger adults' recall levels match that of older adults (the young-matched condition). This allowed the authors to check whether any differences between the age groups are not caused by different levels of memory performance. The results of Tauber and Rhodes' (2012) experiment showed that the UWP effect occurred in all conditions. Further analyses investigated which variables served best as predictors of JOLs. The results suggested that although in all conditions participants relied on past test performance when making JOLs, older adults and younger adults in the young-matched condition relied on MPT to a greater extent than younger adults in the young-same condition did. Analyses for cycle 3 also showed that people only take into account their test performance on the cycle immediately preceding the current one. It was also revealed that in all conditions current trial recall and prior JOLs had influence on JOL levels.

In total, at least partial support was found for the indirect-memory hypothesis of the MPT heuristic, as the pattern of results found in Tauber and Rhodes' (2012) study was unlikely to be caused only by relying on direct memorial evidence for past recall. Nevertheless, it is still possible that this direct memory can, at least partially, influence JOLs: as was found by both by Tauber and Rhodes, and Finn and Metcalfe (2008), it is very unlikely that UWP is caused by a single factor.

England and Serra (2012) compared the contributions of anchoring (Scheck & Nelson, 2005) and MPT to the UWP effect. To this end, in Experiment 1 they manipulated the presence (or absence) of JOL and test phases on cycle 1, thus creating four conditions: SJ-T-SJ-T, S-T-SJ-T, SJ-

SJ-T and S-SJ-T. To equate time spent by participants on cycle 1, they also manipulated presentation times during the study phase, so that they were the shortest in the SJ-T-SJ-T condition, intermediate in the S-T-SJ-T and SJ-SJ-T conditions and the longest in the S-SJ-T condition. Their results showed the UWP pattern in all four conditions, but it was less pronounced in the conditions in which cycle 1 test was present than in those in which it was absent - a pattern which England and Serra assumed to contradict the MPT hypothesis. The authors also split the items in the SJ-T-SJ-T condition according to whether they had been recalled on cycle 1. They noted that for recalled items, JOLs increased from cycle 1 to 2, whereas for unrecalled items, they remained at the same level. UWP was present for both types of items; however, it was stronger for those that were recalled on cycle 1. England and Serra assumed that this confirms the anchoring hypothesis.

However, such an interpretation of England and Serra's results poses a few problems. First, Koriat (1997) and Finn and Metcalfe (2008) found that presentation times influenced the magnitude of the UWP effect. These differed between conditions in England and Serra's experiments and so they may have confounded the results.

Second, England and Serra stated that "according to the past-test hypothesis (...), previously tested items should demonstrate greater underconfidence and relative accuracy across the two study phases relative to non-tested items" (p. 2). Such a prediction, however, does not stem from the MPT hypothesis. The MPT heuristic is only assumed to be a major source of UWP *when information about past test performance is available*. It is conceivable that when information about previous recall is unavailable, people attend to other cues which may be even less diagnostic of future recall. Thus the pattern of results found in England and Serra's Experiment 1 does not provide evidence against MPT as a viable explanation of the UWP effect.

Third, England and Serra did not give any reason to assume that the anchor point is equivalent to the mean of cycle 1 JOLs for unrecalled items. On the contrary, Scheck and Nelson (2005) demonstrated that cycle 1 JOLs were significantly higher than cycle 1 recall both for easy and difficult word pairs in the immediate condition and for difficult pairs in the delayed condition - a pattern of results that would not have been possible if cycle 1 JOL levels served as an anchor. What is more, if an anchor was to be set based on cycle 1 JOLs, it would be difficult to explain the presence of the UWP pattern when no JOLs are elicited on that cycle.

Finally, the results for pairs recalled and unrecalled on cycle 1 are not inconsistent with Finn and Metcalfe's (2007) explanation: as these authors noted, when recall performance is at ceiling (as was the case of items recalled on cycle 1 in England and Serra's (2012) study - recall level was at 94%), any variability in the judgements would lead to UWP, as 100% serves as the upper limit for JOLs (see section 2.2.2 for details). The results of Experiment 1 therefore can be explained both by the MPT and the anchoring hypotheses, thus giving little evidence in support of one of them over the other.

2.1.4 Stability bias

Kornell and colleagues (Kornell & Bjork, 2009; Kornell, Rhodes, Castel, & Tauber, 2011) investigated a phenomenon dubbed the stability bias in human memory. A general conclusion from their research was that people act as if their memories were not to change in the future: they fail to predict both the degree of their learning and forgetting. It was thus possible that the UWP effect could stem from the fact that people do not appreciate the value of additional study and test phases in improving memory. Kornell and Bjork (2009) gave participants in their Experiment 1 four study-test phases and asked them to make immediate predictions of future learning (POLs) on a 0-100% scale during the study phases: they were supposed to predict their performance on either test 1, test 2, test 3

or test 4 on a between-participant basis (thus creating four conditions: S-T, S-T-S-T, S-T-S-T-S-T and S-T-S-T-S-T-S-T). For example, participants in the S-T-S-T-S-T condition were supposed to make predictions of test 3 performance during study phases 1, 2 and 3. The results showed that participants in all conditions except the S-T underestimated their future recall: mean POLs seemed to be influenced by past or current test performance. Even in the S-T-S-T-S-T and S-T-S-T-S-T-S-T conditions, where participants had the chance to experience the extent to which recall increased from cycle to cycle, POLs approached the actual recall levels for the last test only on the last cycle.

The stability bias explanation seems to be consistent with the MPT hypothesis which assumes that people base their JOLs on their performance on the immediately preceding test. In fact, according to Kornell and Bjork (2009), accessing the information about past test performance rather than using knowledge about the benefits of repeated re-exposure to the to-be-learned materials may be the reason for which the UWP effect occurs. It is also not inconsistent with Koriat et al's (2002) notion that people may appreciate the effects of additional study phases on learning, but not that of retrieval practice.

This account may also offer an alternative explanation of different patterns for delayed JOLs found in Koriat et al.'s (2006) Experiments 2 and 3. When only cue was present during the JOL phase in Experiment 2, no UWP was found on cycles 2 and 3, whereas this effect was present when both cue and target were presented to participants at the time of making JOLs in Experiment 3. A comparison of mean JOLs and recall performance in both experiments suggests that whereas JOL levels on each cycle were similar between the experiments, recall was not: it increased to a larger extent in Experiment 3. It is thus possible that participants who were given additional presentations of the study list during the JOL phase of the experiment discounted the effect of this presentation in their judgements of future recall. On the other hand, this

account cannot explain Koriat's (1997) results which show that the UWP pattern can be found in single-cycle procedures when the number of repetitions or presentation times vary within a list.

However, the stability bias explanation describes the UWP effect at a higher level - as an instantiation of a general metacognitive tendency - without explaining which specific mechanisms could be responsible for this pattern of results. It also has to be noted that the relationship between the judgements used in Kornell and Bjork's (2009) study and JOLs is currently not known, so more research would be needed to confirm whether it is discounting future learning that influences the UWP effect on JOLs. Finally, the interpretation of the stability bias has recently been questioned by Ariel, Hines, and Hertzog (2014), who suggested that it may simply stem from misinterpretation of experimental instructions rather than from actual beliefs about stability of memory performance: when instructions were framed in a way that concentrated on benefits of future *study* rather than benefits of future *testing*, the sensitivity of POLs to the number of future study-test cycles increased.

2.1.5 Summary and conclusions

The four mechanisms described in the preceding sections share one quality: they all assume that this effect stems from true underconfidence on cycle 2 and beyond; that is, the UWP pattern results from underestimation of the degree of learning with additional practice. According to Koriat's (1997) cue-utilisation framework, this may stem from the fact that people undervalue mnemonic and extrinsic cues. Koriat and his colleagues (Koriat & Bjork, 2006; Koriat et al., 2006) also suggested that UWP may be a side effect of mnemonic debiasing that occurs after cycle 1 to moderate the effects of illusions of competence. Finn and Metcalfe (2007, 2008) proposed that people base their JOLs on their memory for past test performance; this MPT heuristic would unduly lower their JOLs for items that were not recalled on a preceding test. Finally,

Kornell and Bjork (2009) suggested that the UWP effect is a manifestation of a more general effect found in metacognition - a stability bias; according to this explanation, people base judgements on their current memory state and do not take into account that learning or forgetting may occur.

Of these mechanisms, the MPT heuristic has garnered the most support. It can account well for most of the results found in the literature of the UWP effect. It can also accommodate other explanations that see the UWP effect as a manifestation of participants' underconfidence (switching to mnemonic cues, stability bias). However, it can only explain why UWP occurs for a subset of items: those that were not recalled on a preceding test. Yet the UWP pattern is commonly found also for previously recalled items. Therefore the MPT account needs to be complemented with another mechanism that would explain the UWP effect for recalled items.

The possible explanations of the UWP effect covered in the preceding sections assume that people are truly underconfident. As mentioned before, in order to conclude that this underconfidence can be inferred from correspondence between JOLs and recall performance, a frequentist approach to probability needs to be adopted, according to which JOLs should be used to make frequency judgements for classes of items rather than judgements at an item level.

A serious weakness of this assumption is that there are no data to show that this is indeed what participants strive to do when facing the JOL task. In none of the studies investigating the UWP effect were participants given detailed instructions concerning the use of particular scale values. What is more, even the prompts for JOLs that were provided during the study phase were not always consistent with this understanding of the experimental task. For example, in some of the studies investigating the UWP effect (Scheck & Nelson, 2005; Serra & Dunlosky, 2005), participants were asked to rate their *confidence* in future recall. *Confidence*, in contrast to *probability* or *likelihood*, cannot be objectively defined. Therefore, researchers cannot know the way in which participants

use different values on a confidence scale. And even if seemingly more objective terms like *likelihood* are used to prompt for JOLs, it is not known whether participants use the likelihood scale in the way experimenters expect them to do. Given that participants' adherence to the rules (which are never formulated explicitly) is the *sine qua non* condition for inferring the underlying psychological mechanisms from the JOL/recall correspondence, it has to be concluded that the current data do not allow for conceding that it is psychological underconfidence that drives the UWP effect.

2.2. The UWP effect as an artefact

According to this viewpoint, psychological underconfidence is not necessary to produce the pattern of results commonly found in the studies on the UWP effect. In this vein, people may be well aware of their degree of mastery of the to-be-learned material, but still their JOLs may not match their recall performance. To date only two non-psychological mechanisms have been proposed which share the starting point that percentage scales commonly used to elicit JOLs are prone to systematic distortions. Both of these mechanisms will be described in more detail in the following sections.

2.2.1 Anchoring-and-adjustment hypothesis

The anchoring-and-adjustment explanation of the magnitude of JOLs has first been proposed by Hertzog, Saylor, Fleece, and Dixon (1994) to explain different patterns of results of younger and older adults. Hertzog et al., who asked their participants to provide aggregate predictions of future test performance before and after study and after test, observed that older adults were better calibrated than young adults - possibly because their recall performance was closer to the midpoint of the 0-100% scale used in their experiment. They noted, however, that anchoring was not the only process that could influence predictions of performance in their study -

their data showed that these predictions were also influenced by other factors. They proposed that a two-stage process may take place in which participants first set an anchor and then adjust their predictions depending on their memory self-efficacy, memory monitoring or task appraisal.

The idea that anchoring can influence JOLs was further developed by Connor, Dunlosky, and Hertzog (1997) who collected not only aggregate pre- and post-study and post-test predictions, as Hertzog et al., but also item-by-item JOLs in a single-cycle procedure. The results of their Experiment 1 suggested that anchoring may play a role when making JOLs: younger participants' JOLs, which were close to 50%, showed better calibration than did older participants' JOLs. Although the results of Experiment 3 were not clear with regard to anchoring, Connor et al. concluded that the effects of age on calibration depend at least to some extent on the level of recall performance: the age group whose recall is closer to 50% should also be better calibrated.

Finally, Scheck, Meeter, and Nelson (2004) suggested a dual-factor hypothesis to explain the magnitude of immediate JOLs people make during study. According to the dual-factor hypothesis, the magnitude of JOLs derives from two sources. The first source is an anchor which is set based on people's beliefs about how many items they can recall after study. The second source of information that is used when making JOLs, according to this hypothesis, is on-line monitoring of items. As a result, although participants attempt to track their recall with their JOLs (on-line monitoring), their judgements are pulled towards a pre-set anchor, thus failing to achieve good calibration. The results of the experiments of Scheck et al. provided support for the dual-factor hypothesis with regard to immediate JOLs.

The first anchoring explanation of the results obtained in the multi-cycle procedure was proposed by Scheck et al. (2004). The authors suggested that the UWP effect may, at least to some extent, be caused by the fact that people are not familiar with repeated studying and testing. As

the anchoring part of the dual-factor hypothesis assumes that people set an anchor according to their global predictions about future test performance, lack of experience with the multi-cycle procedure leads to the anchor being incorrectly set, causing the UWP pattern to emerge.

Scheck and Nelson (2005) further explored this idea. In their experiment the authors tested two hypotheses. The first one was built on the observation that Koriat et al. (2002), in several experiments, failed to find a way to eliminate the UWP effect. This was dubbed the UWP-effect-is-pervasive hypothesis. According to this hypothesis, the UWP effect should not depend on the level of recall. The alternative anchoring-and-adjustment hypothesis assumed that calibration of JOLs depends on recall performance: if mean recall level is close to the anchor point, people should show good calibration, and if it is below or above the anchor, recall performance and ratings should diverge. Analysing the results of other studies on JOLs, Scheck and Nelson assumed that the placement of the anchor is usually between 30 and 50%.

In their experiment, the authors manipulated the difficulty of the pairs. “Difficult” pairs were defined as those for which a mean percentage of recall after one presentation was 4% and 21% after two presentations, whereas “easy” pairs were defined as those for which recall equalled 21% and 53%, respectively. The aim of this difficulty manipulation was to vary levels of performance so that they would be close to the anchor point in one condition and above this point in the other. If the UWP-effect-is-pervasive explanation was true, no difference should be found between the conditions. The anchoring-and-adjustment hypothesis would, on the other hand, predict different patterns of results: no UWP should be found when recall levels would be close to the anchor point (difficult items), and the UWP effect should be present when recall would be above the anchor (easy items).

The results gave support for the latter hypothesis. Both for immediate and delayed JOLs no UWP pattern was found on cycle 2 for difficult items;

for delayed JOLs, for which performance was below 20%, participants' JOLs overestimated their recall levels, whereas for immediate JOLs, for which performance was at 30%, participants were perfectly calibrated. On the other hand, for easy items, for which performance was above 50%, the UWP pattern emerged both for immediate and delayed JOLs.

This explanation of the UWP effect does not suggest any mechanisms other than anchoring that could influence the magnitude of JOLs in the multi-cycle procedure. At one point Scheck and Nelson (2005) even suggest that their anchoring-and-adjustment hypothesis may possibly be the sole cause of the UWP effect, departing from the dual-factor hypothesis that was proposed for single-cycle JOLs (p. 128).

The only other study that investigated the anchoring hypothesis in the multi-cycle paradigm was conducted by England and Serra (2012) who manipulated perceived task difficulty in their Experiment 2 by incorrectly informing their participants about the difficulty of a future test. In one group, participants were told that the task was considered by other participants to be easy and that mean recall performance was 90%, whereas the other group was told that the task was considered to be difficult and that mean recall performance was 10%. The authors reasoned that this should influence the setting of the anchor, which then should be higher in the easy than in the difficult condition. They reasoned that the setting of an anchor could be estimated by taking the mean JOL for items unrecalled on cycle 1.

The results of this experiment showed that cycle 1 JOLs were higher than recall in the easy, but not in the difficult condition, and that the UWP pattern appeared on cycle 2 in the difficult, but not in the easy condition. The authors also split the items according to whether they had been recalled on cycle 1 and found UWP for recalled items in both conditions, and for unrecalled items only in the difficult condition. They concluded that this pattern is consistent with the predictions of the anchoring hypothesis.

However, as mentioned before, the authors did not give any reason to assume that the anchor point is equivalent to the mean of cycle 1 JOLs for unrecalled items. Other explanations of their results are also viable. For example, it is possible that the fact that participants were (mis)informed about the difficulty of the upcoming test could have influenced the global level of confidence in future recall. This could have distorted the ratings by making participants more convinced that they would be able to recall the items. Therefore, before the results can be taken as a support for the anchoring hypothesis, further research would be needed to rule out the alternative explanations.

The anchoring-and-adjustment account of the UWP effect is the only account that assumes that the sole cause of the UWP effect does not stem from participants' underconfidence in their future performance but is rather an artefact of using percentage scales to elicit judgements. The problem of psychological anchors has been raised in different areas of psychology (cf. Tversky and Kahneman, 1982; Frederick & Mochon, 2012) and there is no reason for which the effects of anchoring should be absent in studies on metacognition. Currently, there is no evidence that would directly contradict the anchoring-and-adjustment hypothesis of the UWP effect: this account can explain most patterns of results found in the literature and helps explain seemingly discrepant results, such as those found by Koriat (1997) for item difficulty. However, some of these results can be equally well explained by other accounts, such as memory for past test (Finn & Metcalfe, 2007, 2008). Moreover, before assuming that the anchoring-and-adjustment explanation can indeed account for the UWP pattern found in the multi-cycle procedure, it has to be noted that this explanation suffers from a serious drawback. The main problem with the current formulation of this explanation of UWP lies in its circularity. Neither Scheck and Nelson (2005), nor Connor et al. (1997) suggested a precise way to estimate the placement of the anchor. In Scheck and Nelson's study, the estimation of anchor location was based on the results of previous studies. In those

studies, however, placement of the anchor was inferred from the pattern of results. The result is that this pattern of results suggests where the anchor is located, which in turn allows for predicting the same pattern of results. A similar criticism can be applied to the claim made by England and Serra (2012), who, as discussed above, assumed that the placement of the anchor was equivalent to the mean of cycle-1 JOLs. In this case the assumed position of the anchor was also inferred from the pattern of participants' ratings. It has to be concluded, therefore, that the present experimental results do not even allow for verifying whether an anchor is set in the JOL task. Unless an independent way of estimating the existence and position of the anchor is found, this explanation, although plausible, needs to be treated with caution.

For these reasons, more research would be needed to establish the usefulness of the anchoring-and-adjustment hypothesis as an explanation of the UWP effect. There is also no reason to assume that anchoring is the sole cause of UWP, as Scheck and Nelson (2005) proposed; the results of some manipulations that influence the magnitude of the UWP effect cannot be explained by anchoring alone (see Experiment 1 in Koriat et al., 2006, for an example). Therefore the initial dual-factor hypothesis proposed by Scheck et al. (2004) seems to be more plausible.

2.2.2 Reduced variability for the highest JOLs

Finn and Metcalfe's (2007, 2008) MPT account proposes a heuristic that people use on cycle 2 and beyond: they base their JOLs on past test performance. At first glance it seems that, according to this explanation, the UWP effect should be found only for the previously unrecalled word pairs, as it is for this class of items that the degree of learning should be undervalued the most. This was the reason that led Koriat et al. (2002), who found the UWP pattern both for recalled and unrecalled items, to reject the past test account of the UWP effect. However, as Finn and Metcalfe noted, MPT is not the only mechanism that may affect the

magnitude of the UWP effect. As almost all items recalled on cycle 1 are again recalled on cycle 2, mean recall performance for these items is usually at ceiling. For this reason, each previously recalled item should be assigned a very high JOL. However, as the JOL scale ends at 100%, if there was any variability in JOLs assigned to previously recalled items, it could only go one way: downward from 100%. As a result, the mean of JOLs would underestimate mean recall performance, revealing the UWP pattern.

This explanation of the UWP pattern cannot be seen as the sole explanation of the UWP effect as it cannot account for UWP found for previously unrecalled items. Rather, it can be seen as complementary to another mechanism - whether stemming from psychological underconfidence or scale distortions.

2.2.3 Summary and conclusions

The two explanations of the UWP effect that see this pattern of results as an artefact note that it is using percentage scales to elicit JOLs that distorts the results. According to Scheck and Nelson (2005), this distortion is caused by judgements being attached to an anchor. Finn and Metcalfe's (2007) restricted variability explanation suggests that the effect for recalled items is caused by the fact that the percentage scale has an upper end which cannot be exceeded.

Scheck and Nelson's (2005) account, if true, could potentially accommodate different results found in the literature. It could account for some of the inconsistent results, such as those for delayed JOLs or pair difficulty. However, as already mentioned, still too little is known about anchoring in the JOL task. Moreover, some of the results that can be explained by anchoring, can also be accounted for by the MPT account.

The restricted variability account, although limited in scope, casts some serious doubt on the assumptions underlying the explanations based on psychological underconfidence that require participants to rate

the probability of correct recall in a strictly set manner. Even if only the highest JOLs on the scale are distorted, then assessing the correspondence between JOLs and recall in a way that is commonly used in the experiments investigating the UWP effect - by comparing a single mean of all JOLs to global recall levels - may sometimes result in detecting an effect that is driven purely by a response scale artefact.

3. Present research

3.1 Rationale

Current theories concerning the mechanisms of the UWP effect cannot provide a definite answer as to what mechanisms lie at the root of this effect. The explanations assuming that it is psychological underconfidence that is reflected in the pattern of results found on cycle 2 and beyond require an assumption that, in the 0-100% JOL task, participants use scale JOLs to indicate the assessed frequencies of correct recall. However, there is currently no evidence to support this assumption. The explanations that assume that the UWP effect is a response scale artefact also lack convincing verification, as discussed above. Current empirical results can be explained by theories belonging to both of these groups, so they cannot offer a solution to this conundrum.

The starting point of this thesis is that the mechanism behind the UWP pattern in the multi-cycle procedure cannot be properly understood without first understanding what information can be extracted from 0-100% JOLs assigned in the multi-cycle procedure. At least three different interpretations of 0-100% JOLs can be proposed. According to the *probability* interpretation, 0-100% JOLs convey information regarding the assessed probability of future recall. In this case, the assessed probability of future recall is supposed to be mapped directly onto the rating scale, so any differences between the magnitude of JOLs and recall performance (i.e. over- or underconfidence) are thought to be stemming from inaccurate predictions. This interpretation of JOLs is implicitly assumed by the proponents of the probability interpretations of UWP, as it is the only one that allows for calculating and interpreting calibration measures.

Although the remaining two interpretations assume that comparing the mean of JOLs to mean recall performance is meaningless, consistent with the artefactual accounts of the UWP effect, they make this assumption for different reasons. The *distorted-rating* interpretation of 0-100% JOLs assumes that even though people aim at rating subjective

probability of future recall in the JOL task, there are factors that can distort the ratings. In other words, people may correctly assess the degree to which the to-be-learned material has been mastered, but for some reason this is not reflected in the ratings. One interpretation of anchoring can serve an example. According to the scale-distortion theory of anchoring (Frederick & Mochon, 2012; Mochon & Frederick, 2013), an anchor may cause a change in the use of the response scale without affecting the internal assessment of to-be-rated stimuli. In the case of the scale-JOL task, a person can be, therefore, absolutely certain that some items will be recalled at test (a state that should warrant a JOL of 100%), but, as the ratings are drawn towards a relatively low anchor (see section 2.2.1 above), the mean of ratings is lower than recall performance for those items.

Finally, the *ranking* interpretation posits that 0-100% JOLs are ordinal ratings of evidence for future recall. According to this interpretation, participants do not aim at all at rating probability in the 0-100% task; instead, they use JOLs merely to convey information about the relative position of items within the study list in terms of that evidence. For example, a person may be equally certain that targets from two pairs will be recalled at test, but decides to assign different ratings to these pairs because one pair was easier to learn than the other (e.g., because it required less repetitions, consists of related words, etc.).

Even though the correct interpretation of the meaning of 0-100% JOLs in the multi-cycle procedure is crucial for understanding the UWP effect, there is little research that tackled that problem. A step in this direction has recently been made by Hanczakowski, Zawadzka, Pasek, and Higham (2013). Their starting point was that there is little evidence that people indeed aim at rating probability in tasks such as the 0-100% JOL task. First, usually the instructions used in the 0-100% JOL task introduce the issue of probability only in the prompt used for eliciting JOLs, which usually starts with “Rate the likelihood...”, and no other instructions

clarifying the issue of probability are provided. This is unfortunate, as calculating calibration requires a specific, frequentist approach to probability, which may not be intuitive to participants. Second, as research in the domain of judgement and decision making shows, people often do not aim at maximising calibration even if it is the main focus of the task that is given to them. For example, Keren and Teigen (2001) conducted a study in which participants were asked to indicate which of two weather forecasters they considered to be better. The first weather forecaster predicted a 75% chance of rain for the upcoming four days, while the second forecaster's prediction was 90%. Participants were then informed that it rained on three days out of four. Even though the first forecaster obtained perfect calibration, while the second revealed overconfidence, participants preferred the second forecaster. This suggests that participants preferred informativeness to perfect calibration even in a calibration rating task: even though the second forecaster's prediction was less accurate in terms of calibration, it was also less ambiguous.

For these reasons, Hanczakowski et al. (2013) investigated whether the UWP effect can be generalised to measures other than scale JOLs. They assumed that if repeating study-test cycles in the UWP task indeed impairs assessments of probability and, in turn, produces psychological underconfidence, the UWP effect should be still found if other measures were used instead of 0-100% scales. To this end, Hanczakowski et al. compared participants' decisions to their recall performance in the multi-cycle procedure. In their experiments, they employed three rating tasks differing only in the response format. Experiment 1 employed typical 0-100% JOLs in a procedure consisting of three study-test cycles and revealed the usual UWP pattern. In Experiments 2 and 3, two types of binary responses were employed instead of the 0-100% ratings: yes/no JOLs and betting decisions. In the binary-JOL task, participants were asked to predict future recall of each target by responding "yes" if they thought they would recall the target at test or "no" if they thought the target

would not be recalled. In the betting task, participants were asked either to bet that they will recall the target or to refrain from betting. They were instructed that for correct bets (target recalled) they would gain a point, for incorrect bets (target not recalled or recalled incorrectly) they would lose a point, and no points would be gained or lost for refraining from betting. Aside from the response format, all other aspects of the procedure were kept intact. To assess calibration, in the binary-JOL task, the proportion of "yes" responses was compared to recall performance. In the betting task, recall performance was compared to the proportion of bets.

What Hanczakowski et al. (2013) found in their Experiments 2 and 3 is that the UWP pattern did not emerge when binary tasks were used. To eliminate a hypothesis that UWP is absent from binary tasks because different information is retrieved when binary questions are asked, the authors asked their participants in Experiment 4 to perform both the 0-100% JOL task and the betting task. For each pair, participants were first asked to provide a 0-100% rating, and then make a bet. However, even in this situation the effect was only found for the former task, while in the latter task participants were well calibrated. In a further analysis, the 0-100% scale was dichotomised: judgements of 50% and above were treated as "yes" JOLs, while judgements lower than 50% were treated as "no" JOLs. These binary JOLs were then compared to recall performance. The results revealed that the binary JOLs derived from the percentage scale were no different than binary bets: no UWP was found for these ratings.

Hanczakowski et al. (2013) argued that their results cast a serious doubt on the probability interpretation of the UWP effect: if it were assessments of probability of future recall that were impaired by practice, binary ratings should be affected as well. Furthermore, the fact that correct assessments of probability of future recall can even be derived from the 0-100% JOL scale when it is dichotomised, suggests that it is the percentage scale that is responsible for the artefactual UWP pattern. For

these reasons, the authors assumed that it may be *confidence* rather than probability that people rate in the 0-100% JOL task. Their interpretation of 0-100% JOLs as confidence ratings was akin to the ranking interpretation described above. The authors noted, however, that assuming the confidence interpretation of UWP does not mean that people do not become truly underconfident with practice. This interpretation of UWP does not speak to the issue of true, psychological confidence; it only states that the 0-100% JOL task is not suitable for assessing calibration for methodological reasons.

In general, the results of Hanczakowski et al. (2013) demonstrate that any conclusions concerning the psychological underconfidence mechanism of UWP are premature, as they depend on untested assumptions concerning the use of percentage scales. It has to be noted, though, that the lack of generalisability of the UWP pattern to binary measures does not rule out entirely the probability interpretation of JOLs. As the authors put it, the fact that measures supposed to reflect the same cognitive processes produce divergent results merely "casts doubt on whether the underlying representation of probability was affected by repeated study-test cycles", which, in turn, makes it "safer to subscribe to the more parsimonious confidence interpretation" of UWP in order to avoid misinterpreting the data, at least until more evidence is gathered (p. 440). At present it is also unknown what produces the dissociations between the results of the scale and binary JOL tasks, which makes it difficult to assess which of these measures (if any) more accurately reflects participants' internal assessments of probability. Therefore the present state of knowledge does not allow for formulating strong conclusions concerning the interpretation of the 0-100% JOL scale.

The aim of the experiments described in this thesis is therefore to further the understanding of the UWP effect by investigating how immediate 0-100% JOLs assigned in the multi-cycle procedure should be interpreted and what mechanisms govern their assignment. The aim of the

experiments described in the first two papers comprising this thesis was to establish which of the three interpretations of 0-100% JOLs - probability, distorted rating, or ranking - applies to JOLs elicited on cycle 2 and beyond. This was achieved by concentrating on judgements assigned to pairs, which were equal in terms of their recall on a subsequent test, but differed in some other aspects. The experiments described in Paper 3 follow up on the previous experiments by investigating a particular mechanism that has the potential to govern the assignment of 0-100% JOLs in the UWP paradigm. Taken together, the findings presented in this thesis are able to resolve the debate whether the UWP effect is a manifestation of psychological underconfidence, or an artefact of a particular measurement scale.

3.2 Experimental overview

This thesis consists of seven experiments forming three separate papers. All experiments used the multi-cycle procedure akin to that used in the research on the UWP effect. On each cycle, participants first studied a list of word pairs. 0-100% JOLs or binary bets were elicited immediately after the presentation of each pair. Each study phase was followed by a cued-recall test, on which participants were presented with cues only, and their task was to type in the target that accompanied that cue at study. The procedure consisted either of two (Paper 2, Experiment 3; Paper 3, Experiment 2) or three (the remaining experiments) study-test cycles. Study materials used in each of the experiments were created from the same list of 60 unrelated word pairs (see Appendix A). Additional pairs used in experiments described in Paper 3 are presented in Appendix B.

3.2.1 Paper 1

This paper consists of two experiments examining the relationship between cycle-3 recall performance, 0-100% JOLs and betting decisions for items previously recalled never (neither on cycle 1 nor on cycle 2),

once (either on cycle 1 or on cycle 2) or twice (both on cycle 1 and cycle 2). Based on the results of previous research (e.g., Finn & Metcalfe, 2007; Hanczakowski et al., 2013), both recall performance and JOLs / betting decisions were predicted to differ between items not recalled before and those previously recalled at least once. Furthermore, as even one successful past recall attempt makes future recall success very likely, recall performance was predicted to be at ceiling for items previously recalled at least once. Therefore, for items previously recalled once and twice, the number of successful recall attempts on the preceding cycles was an invalid cue for probability ratings. The question remained whether cycle-3 0-100% JOLs and betting decisions would be impervious to this cue.

This approach allowed for answering two questions concerning the interpretation of 0-100% JOLs. The main purpose of the study was to shed light on the mechanisms behind the scale / betting dissociation reported by Hanczakowski et al. (2013). First, by splitting items into three distinct categories (never, once, and twice recalled) based on participants' assessments of past recall performance, it was possible to observe whether the good calibration found by Hanczakowski et al. in the betting task at list level would also be present at category level for categories differ in recall performance. If true, that would strengthen the claim that in the binary tasks, participants aim at making assessments of probability of future recall. Second, this design allowed for investigating whether cues irrelevant to future recall performance are incorporated into high JOLs assigned to previously recalled items. As participants should be able to correctly predict that even one past recall success makes the probability of future recall close to 100%, if they incorporated the cue of the number of past recall successes (greater than zero) into their 0-100% JOLs but not into betting decisions, this would strengthen the support of the ranking interpretation of 0-100% JOLs in the UWP paradigm.

To preview, Experiment 1, in which 0-100% JOLs were employed, demonstrated that scale JOLs are determined by the number of previous successful recall attempts, showing a never < once < twice pattern. This was not accompanied by a similar difference in recall performance, as, even though recall was the lowest for items not recalled before, for both types of previously recalled items it was comparable and at ceiling. In Experiment 2, 0-100% JOLs were substituted with binary bets. This time, only the difference between items not recalled on the preceding cycles and items recalled at least once emerged in both measures. Past successful recall failed to exert any effect on the proportion of bets and on recall performance. The results were consistent with the interpretation of the betting decisions as probability assessments, as well as with the ranking interpretation of 0-100% JOLs.

3.2.2 Paper 2

The findings reported in Paper 1 supported the ranking interpretation of 0-100% JOLs by showing that, while making 0-100% JOLs, participants utilise cues that are not predictive of future recall performance on an immediate test. The question remained, however, what kind of cues can influence the assignment of high JOLs without affecting recall performance. The aim of three experiments comprising Paper 2 was to investigate the influence of one particular cue that participants can utilise on cycle 2 and beyond in the 0-100% JOL task for the purpose of ranking the evidence for future recall: the presence of self-generated contextual details (*cognitive context*; e.g., Diana, Yonelinas, & Ranganath, 2012, 2013). In this paper, cognitive context is understood as any information associated with the studied pair, such as thoughts at the time of encoding, images linking the cue with the target, etc. The assumption underlying this set of experiments was that if presentation of items during the study/JOL phase can be accompanied by retrieval of such cognitive context, context retrieval could add to the volume of memorial information available for the

cue-target pair. If participants indeed rank order the pairs in terms of memorial evidence available, as suggested by the previous experiments, then items rich in contextual details should, on average, be assigned higher JOLs than items for which these details are not present or are of lower quality. According to the ranking interpretation, this pattern of JOLs should be found even when recall is equated between these items.

In the first two experiments, the remember/know (R/K) procedure (see e.g., Daniels, Toth, & Hertzog, 2009; McCabe, Roediger, & Karpicke, 2011, for a version of the R/K procedure adapted for use in recall tasks) was employed to gain insight into the presence and number of contextual details available to participants at test. In this procedure, participants are asked to assess for each recalled item whether they "remember" it (i.e., recall of this item is associated with recall of contextual details associated with this item), or if they merely "know" it (i.e., no contextual details are retrieved). In Experiment 1, a standard version of the task was used, in which the only available options were "remember" and "know". In Experiment 2, participants had to clarify their "remember" responses: they were asked to indicate how many contextual details were retrieved for each pair (one, two, three or more). In Experiment 3, the R/K task was substituted with a rating scale ranging from "-" to "+++", on which participants were supposed to rate not only the quantity, but also the quality of the retrieved contextual details.

To preview, the results were consistent with the prediction that contextual details are used as cues for 0-100% JOLs even when their retrieval does not lead to better performance at test. Experiment 1 demonstrated that the mere presence of contextual details (as indicated by "remember" responses) is enough to elevate JOLs above those assigned to items for which contextual details are not available (as indicated by "know" responses). Experiment 2 extended this finding to the number of available contextual details: items for which only one detail was available were assigned lower JOLs than those with two or more accompanying

details. Finally, as Experiment 3 showed, the quality of the contextual details may matter as well for JOL assignment. Crucially, none of the experiments revealed any difference in recall performance between items differing in the quality or quantity (greater than zero) of contextual details. This again supports the ranking interpretation of 0-100% JOLs.

3.2.3 Paper 3

The first two papers concentrated on cues that participants can potentially incorporate into their 0-100% JOLs. The final paper concentrates on the mechanism which may underlie the rating of items in the multi-cycle procedure in terms of evidence for future recall available. This mechanism, called *recalibration*, is also proposed as an explanation for the presence of the UWP pattern in multi-cycle experiments.

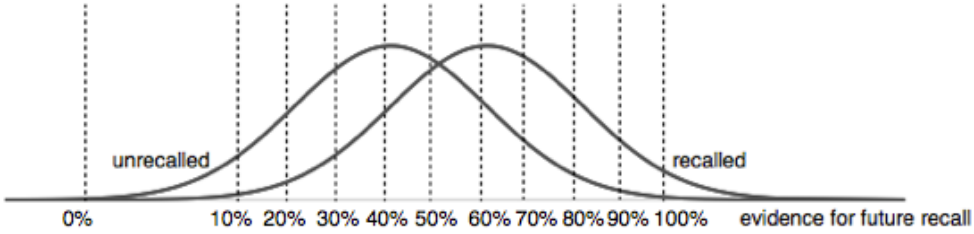

Figure 3.1. A signal-detection model of responding in the JOL task. All studied items are positioned on the evidence-for-future-recall dimension. Items recalled at test have on average stronger evidence for future recall than unrecalled items, hence the recalled items distribution is positioned on the right of the unrecalled items distribution. Vertical lines represent confidence criteria.

The recalibration account has its roots in signal detection theory (SDT). According to SDT, 0-100% JOLs can be thought of as confidence criteria, with a separate criterion for each confidence level (e.g., in increments of 10%: 0%, 10%, 20%, ..., 100%; see Figure 3.1). The values that the criteria represent do not pertain to probabilities of future recall (as per the ranking interpretation of JOLs). As these values do not have any

objective referents, the amount of evidence for future recall that is needed for a rating of, say, 40%, is entirely subjective and can be adjusted according to situational demands. Consequently, the meaning of the 40% JOL can potentially differ between cycles within the multi-cycle procedure, or between experimental groups subjected to different manipulations. It is the latter possibility that is explored in the two experiments presented in Paper 3.

Both experiments were conducted in a between-participants design. In the control groups, participants underwent the typical multi-cycle procedure, with the same items studied on all cycles. In the experimental groups, after the first cycle two-thirds of items comprising the cycle-1 study list were replaced with the same number of new word pairs. In Experiment 1, these new pairs were more difficult than the *critical* pairs (that is, pairs presented on all cycles), while in Experiment 2, the new pairs were easier. This manipulation was supposed to extend the range of evidence for future recall experienced by participants in the experimental conditions. In addition to that, the inclusion of new pairs was supposed to lead to a greater number of items crammed at either the low end (Experiment 1) or the high end (Experiment 2) of the evidence-for-future-recall dimension. It was assumed that this would induce more fine-grained discrimination between items by the means of 0-100% JOLs. Consequently, in Experiment 1, low JOL values were supposed to be reserved for differentiating between difficult new items, while in Experiment 2, high JOL values were supposed to be reserved mostly for the easy new items. This was predicted to influence the assignment of JOLs for the critical pairs in the experimental groups without affecting their recall: the mean of 0-100% JOLs for these pairs should increase when difficult items are present, and decrease when easy items are present. In other words, in Experiment 1 the discrepancy between the mean of 0-100% JOLs and recall performance (traditionally interpreted as the magnitude of the UWP effect)

was predicted to decrease as compared to the control group, while in Experiment 2 this pattern was supposed to be reversed.

To preview, UWP indeed was less pronounced in the experimental than in the control group of Experiment 1, and the reverse was true for Experiment 2. Subsequent analyses revealed that this was caused by differences in the interpretation of JOL values between the groups. In Experiment 1, less evidence was needed for low JOLs in the experimental than in the control group, while in Experiment 2, more evidence was required in order for high JOLs to be assigned. Crucially, the perception of the critical items was not distorted in the experimental group: the inclusion of new pairs did not make critical pairs seem easier (Experiment 1) or more difficult (Experiment 2) than they really were. These results rule out the interpretation of the UWP effect as a result of psychological underconfidence. Instead, a recalibration account of the UWP effect is postulated.

4. Authorship

All papers presented in this thesis have been authored by Katarzyna Zawadzka (first author) and Philip A. Higham (second author). The ideas behind the line of research described in this thesis, and the overall plan of experiments have been formulated collaboratively over several years of supervisory meetings. The first author conducted all experiments and carried out all statistical analyses. The drafts of all manuscripts have been written by the first author, and the final versions incorporate comments and suggestions made by the second author.

5. Paper 1

Judgements of learning index relative confidence not subjective probability

Metacognitive theorists use a variety of different judgements to investigate how people assess their own memory processes. One common one is the judgement of learning (JOL) for which people assess their future memory performance. In a typical experiment employing JOLs, participants study a list of single words or word pairs. After the presentation of each item, a prompt appears instructing participants to rate the likelihood of future recall of that item on a scale from 0% to 100% - the JOL. After the study phase, a recall test for the whole list follows. By comparing JOLs to recall performance, two measures can be calculated. First*, resolution* is the degree to which JOLs distinguish between items that will and will not be recalled at test. In order for resolution to be maximized, later recalled versus later unrecalled items should be assigned high versus low JOLs, respectively. Second, *calibration* is the difference between mean JOLs and mean recall performance. If the two measures are equal, assessments of future recall are said to be *realistic*. Mean JOLs lower versus higher than recall performance indicate *underconfidence* versus *overconfidence*, respectively.

Although most JOL studies reveal overconfidence (e.g., see Koriat, 2012), there are exceptions. For example, the underconfidence-with-practice (UWP) effect (e.g., Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002) is a common finding in JOL research involving repeated study and recall of the same list over at least two cycles. In most UWP studies, recall performance increases with each additional study-test cycle, as does resolution. However, although JOLs are typically similar to recall on the first cycle, they do not increase as much as recall on subsequent cycles, causing calibration to worsen with practice - the UWP effect.

One explanation of the UWP pattern is based on people's memory

for past test performance (Finn & Metcalfe, 2007, 2008; Tauber & Rhodes, 2012). According to the memory-for-past-test (MPT) account of the UWP effect, after cycle 1, people base their immediate JOLs on their performance on the last test. Previously recalled items tend to get high JOLs, as their future recall seems very likely. Conversely, previously unrecalled items are assigned low JOLs, as people remember their failed recall attempt. What people fail to appreciate, though, is that additional learning occurs between the two tests. The additional learning means that some of these previously unrecalled items are recalled on a subsequent test, increasing the discrepancy between mean JOLs and mean recall performance, thus producing UWP.

The MPT account localizes the UWP effect mostly in unduly low judgements assigned to previously unrecalled items. However, Koriat et al. (2002), Finn and Metcalfe (2007), and Hanczakowski, Zawadzka, Pasek, and Higham (2013) reported the presence of the UWP pattern for previously recalled items as well. Finn and Metcalfe argued that these items may contribute to the UWP effect because of variability present in JOLs. Subsequent recall of items that were successfully recalled on a previous cycle(s) is typically excellent and they attract very high JOLs. However, because the JOL scale ends at 100%, any variability is necessarily downward, resulting in mean JOLs that underestimate mean recall performance (i.e., underconfidence). But what produces this downward variability? Finn and Metcalfe remain agnostic of its source. One option is that it may be simply random, not stemming systematically from any characteristics of the rated items. For example, people may be reluctant to use the 100% rating too often, therefore assigning lower ratings to some items even though they believe that they are extremely likely to be later recalled. However, it is equally plausible that the JOL variance for recalled items depends on item-specific information. For example, previously recalled items may have qualitatively different retrieval characteristics, or participants may remember how many times

items were previously recalled and use that that information to guide their JOL assignments. The aim of the present experiments was to examine in more detail the way in which people assign JOLs to previously recalled items.

Experiment 1

In Experiment 1, we employed a common UWP procedure involving three study-test cycles. This methodology allowed us to investigate cycle 3 JOLs for items previously recalled once and twice. Previous research has shown that even a single successful recall attempt makes future recall success extremely likely (e.g., Koriat et al., 2002); therefore, no differences in recall performance between these classes of items were predicted (both should be near ceiling). If it was only the recall success on the preceding cycle(s) that determined scale JOLs for previously recalled items, and if the variability in JOLs for these items did not stem from their characteristics, no difference between items previously recalled on two cycles versus items recalled only on one cycle should be found. If, however, the two item types have different retrieval characteristics, JOLs should differ between these two classes of items on cycle 3. These predictions were tested in Experiment 1.

Method

Participants. Twenty-seven students of the University of Southampton participated in this study for course credit.

Materials and procedure. Sixty pairs of unrelated words were created from a set of 120 English nouns of medium frequency, ranging from four to eight letters in length. The same set of pairs was used for all study-test cycles. All pairs were randomly ordered anew for each participant on each study and test phase.

Before the study phase, participants were provided with instructions describing the JOL task. They were told that their task would be to memorise a set of word pairs for a future cued-recall test. List length was

not mentioned in the instructions. At study, each pair was presented on a computer screen for 1.5 seconds. After the presentation of each pair, the target was replaced by a prompt instructing participants to judge the likelihood of recalling the target from that pair at test when presented with the cue only. Participants were allowed to type in any value between 0% and 100%. This JOL assignment step was self paced.

At test, participants were presented with one cue at a time and asked to type in the target that accompanied this cue during the study phase. If they could not recall the target, they were instructed to press "Continue" to skip to the next cue. To ascertain that people had access to the information about the number of successful recall attempts for each pair, an additional task was implemented at Test 3. When presented with the cue, participants were asked to recall not only the target, but also the number of successful recall attempts for that target on the preceding cycles. The options presented to participants were: 2 (on Tests 1 and 2), 1 (on Test 1 or Test 2), and 0 (on neither Test 1 nor Test 2). This judgement was made for all pairs, independently of whether they were recalled at Test 3 or not.

Before subjecting the results to analysis, participants' recall scores were checked manually. Responses were scored as correct whenever the stem of the word typed in by a participant matched the stem of the target (e.g., *silent* was considered correct if *silence* was the target). Misspelled words (e.g., *slience*) were also counted as correct responses.

Results and Discussion

The means for JOLs, recall, and resolution (A_g) are presented in Table 5.1.[5] Resolution scores were influenced by cycle, as evidenced by a one-way repeated-measures Analysis of Variance (ANOVA), $F(2, 52) =$

[5] A_g is a nonparametric measure of resolution that can be calculated from confidence ratings.

53.912, $MSE = .006$, $p < .001$, $\eta_p^2 = .675$. Resolution increased from cycle 1 to cycle 2, $t(26) = 7.622$, $SE = .018$, $p < .001$, $d = 1.47$, and from cycle 2 to cycle 3, $t(26) = 3.680$, $SE = .021$, $p = .001$, $d = 0.71$.[6]

A 2 (measure: JOL, recall) x 3 (cycle: 1, 2, 3) repeated-measures ANOVA revealed a significant main effect of cycle, $F(2, 52) = 75.206$, $MSE = .014$, $p < .001$, $\eta_p^2 = .743$, showing that, in general, both JOLs and recall performance increased from cycle to cycle. The interaction was also significant, $F(2, 52) = 62.918$, $MSE = .007$, $p < .001$, $\eta_p^2 = .708$. Whereas on cycle 1, participants' mean JOLs were higher than their recall performance, $t(26) = 5.069$, $SE = .042$, $p < .001$, $d = 0.98$, this pattern was reversed on cycles 2 and 3, $t(26) = 3.131$, $SE = .039$, $p = .004$, $d = 0.61$ and $t(26) = 2.647$, $SE = .033$, $p = .014$, $d = 0.48$, respectively, revealing the UWP effect. The main effect of measure was not significant, $F < 1$.

Mean recall and mean JOLs for items assessed as being recalled never, once, or twice, are presented in the top panel of Figure 5.1. Also shown in the figure are analogous means for actual (rather than assessed) recall performance. On average, at Test 3, participants were able to correctly recall the number of previous successful recall attempts in 90% of cases. The mean gamma correlation computed for each participant between assessed and actual number of successful recall attempts was .986, confirming that participants were highly accurate in their assessments.

[6] No corrections for multiple comparisons were made for the analyses reported in this thesis.

Table 5.1

Means (SDs) for Recall Performance, JOLs and Proportion of Bets, and Resolution (A_g for JOLs and d' for Bets) as a Function of Cycle in Experiments 1 and 2.

	Measure				
Experiment and Cycle	Recall	JOL	Betting	A_g	*d'*
Experiment 1					
cycle 1	.22 (.15)	.43 (.19)	-	.61 (.12)	-
cycle 2	.51 (.24)	.39 (.20)	-	.75 (.08)	-
cycle 3	.65 (.23)	.56 (.20)	-	.83 (.12)	-
Experiment 2					
cycle 1	.29 (.22)	-	.43 (.18)	-	1.10 (0.70)
cycle 2	.56 (.24)	-	.52 (.21)	-	1.84 (0.71)
cycle 3	.70 (.22)	-	.69 (.21)	-	2.11 (0.57)

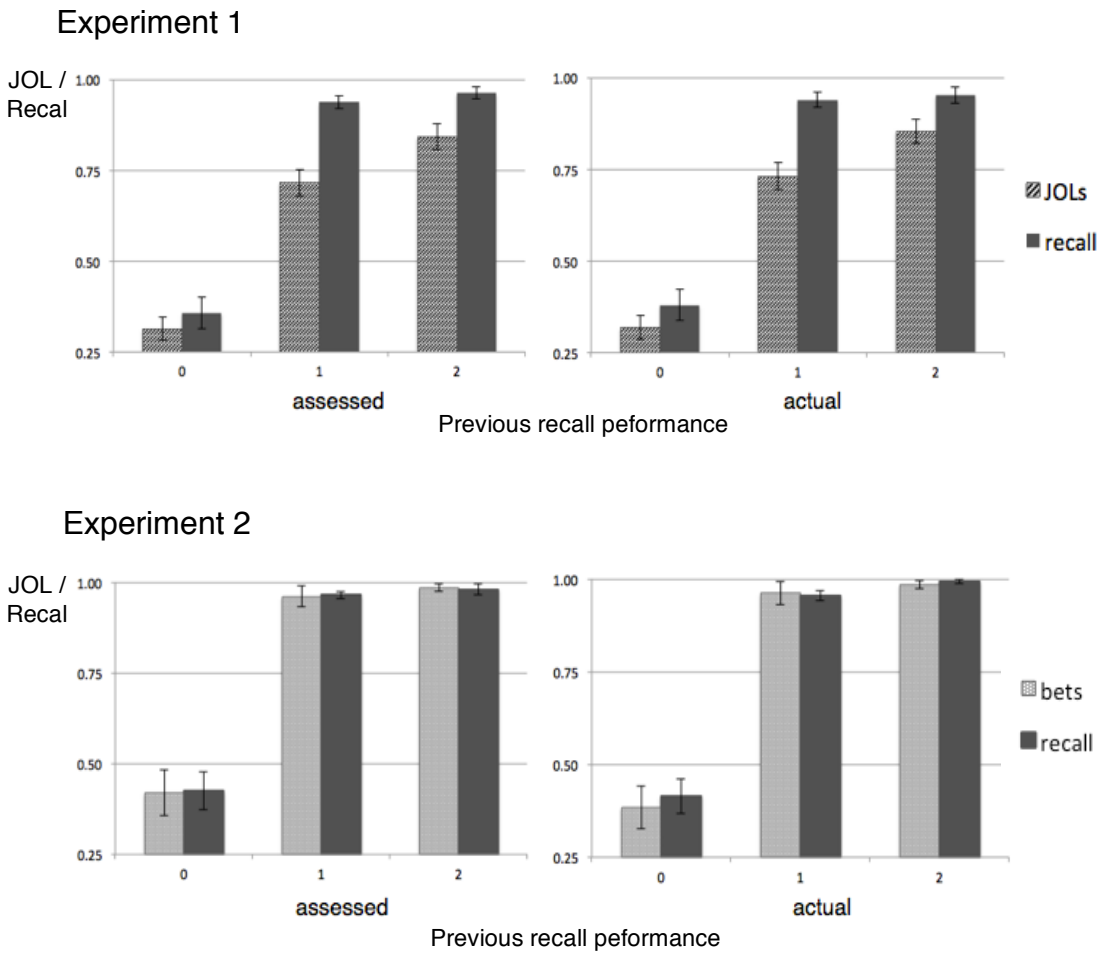

Figure 5.1. Mean JOLs (Experiment 1), proportion of bets (Experiment 2) and recall performance (both experiments) on cycle 3 as a function of the assessed (left panel) and actual (right panel) number of previous recall successes on cycles 1 and 2. Error bars indicate standard error of the mean.

Finally, mean JOLs and mean recall performance on cycle 3 for items assessed as being previously recalled never, once or twice (see the top left panel of Figure 5.1) were analysed with two separate one-way ANOVAs.[7] The analyses were conducted on the data set based on participants' assessments of their past recall performance (rather than

[7] We decided against performing a 2 (judgement: JOL, recall) x 3 (previous recall performance: never, once, twice) ANOVA on these data to avoid getting a spurious interaction caused by ceiling performance for previously recalled items.

their actual past recall performance) as it is this information that participants can access and, potentially, make use of when assigning JOLs. For JOLs, the number of successful recall attempts was significant, $F(2, 52) = 113.520$, $MSE = .018$, $p < .001$, $\eta_p^2 = .814$. Mean JOLs differed between items judged as being recalled never versus once, $t(26) = 10.141$, $SE = .039$, $p < .001$, $d = 1.96$, never versus twice, $t(26) = 11.972$, $SE = .044$, $p < .001$, $d = 2.31$, as well as once versus twice, $t(26) = 5.656$, $SE = .022$, $p < .001$, $d = 1.09$.[8] For recall performance, the effect of the number of successful recall attempts was significant as well, $F(2, 52) = 169.478$, $MSE = .019$, $p < .001$, $\eta_p^2 = .867$. As in the case of JOLs, recall differed between items judged as being recalled never versus once, $t(26) = 14.622$, $SE = .040$, $p < .001$, $d = 3.48$, and never versus twice, $t(26) = 13.050$, $SE = .046$, $p < .001$, $d = 2.78$. Critically, however, there was no difference in recall performance for items judged as being recalled once versus twice, $t(26) = 1.294$, $SE = .020$, $p = .21$, $d = 0.24$.

In this study, we intentionally decided to concentrate on cases in which recall performance was at ceiling, as this very high performance should be easy for participants to predict. However, a critic could argue that the ceiling effect makes it impossible to correctly assess the impact of the number of past recalls on JOLs. Had the test been more difficult, perhaps recall performance would have mirrored the JOL pattern rather

---

8 One potential reason for the difference in JOLs between items judged as being recalled once and twice is that lower JOLs for the former class of items are caused by the presence of items recalled on cycle 1, but forgotten on cycle 2. In such a case, failure to recall an item on the immediately preceding cycle may, in theory, make participants treat such items as previously unrecalled and, in turn, make them assign unduly low JOLs. To test this explanation, we split items judged as previously recalled once into two groups: recalled only on cycle 1 and recalled only on cycle 2. Only four participants recalled at least one item only on cycle 1, and the mean of JOLs for these items was numerically *higher* ($M = .83$, $SD = .05$) rather than lower than the mean for all items recalled once ($M = .71$, $SD = .19$), ruling out this hypothesis.

than dissociating from it. To address this concern, we performed an additional analysis analogous to one reported by Wixted and Mickes (2010). Specifically, we divided participants into three subgroups according to their pattern of recall performance on cycle 3 (see Table 5.2). The first group had higher recall on cycle 3 for items assessed as previously recalled twice than once (once<twice), the second group exhibited the opposite pattern (once>twice), while the third group performed equally well for both classes of items (once=twice). We then compared cycle-3 JOLs and recall in each subgroup using three separate 2 (measure: recall, JOL) x 2 (judged previous recall: once, twice) ANOVAs.

Table 5.2

Means (SDs) for Cycle 3 Recall Performance in Experiments 1 and 2, JOLs in Experiment 1, and Proportion of Bets in Experiment 2, for Items Judged as Previously Recalled Once and Twice as a Function of Cycle-3 Recall Pattern.

	Recall		JOL		Betting	
Experiment and Cycle-3 Recall Pattern	once	twice	once	twice	once	twice
Experiment 1						
once<twice (*n* = 10)	.87 (.09)	1.00 (.00)	.71 (.23)	.85 (.23)	-	-
once=twice (*n* = 11)	1.00 (.00)	1.00 (.00)	.75 (.14)	.89 (.11)	-	-
once>twice (*n* = 6)	.93 (.11)	.84 (.14)	.66 (.19)	.75 (.18)	-	-
Experiment 2						
once<twice (*n* = 8)	.92 (.03)	.99 (.03)	-	-	.91 (.22)	.96 (.08)
once=twice (*n* = 12)	1.00 (.00)	1.00 (.00)	-	-	.99 (.03)	1.00 (.00)
once>twice (*n* = 2)	.94 (.08)	.82 (.21)	-	-	1.00 (.00)	1.00 (.00)

If the observed dissociation was actually due to a ceiling effect on recall which was masking a positive relationship between recall and JOLs as the number of assessed previous recalls increased, then the JOL advantage for twice versus once recalled items observed in the full dataset should be present in the once<twice subgroup, but attenuated or reversed in the once=twice and once>twice subgroups. However, this pattern was not observed. For the once<twice subgroup, main effects of measure, $F(1, 9) = 5.855$, $MSE = .043$, $p = .039$, $\eta_p^2 = .394$, and previous recall, $F(1, 9) = 32.080$, $MSE = .005$, $p < .001$, $\eta_p^2 = .781$, were significant. The interaction was not significant, $F < 1$. Conversely, for the once>twice subgroup, there was no main effect of previous recall, $F < 1$, but there was a significant main effect of measure, $F(1, 5) = 8.060$, $MSE = .024$, $p = .036$, $\eta_p^2 = .617$, which was qualified by a significant interaction, $F(1, 5) = 13.514$, $MSE = .004$, $p = .014$, $\eta_p^2 = .730$. The interaction was significant as well in the once=twice subgroup, $F(1, 10) = 16.483$, $MSE = .003$, $p = .002$, $\eta_p^2 = .622$, as were the main effects of measure, $F(1, 10) = 26.213$, $MSE = .013$, $p < .001$, $\eta_p^2 = .724$, and recall, $F(1, 10) = 16.483$, $MSE = .003$, $p = .002$, $\eta_p^2 = .622$. In total, these results confirm that JOLs were higher for twice- versus once-recalled items regardless of the recall pattern; even participants who recalled once-recalled items better than twice-recalled items, the pattern for JOLs was still twice>once. This pattern of results suggests that there was a true dissociation between recall and scale JOLs in the complete dataset rather than one that was produced by a ceiling effect on recall.

The results of Experiment 1 replicated two effects already present in the literature. For the aggregate data, the UWP effect was present on cycles 2 and 3. The differences in mean JOLs between items judged as unrecalled and recalled at least once were also found, consistent with the MPT account. However, on cycle 3 we also found a 13% difference in mean JOLs between items judged as being previously recalled once

versus twice. Interestingly, this difference was not accompanied by a difference in recall performance, producing a dissociation. As both kinds of items had been successfully recalled at least once on cycles 1 and 2, if it was only the successful recall on at least one of the preceding cycles that determined the highest JOLs in a multi-cycle procedure, no difference in JOLs should be found.

What, then, could have lead to such a pattern of results? One possibility is that it stems from people's predictions regarding the differences in the amount of forgetting from one cycle to another. The present results show that people remember very well the number of successful recall attempts for each item. It is thus possible that, for items previously recalled once, participants remember a failed recall attempt and therefore assumed that they may fail to recall that item again. For items previously recalled twice, on the other hand, forgetting the target at Test 3 can be seen as less likely. As recall performance does not differ between these two classes of items, it would therefore be the overestimation of forgetting for items recalled once that would be responsible for the present result. However, if this were true, this finding would stand in opposition to the extant literature on the metacognition of forgetting, which shows that people tend to underestimate rather than overestimate the amount of forgetting they will experience (e.g., Koriat, Bjork, Sheffer, & Bar, 2004; Kornell & Bjork, 2009; Serra & England, 2012; although see Ariel, Hines, & Hertzog, 2014).

The aim of Experiment 2 was to test the forgetting overestimation hypothesis. To this end, the scale JOL task was substituted with a binary betting task. Hanczakowski et al. (2013) have recently demonstrated that binary tasks can be used to assess the accuracy of people's predictions concerning their future memory performance. We predicted that if it was the overestimation of forgetting that unduly lowered scale JOLs for once-recalled items in Experiment 1, the same pattern should be found in the betting task in Experiment 2 (i.e., the proportion of bets would be

inappropriately low for once-recalled items). If, however, no difference was found between the proportions of bets for items judged as previously recalled once versus twice, this would mean that people can precisely assess their future recall performance and the forgetting overestimation explanation would be eliminated.

Experiment 2

Method

Participants. Twenty-two students of the University of Southampton participated in this study for course credit.

Materials and procedure. The materials and procedure were the same as in Experiment 1 with one exception: instead of providing scale JOLs, participants were given a binary-betting task. Good calibration on the binary-betting task would be obtained if the proportion of bets equaled the proportion of recalled items. Hanczakowski et al. (2013) have recently demonstrated that the binary-betting task produces the same results as the binary ("yes/no") JOL task, without suffering from a potentially serious drawback that characterizes the latter task. In the binary JOL task, people may assign different subjective values to two types of incorrect answers: metacognitive misses (correctly recalling an item assigned a "no" JOL) and metacognitive false alarms (failing to recall an item assigned a "yes" JOL). This introduces a source of potential bias to the measure: participants may value misses and false alarms differently in the yes/no task. This may make them more inclined to use less often the response that is associated with a higher aversive value (e.g., if participants prefer making a miss to a false alarm, they could be biased toward saying "no" more often than it would follow from their internal assessments of future memory performance). In the betting task, on the other hand, penalties and rewards for different types of answers are objectively defined by the experimenter and equated, minimizing the possibility of biased responding.

For each pair in the betting task, participants were asked whether

they would like to bet they would later recall the target from that pair when presented with the cue. They were instructed that for correct bets they would gain a point, whereas for incorrect bets they would lose a point. If they refrained from betting, no points would be gained or lost. Participants were not shown their point count during the experiment.

Results and Discussion

The means for the proportion of bets, recall performance, and resolution (*d'*, a signal-detection measure of discrimination calculated from binary data) are presented in Table 5.1. Again, resolution was influenced by cycle, $F(2, 42) = 20.939$, $MSE = .284$, $p < .001$, $\eta_p^2 = .499$, and increased from cycle 1 to cycle 2, $t(21) = 4.418$, $SEM = .167$, $p < .001$, $d = 0.74$, and from cycle 2 to cycle 3, $t(21) = 2.123$, $SEM = .126$, $p = .046$, $d = 0.27$.

A 2 (measure: proportion of bets, recall performance) x 3 (cycle: 1, 2, 3) repeated-measures ANOVA revealed a significant main effect of cycle, $F(2, 42) = 123.270$, $MSE = .010$, $p < .001$, $\eta_p^2 = .854$, indicating an increase from cycle to cycle both in the proportion of bets and recall performance. The interaction was also significant, $F(2, 42) = 22.888$, $MSE = .004$, $p < .001$, $\eta_p^2 = .522$. Whereas on cycle 1, participants bet on a greater proportion of items than they later recalled, $t(21) = 4.084$, $SEM = .033$, $p = .001$, $d = 0.91$, there was no difference between these measures on cycles 2 and 3, $t(21) = 1.239$, $SEM = .033$, $p = .23$, $d = 0.26$ and $t < 1$, respectively. The main effect of measure was not significant, $F(2, 42) = 1.153$, $MSE = .0122$, $p = .29$, $\eta_p^2 = .052$. These results replicate the findings of Hanczakowski et al. (2013), again demonstrating that participants can accurately track future recall performance with their betting decisions.

Participants' assessments of the number of successful recall attempts were as good as they were with scale-JOLs (see the bottom panel of Figure 5.1). At Test 3, in 91% of cases participants were able to

correctly assess the number of previously successful recall attempts for each pair, producing gamma = .977.

The crucial comparisons concern the proportion of bets and recall performance on cycle 3 for items judged as being previously recalled never, once and twice (presented in the bottom left panel of Figure 5.1). We again performed two separate one-way ANOVAs on the proportion of bets and recall. For bets, there was a significant effect of number of successful recall attempts, $F(2, 40) = 82.892$, $MSE = .028$, $p < .001$, $\eta_p^2 = .806$. The proportion of bets differed significantly between items judged as being recalled never versus once, $t(20) = 9.108$, $SE = .063$, $p < .001$, $d = 1.55$, and never versus twice, $t(20) = 9.173$, $SE = .062$, $p < .001$, $d = 2.92$, but, importantly, not between items judged as being recalled once versus twice, $t(21) = 1.260$, $SE = .019$, $p = .22$, $d = 0.56$.

For recall performance, the effect of the number of successful recall attempts was significant as well, $F(2, 40) = 113.251$, $MSE = .019$, $p < .001$, $\eta_p^2 = .850$. This time, recall performance behaved in the same manner as bets, differing only between items judged as being recalled never versus once, $t(20) = 10.898$, $SE = .050$, $p < .001$, $d = 3.24$, and never versus twice, $t(20) = 10.754$, $SE = .052$, $p < .001$, $d = 2.78$, but not between items judged as being recalled once versus twice, $t(21) = 1.11$, $SE = .014$, $p = .28$, $d = 0.27$. This suggests that participants were able to predict the very high probability of successful recall for items previously recalled at least once, as well as the lack of difference in future recall performance for these items, and executed their betting decisions accordingly.

As in Experiment 1, we divided participants into subgroups based on their recall performance for the once/twice cue (see Table 5.2). In the once<twice and once=twice subgroups, bets closely tracked recall performance, but there was no evidence that betting was sensitive to the once/twice recall cue. This result was confirmed by two 2 (measure: recall, proportion of bets) x 2 (judged previous recall: once, twice) ANOVAs. For

the once>twice subgroup, the ANOVA revealed only a marginally significant main effect of previous recall, $F(1, 7) = 4.508$, $MSE = .007$, $p = .071$, $\eta_p^2 = .392$. The main effect of measure and, most critically, the interaction were not significant, both Fs < 1. For the once=twice subgroup, neither of the main effects, nor the interaction was significant, all Fs = 1.[9]

In total, the results suggest that participants in the betting task were able to predict the very high probability of successful recall for items previously recalled at least once, as well as the lack of difference in future recall performance for these items, and executed their betting decisions accordingly. This result rules out the hypothesis that the difference between scale JOLs for items judged as being recalled once versus twice stems from an overestimation of forgetting for the former group of items.

General Discussion

The present study investigated the way in which JOLs are assigned in a multi-cycle procedure. In particular, we focused on JOL assignment to items successfully recalled on a previous test. Experiment 1 demonstrated that the assignment of JOLs to these previously recalled items depended on the number of previous successful recall attempts, whereas recall performance did not. Conversely, in Experiment 2, betting decisions tracked recall performance almost perfectly, suggesting that people are able to predict their future recall performance with very high accuracy if given the appropriate task. The results of Experiment 2 demonstrate that the difference between JOLs assigned to items previously recalled once versus twice that we observed in the scale task cannot be attributed to an inability to make correct predictions concerning future recall performance for these items. In general, the results reveal a dissociation between scale JOLs and binary judgements that presumably are meant to be tapping the

[9] No analysis was performed on the once>twice data in this experiment, as only two participants displayed this pattern.

same underlying representations.

Two potential explanations of these results can be postulated. According to the first one, participants in the scale JOL task, but not in the betting task, incorporate *invalid cues* into their probability ratings. By an invalid cue, we mean a cue that, when incorporated into a metacognitive judgement, reduces the metacognitive accuracy of this judgement. This can happen either because the cue is unrelated to recall performance, or because it actually predicts an opposite effect on recall. Note, however, that the validity of a cue strongly depends on the experimental task. For example, some cues may be valid when the test requires recall, but not recognition, or when a test is immediate rather than delayed. In the present case, we define an invalid cue as one that does not predict the likelihood of recalling an item on a cued-recall test administered immediately after the study phase, as it is performance on this type of test that participants' judgements are meant to be predicting.

As our results show, recall performance on the last test did not differ for items previously recalled once and twice. Therefore, from the perspective of the scale-JOL and binary-betting tasks, the assessed number of successful recall attempts was an invalid cue. Yet participants apparently incorporated this cue into their scale-JOL ratings (but not their binary-betting decisions), lowering their predictive value. Thus, it could be assumed that participants in the scale-JOL task were misled by their memory for performance on past tests, which made them perform the task suboptimally.

However, we have also shown that participants are able to correctly predict future recall for items previously recalled once and twice when given the binary-betting task. This speaks against the scale-JOL difference for once versus twice recalled items being based on a true psychological difference in subjective probability: if participants believed that items previously recalled once are indeed less likely to be recalled again than items previously recalled twice, this would be evident both in the betting

and scale-JOL tasks. Therefore we believe that another explanation, one based on the recent findings of Hanczakowski et al. (2013), is more viable: namely, that participants' scale JOLs may not be assessments of recall probability, but rather represent confidence judgements (see also Higham, Zawadzka & Hanczakowski, in press).

The difference between *probability* and *confidence* ratings has profound consequences for calibration research for which it is common practice to directly compare mean JOLs and mean recall. For this comparison to be meaningful for the assessments of realism of JOLs, there is an assumption that intervals on the scale on which ratings are made are comparable to the intervals on the underlying psychological dimension that the scale values are meant to index. However, confidence scales are likely only ordinal; that is, JOLs may simply represent a rank ordering of the recallability of items, not recall probabilities. For the latter, participants must ensure that the psychological distance between 70% and 80% is the same as that between 20% and 30% (or any other pairs of values that differ by 10%), which seems unlikely (e.g., see Poulton, 1979). That being the case, direct comparisons between mean of scale JOLs and mean recall provides little to no information about the realism of people's judgements.

The present findings confirm and extend those of Hanczakowski et al. (2013). First, they replicate their main finding that binary-betting and scale-JOL tasks give rise to different results, even though they are supposed to measure the same underlying construct: probability of future recall. Second, by concentrating on a particular cue unrelated to immediate recall performance, our findings directly demonstrate different bases of binary and scale ratings. Betting decisions turn out to be impervious to the number of successful recall attempts, suggesting that participants are aware of the lack of effect of this cue on future memory performance. Yet this does not prevent incorporating this cue into scale JOLs. In total, our results strengthen the conclusion that scale JOLs, as

compared to betting decisions, do not measure subjective probability, but rather are ratings of confidence in future recall.

How might an account that considers scale-JOLs to be confidence ratings rather than probability judgements account for our results? According to the confidence account of scale JOLs, even though participants in Experiment 1 were aware that the *probability* of recalling a previously recalled item is similar for the two classes of items and close to 100%, they may have assumed that each additional successful recall attempt warrants an increase in *confidence* in future recall. Confidence ratings are not aimed at providing numerical assessments of probability. Therefore it is not incorrect to use different values on a confidence scale to refer to items sharing the same probability of being recalled, but differing in some other aspects. In other words, even though the number of successful recall attempts on the preceding cycles is an invalid cue for ratings of probability of future recall, it can serve as a valid cue for confidence ratings.

If scale JOLs are in fact confidence ratings, subjective probability of future recall is only one of the cues that might be used for making a JOL and other cues may be used to discriminate amongst items that participants nonetheless believe will be later recalled successfully. Even if the subjective probability of later recall is equal for different subsets of items, and, as in our study, is close to 100%, participants may still base their JOLs on cues that allow them to demonstrate that they are aware of differences amongst the subsets of highly recallable items - such as memory for recall performance on the preceding tests.

What information can, therefore, serve as a basis for high JOLs in our study? It may be that it is simply memory for past-test performance. Ratings for items previously recalled once versus twice may have differed because participants wanted to distinguish between these classes of items using their JOLs. In such a case it would be the number of successful recall attempts that would serve as a cue for high JOLs. Another option is

that the two classes of items also differ on some other dimension apart from the number of successful recall attempts. Items recalled on two cycles are likely to be subjectively easier to learn and later recall than items recalled only on cycle 2, which could as well lead to higher JOLs for the former class of items. It has to be noted as well that the number of successful retrievals can be predictive of future recall under certain circumstances. For example, the more times an item is recalled, the better the memory for that item after a delay (e.g., Vaughn & Rawson, 2011), and scale JOLs are known to be sensitive to this cue (Pyc & Rawson, 2012). Even though participants are aware that their task is to predict recall on an immediate test, rendering long-term predictions of retention irrelevant for the task they face, the number of previous successful recalls may be incorporated into ratings to demonstrate that items previously recalled once and twice differ in terms of memorial evidence. In any case, the quality and/or quantity of evidence for future recall differs between items recalled once versus twice, leading to different scale ratings in spite of equated recall performance.

6. Paper 2

Cognitive context drives judgements of learning

Judgements of learning (JOLs) are predictions of future memory performance elicited after a study trial. In a typical experiment employing JOLs, participants are presented with a study list, and after each item on that list they are asked to rate the likelihood of recalling this particular item at test (so called *immediate* JOLs). After that, a memory test follows. The comparison of JOLs and recall performance allows two measures to be calculated. *Resolution* is the degree to which JOLs distinguish between subsequently recalled and unrecalled items. *Calibration* assesses the difference between the mean of JOLs and recall performance. If the mean of JOLs and recall performance are equal, people are thought to be *well calibrated* or *realistic* in their assessments. If JOLs exceed recall, this reveals *overconfidence*, whereas if they underestimate recall, *underconfidence* is revealed.

The underconfidence-with-practice effect (UWP, cf. Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002) is a finding from calibration studies in which people study and are tested on the same list in at least two study-test cycles. The usual pattern of results is that of overconfidence or good calibration on cycle 1, but underconfidence on cycle 2 and beyond. In other words, with practice, calibration becomes impaired - hence "underconfidence with practice". Impaired calibration stands in contrast to a general improvement of performance which occurs from one cycle to another: both memory for studied items and resolution increase with practice.

Several theories have been put forward to explain the UWP effect, the most popular of which are the anchoring (e.g., England & Serra, 2012; Scheck & Nelson, 2005) and memory-for-past-test accounts (e.g., Finn & Metcalfe, 2007, 2008). However, in contrast to these psychological

theories, Hanczakowski, Zawadzka, Pasek, and Higham (2013) recently proposed that the UWP effect may be an artefact caused by an incorrect interpretation of the results of the experimental task. In four experiments, the authors demonstrated that the UWP effect is found only when JOLs are made on a scale from 0% to 100%. When binary tasks (a yes/no JOL task and a betting task) were employed, the effect disappeared, even though responses in these binary tasks were supposed to reflect the same assessments of future recall. On the basis of these discrepant findings for scales and binary tasks, Hanczakowski et al. concluded that it was unlikely that participants in the 0-100% JOL task were rating *probability* of future recall. Instead, participants in the scale-JOL task likely rate their *confidence* in future recall. The difference between these terms does not lie only on the semantic level. In order for the mean of JOLs to be meaningfully interpreted, the scale on which these judgements should be at least interval - an assumption that is met by the probability interpretation. Confidence, on the other hand, may well be measured on an ordinal scale. If it were confidence that people rated in the JOL task, calculating calibration would be incorrect from a methodological point of view.

The results of Hanczakowski et al. (2013) suggested that JOLs are not pure ratings of probability. This prediction was tested in the experiments described in Paper 1, which investigated how people assign 0-100% JOLs to items recalled on the preceding cycles. The results revealed that, on cycle 3, participants assigned lower JOLs to items they rated as having been recalled only once on the preceding cycles (that is, either on cycle 1 or on cycle 2) than to items rated as having been recalled twice (both on cycle 1 and cycle 2). Crucially, the probability of recalling these classes of items was equal and at ceiling, exceeding 90%. This difference in JOLs did not stem from an inaccurate estimation of probability of future recall: when given a binary betting task, participants were able to assess the probability of recalling the target with very high

accuracy (see also Hanczakowski et al., 2013). Thus, as the information about the probability of future recall was available at the time of the JOL assignment, participants must have incorporated information into their JOLs other than the mere likelihood of recall.

The aim of the present study is to test another potential source of variability in high JOLs - namely, the quality of memorial information that accompanies processing of the cue-target pair during the experimental procedure. One group of factors that can affect the perceived quality of retrieved information are contextual cues. These cues accompany retrieval of an event, providing episodic details about encoding or previous retrieval attempts. It is conceivable that memorised items may differ with regard to the quality and quantity of contextual information that was encoded together with the items. These qualitative and quantitative differences may potentially introduce variability to JOLs made for recalled items. As a result, the highest JOLs would be reserved only for items with the strongest accompanying evidence, lowering the JOLs for items for which this evidence is weaker or even absent. This, in turn, could contribute to the UWP effect.

One type of context that can be particularly relevant here is *cognitive context* (e.g., Diana, Yonelinas, & Ranganath, 2012, 2013). Any internally produced contextual information can be considered cognitive context: for example, moods, reactions or thoughts generated at the time of study. Suppose participants are presented at study with a pair *nurse-museum*. One person can imagine a nurse visiting a museum, thus linking the two words from the pair, while another can think at the same time that her mother is a nurse. These internally generated contextual details can later be retrieved at test.

To date, two studies have investigated the relationship between internally generated contextual details and the magnitude of JOLs. Daniels, Toth, and Hertzog (2009) used a cued-recall task for this purpose. The experiment conducted by Daniels et al. consisted of a single study-

test cycle. At study, the presentation of each item was followed by an immediate JOL. At test, an additional task was implemented to investigate the presence of contextual details associated with the retrieved items. This additional task was based on a distinction taken from dual-process models of memory (see Yonelinas, 2002, for a review), in which two distinct types of retrieval can be identified. *Recollection* occurs when retrieval of an event is accompanied by contextual details such as thoughts at the time of encoding or perceptual details; if such episodic details do not accompany the act of retrieval, the event is considered to be *familiar*. To compare JOLs assigned to items accompanied at test by contextual details, and those for which no context is present, Daniels et al. asked their participants to indicate for each retrieved item whether it was "recollected" or "familiar", or if they had "no memory" for that item. The authors then compared mean immediate JOLs assigned to all item classes. Their results suggested that JOLs elicited during the study phase were higher for items that were later assigned a "recollect" judgement than for those rated as "familiar", which were, in turn higher than those for which a "no memory" option was chosen.

In a similar vein, Skavhaug, Wilding, and Donaldson (2013) examined the so-called electrophysiological correlates of familiarity (the mid-frontal effect) and recollection (the left-parietal effect) in a JOL task. At study, participants were presented with a list of cue-target pairs and assigned to each of the pairs an immediate JOL. At test, they were given an old/new recognition task for cues only. Electrophysiological data were collected at test, and the magnitudes of the mid-frontal and the left-parietal effects were separately assessed for items which were earlier assigned high and low JOLs. Skavhaug et al. found that although for both items assigned high and low JOLs the mid-frontal and left-parietal effects were present, suggesting that familiarity and recollection occurred at test, the left-parietal effect was stronger for items assigned high than low JOLs.

The studies by Daniels et al. (2009) and Skavhaug et al. (2013) suggest that, at least in single-cycle procedures, high JOLs are assigned to items for which it is later possible to retrieve contextual details at the time of the test. Two questions arise from these findings. First, does this pattern of results generalise to the multi-cycle procedure? Second, if the answer to the first question is positive, does cognitive context contribute to the magnitude of the UWP effect? If high JOLs were reserved for items for which contextual details were available, it could unduly lower JOLs assigned to items not accompanied by cognitive context. This could, in turn, lower the mean of JOLs assigned at study, increasing the discrepancy between JOLs and recall performance. In order to answer these two questions, we investigated the influence of contextual details on the magnitude of JOLs using the UWP procedure consisting of multiple study-test cycles.

To investigate the role of context in memory tasks, remember/know judgements are often elicited (cf. Donaldson, 1996; Tulving, 1985). In this task, *remember* judgements are theorized to be based on recollective experience, whereas *know* judgements are thought to be reflections of familiarity. Although this task is commonly used to investigate recognition memory (cf. Gardiner, 1988; Rajaram, 1993), it can also be used on recall tests (e.g., McCabe, Roediger, & Karpicke, 2011; Tulving, 1985).

The remember/know task seems suitable for examining the possible influence that contextual cues can exert on the magnitude of JOLs. The present study concentrated on a general class of contextual cues, without specifying their type. As recollection of any contextual cue for a given item should be reflected in remember judgements, comparing JOLs assigned to remember and know responses should allow for testing the prediction that the presence of contextual cues can affect the magnitude of JOLs in the UWP procedure.

Experiment 1

Method

Participants. The participants were 21 undergraduate students at the University of Southampton who participated in this experiment in exchange for course credits. The data of one participant were excluded from the analyses as this person did not understand correctly the difference between remember and know judgements (see below).

Materials and Procedure. Sixty pairs of unrelated words were created from a set of 120 English nouns chosen from the MRC database, of medium frequency and ranging from four to eight letters in length. The procedure consisted of three study-test cycles, and the same set of pairs was used for all cycles. All pairs were randomly ordered anew for each participant and on each cycle.

During the study phase participants were presented with all 60 pairs and instructed to memorise them for a future test. Each pair was presented on a computer screen for 1.5 seconds. After the presentation of each pair, the target disappeared from the screen and a prompt appeared instructing participants to judge the likelihood of recalling the target from that pair at test when presented with the cue only. Participants were allowed to type in any value between 0 and 100%. At test participants were presented with one cue at a time and asked to type in the target that accompanied this cue during the study phase. If they could not recall the target, they were instructed to press “Continue” to skip to the next cue. Both the JOL task and the test were self-paced.

An additional task was administered for each recalled item at Test 3. Before starting this test, participants were given instructions describing the difference between remember and know judgements which were adapted from Rajaram (1993) for use in a cued-recall task. The instructions included examples of remember and know judgements and a brief test at the end. On this test, participants were asked to read four examples of statements that could have been made in a task similar to that they were

about to face (e.g., "*When I saw the pair PEN - WALL I imagined writing with a pen on a wall*") and choose what experience these statements indicated from two options: "remember (+ context)" or "know (- context)" experience. After each answer, participants were informed whether it was correct. They were also encouraged to ask the experimenter for clarification before starting the test if they were still in doubt about the difference between "remembering" and "knowing." At Test 3, for each target that was typed in, a prompt appeared asking whether this answer was based on "remembering" or "knowing" and participants were allowed to advance to the next pair only after giving the response. After completing the whole procedure, all participants were asked by the experimenter to describe in their own words the difference between "remembering" and "knowing" to ascertain that they correctly understood the experimental task.

Results and Discussion

Resolution. The means for A_g (a nonparametric measure of resolution) are presented in Table 6.1. A one-way repeated measures Analysis of Variance (ANOVA) investigating the effects of cycle on resolution revealed a significant difference in resolution between the cycles, $F(2, 36) = 51.966$, $MSE = 0.004$, $p < .001$, $\eta_p^2 = .743$. Follow-up *t*-tests showed that resolution (A_g) increased from cycle 1 to cycle 2, $t(19) = 6.706$, $SE = 0.019$, $p < .001$, $d = 1.68$, and from cycle 2 to 3, $t(25) = 3.982$, $SE = 0.021$, $p = .001$, $d = 1.40$. This is consistent with most of the results from studies conducted in the UWP paradigm (although see Hanczakowski et al., 2013, Experiment 4, for an exception).

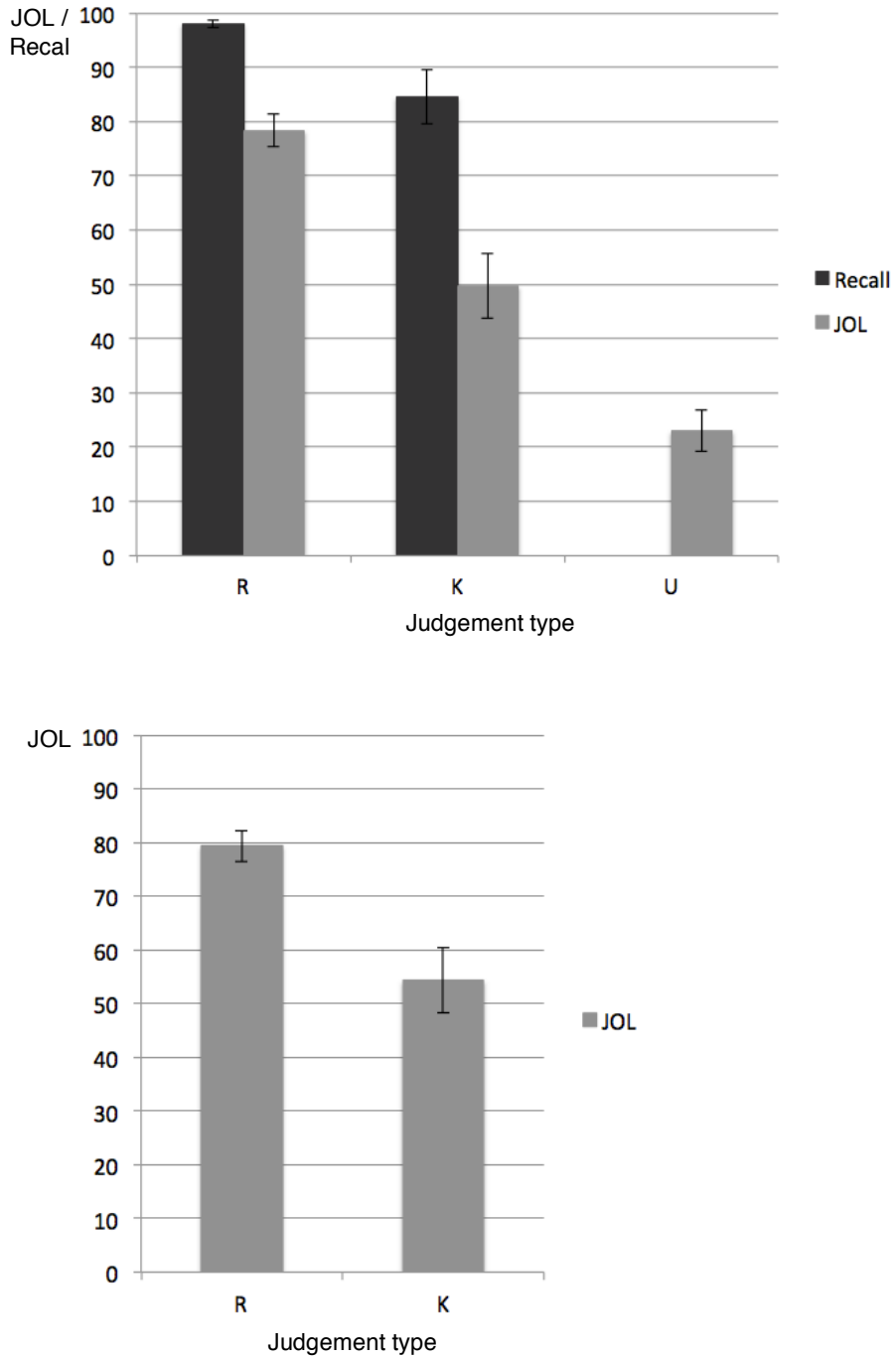

Figure 6.1. Recall performance and JOLs for recalled items assigned Remember (R) and Know (K) ratings and for unrecalled items (U) in Experiment 1. The top panel presents the results for all word pairs. The bottom panel presents JOL data for correctly recalled items only. Error bars denote standard error of the mean.

Calibration. Mean JOLs and recall performance on cycles 1-3 are presented in Table 6.1. A 2 (measure: JOL, recall) x 3 (cycle: 1, 2, 3) repeated-measures ANOVA revealed significant main effects of cycle, $F(2, 38) = 132.021$, $MSE = 89.31$, $p < .001$, $\eta_p^2 = .874$, and measure, $F(1, 19) = 7.838$, $MSE = 305.69$, $p = .011$, $\eta_p^2 = .292$, which were qualified by a significant interaction, $F(2, 38) = 49.577$, $MSE = 38.34$, $p < .001$, $\eta_p^2 = .723$. Follow-up *t*-tests showed that whereas on cycle 1 participants' JOLs were numerically higher than their recall performance, $t(19) = 1.475$, $SE = 4.12$, $p = .16$, $d = 0.38$, the difference was significantly reversed on cycles 2 and 3, $t(19) = 5.922$, $SE = 3.55$, $p < .001$, $d = 1.02$ and $t(19) = 4.055$, $SE = 2.93$, $p = .001$, $d = 0.56$, respectively, revealing the UWP effect.

Table 6.1

Means (SDs) for Recall Performance, JOLs and Resolution (A_g) as a Function of Cycle in Experiments 1, 2 and 3.

	Measure		
Experiment and Cycle	Recall	JOL	A_g
Experiment 1			
cycle 1	28.66 (16.56)	34.75 (16.56)	.68 (.08)
cycle 2	62.10 (20.13)	41.09 (20.71)	.81 (.07)
cycle 3	71.84 (20.16)	59.95 (22.15)	.89 (.06)
Experiment 2			
cycle 1	30.48 (18.77)	42.33 (14.80)	.65 (.12)
cycle 2	59.81 (22.95)	44.83 (15.81)	.77 (.09)
cycle 3	70.70 (20.75)	63.53 (18.07)	.77 (.19)
Experiment 3			
cycle 1	67.39 (16.90)	49.79 (13.62)	.73 (.09)
cycle 2	83.89 (12.08)	69.83 (16.29)	.87 (.13)

Remember/know. Mean JOLs and recall performance for Remember and Know responses are presented in the top panel of Figure 6.1. Remember responses ($M = 33.60$, $SD = 11.29$) constituted 74% of all volunteered responses ($M = 45.30$, $SD = 11.65$). Ninety-eight percent of responses assigned to this category were correct ($M = 33.00$, $SD = 11.24$). Out of the remaining Know responses ($M = 11.70$, $SD = 5.77$), 86% were correct ($M = 10.10$, $SD = 6.35$).

A comparison between JOLs and recall revealed that JOLs were lower than recall performance both for Remember and Know items, $t(20) = 6.753$, $SE = 2.91$, $p < .001$, $d = 1.95$, and $t(19) = 5.249$, $SE = 6.62$, $p < .001$, $d = 1.21$, respectively. A repeated-measures ANOVA with judgement type (Remember, Know, no judgement) as a factor was performed on the JOL data from cycle 3.[10] The ANOVA was significant, $F(2, 34) = 118.325$, $MSE = 113.14$, $p < .001$, $\eta_p^2 = .874$. Follow-up t-tests showed that JOLs assigned to items for which a recall attempt was made later on cycle 3 (that is, Remember and Know items) were higher than JOLs assigned to items that participants did not attempt to recall, $t(18) = 21.725$, $SE = 2.50$, $p < .001$, $d = 5.24$, and $t(17) = 7.255$, $SE = 3.43$, $p < .001$, $d = 2.03$, respectively. Crucially, JOLs for Remember items were also higher than those assigned to Know items, $t(18) = 6.567$, $SE = 4.31$, $p < .001$, $d = 1.90$.

To ensure that the difference in mean JOLs found for Remember and Know items was not caused by a higher percentage of correct Remember than Know answers, we also ran a similar analysis on a restricted data set, comparing mean JOLs only for correctly recalled Remember and Know items (presented in the bottom panel of Figure 6.1). Restricting the sample in this way did not eliminate the difference between Remember and Know

[10] Remember and Know judgements were elicited only for recalled items (independent of their correctness). No judgement was elicited when participants refrained from typing in an answer on the test trial.

JOLs, which was still significant, $t(18) = 5.167$, $SE = 4.79$, $p < .001$, $d = 1.39$.

The pattern of results which revealed the lowest JOLs for subsequently unrecalled items, higher for items later judged as “known”, and the highest for later “remembered” items resembles the one found for “recollected” and “familiar” items in the single-cycle study by Daniels et al. (2009). The pattern is consistent with the notion that people base their JOLs on the quality of memorial information that is available to them during the JOL stage of the procedure. As remember judgements require retrieval of contextual details, the quality of this information is higher than when the retrieved target is not accompanied by retrieval of contextual information, as it is in the case of know judgements. To eliminate one possible explanation of this difference in mean JOLs - that items judged as “remembered” were correct more often (in 98% of the cases) than “known” items (86%) - the same analysis was performed only on correct answers, but the results did not change. Therefore it cannot be assumed that it is only the fact that items for which remember judgements are made are more often correct that causes the difference in the JOL magnitude.

It has to be noted, however, that even though “remembered” items were assigned much higher JOLs than items merely “known”, the UWP pattern was found for both classes of items. Therefore our assumption that the UWP effect is driven by the subset of items for which no context is present found no support in the data. One possible explanation of this result can be based on the findings of Rotello, Macmillan, Reeder, and Wong (2005) who found that the specific wording of instructions can influence how conservative people are in using remember responses. The instructions used in the present experiment were adapted for the cued-recall task from the standard instructions first used by Rajaram (1993) which, according to results of Rotello et al., induced a liberal understanding of remembering. For this reason, in Experiment 2 we changed the instructions used in the remember/know task. The

instructions were altered to resemble those used by Yonelinas (2001) and asked participants to respond “remember” only if they could describe to the experimenter the “remembered” aspects of studying a pair when asked to do so. The present instructions aimed at inducing conservative responding in this task in order to include in the “remember” class only the most confident responses. This was done in an attempt to eliminate the UWP pattern for “remembered” items.

The results of Experiment 1 demonstrated that JOLs assigned to items for which contextual details are later available were higher than JOLs assigned to items for which such details were not present. This generates another question concerning the relationship between cognitive context and JOLs: is it the mere presence of contextual details that plays a role in the assignment of JOLs? Or, alternatively, can the ratings also be influenced by the number of contextual details for each item? The second purpose of Experiment 2 was therefore to investigate this issue.

Experiment 2

Method

Participants. Twenty-four students of the University of Southampton participated in this study for course credit or monetary compensation.

Materials and procedure. The same materials were used as in Experiment 1. The procedure was identical to Experiment 1, with two exceptions. First, the instructions were changed in order to induce more conservative responding in the remember/know task: participants were asked to respond “remember” on Test 3 only if they could describe the “remembered” details to the experimenter when asked to do so. Second, each time participants chose the “remember” response, a second question appeared. In that next step, participants had to decide *how many* contextual details they could recall. The purpose of this additional question was to examine the JOLs assigned to “remembered” answers in more detail. The available options were: one, two, three or more. If JOLs depend

not only on the presence of contextual details available for a given pair, but also on their quantity, they should increase with an increment in the number of recalled contextual details. As inferred from the results of Experiment 1, recall for all "remembered" items should be at ceiling. Therefore a dissociation between JOLs and recall for "remembered" items differing in the number of recalled contextual details was predicted.

As in Experiment 1, after completing the whole procedure, all participants were asked by the experimenter to describe in their own words the difference between "remembering" and "knowing" to ensure that they correctly understood the experimental task.

Results and Discussion

Resolution. The means for A_g are presented in Table 6.1. A one-way repeated-measures ANOVA with cycle (1, 2, 3) as a factor revealed a significant effect, $F(2, 42) = 5.520$, $MSE = 0.004$, $p = .019$, $\eta_p^2 = .208$. A_g increased from cycle 1 to cycle 2, $t(23) = 3.870$, $SE = 0.032$ $p = .001$, $d = 0.80$, but not from cycle 2 to 3, $t < 1$.[11]

Calibration. Mean JOLs and recall performance on cycles 1-3 are presented in Table 6.1. A 2 (measure: JOL, Recall) x 3 (cycle: 1, 2, 3) repeated-measures ANOVA revealed a significant main effect of cycle, $F(2, 46) = 89.746$, $MSE = 126.21$, $p < .001$, $\eta_p^2 = .796$, which was qualified by a significant measure x cycle interaction, $F(2, 46) = 28.569$, $MSE = 80.01$, $p < .001$, $\eta_p^2 = .554$. Follow-up *t*-tests showed that whereas on cycle 1, participants' JOLs were significantly higher than their recall performance, $t(23) = 2.230$, $SE = 5.31$, $p = .036$, $d = 0.46$, this pattern was

[11] This lack of a significant difference between cycle 2 and 3 resolution was caused by cycle 3 results of one participant who assigned a JOL of 100% to all items on cycle 3, and later failed to recall one target (out of 60). This disproportionately affected his or her score: A_g equalled 0 (where chance level is .5). When this person's result was excluded, resulting in a group A_g mean of .81 ($SD = .08$), the difference between cycles 2 and 3 was significant, $t(20) = 2.213$, $SE = 0.019$ $p = .039$, $d = 0.48$.

reversed on cycles 2 and 3, $t(23) = 4.600$, $SE = 3.25$, $p < .001$, $d = 1.03$ and $t(23) = 2.518$, $SE = 2.85$, $p = .019$, $d = 0.53$, respectively, again revealing the UWP effect. The main effect of measure was not significant, $F(1, 23) = 1.053$, $MSE = 403.65$, $p = .32$, $\eta_p^2 = .044$.

Remember/know*.* Mean JOLs and recall performance for Remember and Know responses are presented in Table 6.2. Remember responses ($M = 24.79$, $SD = 15.69$) constituted 56% of all volunteered responses ($M = 44.71$, $SD = 11.67$). This is consistent with the findings of Rotello et al. (2005) in showing that the modified instructions induced more conservative responding; in Experiment 1 the mean number of Remember responses was 33.60, which constituted 74% of all responses to recalled items. Ninety-seven percent of responses assigned to this category were correct ($M = 24.13$, $SD = 15.88$). Out of the remaining Know responses ($M = 19.92$, $SD = 12.27$), 92% were correct ($M = 18.29$, $SD = 12.12$). The mean number and the percentage of correct Know responses increased numerically as compared to those from Experiment 1 (where they equaled 11.70 and 86%, respectively). The increase in the correctness of Know responses from Experiment 1 to Experiment 2 was driven by the fact that many correct responses that under more liberal instructions (such as those used in Experiment 1) would have been classified as “remembered”, were included in the Know category in Experiment 2.

As in Experiment 1, JOLs underestimated recall performance both for Remember items, $t(23) = 5.224$, $SE = 3.52$, $p < .001$, $d = 1.16$, and for Know items, $t(23) = 5.769$, $SE = 4.46$, $p < .001$, $d = 1.26$ (see Table 6.2). This shows that more conservative responding in the remember/know task did not eliminate the UWP effect for Remember items.

A repeated-measures ANOVA with judgement type (Remember, Know, no judgement) performed on the JOL data revealed a significant effect, $F(2, 42) = 38.035$, $MSE = 243.82$, $p < .001$, $\eta_p^2 = .644$. Follow-up *t*-tests showed that JOLs assigned to items later judged to be remembered were higher than those assigned to items later judged as known, $t(23) =$

2.368, *SE* = 4.92, *p* = .027, *d* = 0.49. JOLs for both Remember and Know items were higher than JOLs assigned to items that were not recalled on cycle 3, *t*(21) = 8.752, *SE* = 4.54, *p* < .001, *d* = 1.87, and *t*(21) = 6.660, *SE* = 4.32, *p* < .001, *d* = 1.43, respectively (see Table 6.2). This pattern replicates the one observed in Experiment 1.

Table 6.2

Means (SDs) for Recall Performance and JOLs as a Function of Rating Type in Experiment 2.

	Measure	
Item Set and Rating Type	Recall	JOL
All items		
Remember	96.01 (8.08)	77.59 (16.80)
Know	91.71 (12.28)	65.93 (21.96)
Unrecalled	-	36.96 (20.97)
Correctly recalled		
Remember	-	78.92 (16.78)
Know	-	67.67 (21.96)

We further split Remember judgements according to the number of contextual details that participants were able to recall for a given pair. This allowed for a more in-depth comparison of JOLs for items for which participants reported access to contextual details. Participants were given three options for each remembered item: one, two, or three or more

details, henceforth referred to as Remember 1, Remember 2 and Remember 3+. As only seven participants reported recalling three or more contextual details for at least one pair, Remember 3+ and Remember 2 categories were binned, producing a category of Remember 2+. The results are presented in the bottom panel of Figure 6.2.

The following analyses are restricted to correctly recalled items only. The main reason for this was that recall performance for Know, Remember 1 and Remember 2+ items was at ceiling, exceeding 90% for all item types. In an ANOVA performed both on JOLs and recall performance this ceiling effect could produce a spurious interaction. By equating recall performance between items, it was possible to analyse JOLs only. A repeated-measures ANOVA with judgement type (Remember 2+, Remember 1, Know) as a factor revealed a significant effect, $F(2, 34) = 4.662$, $MSE = 257.37$, $p = .016$, $\eta_p^2 = .215$. The linear trend was significant, $F(1, 17) = 6.333$, $MSE = 372.49$, $p = .022$, $\eta_p^2 = .271$, revealing that JOLs increased with an increase in the amount of contextual details. The quadratic trend was not significant, $F < 1$.

The results of Experiment 2 replicate the results of Experiment 1, again demonstrating that JOLs tend to be higher for items for which contextual details are later retrieved at test. They further demonstrate that it is more than mere *presence* of these details that matters: JOLs also increased for items for which more than one contextual detail was available, as compared with items for which only one detail was accessed at the time of the test. Therefore the detailedness of the contextual information can also serve as a cue for JOLs.

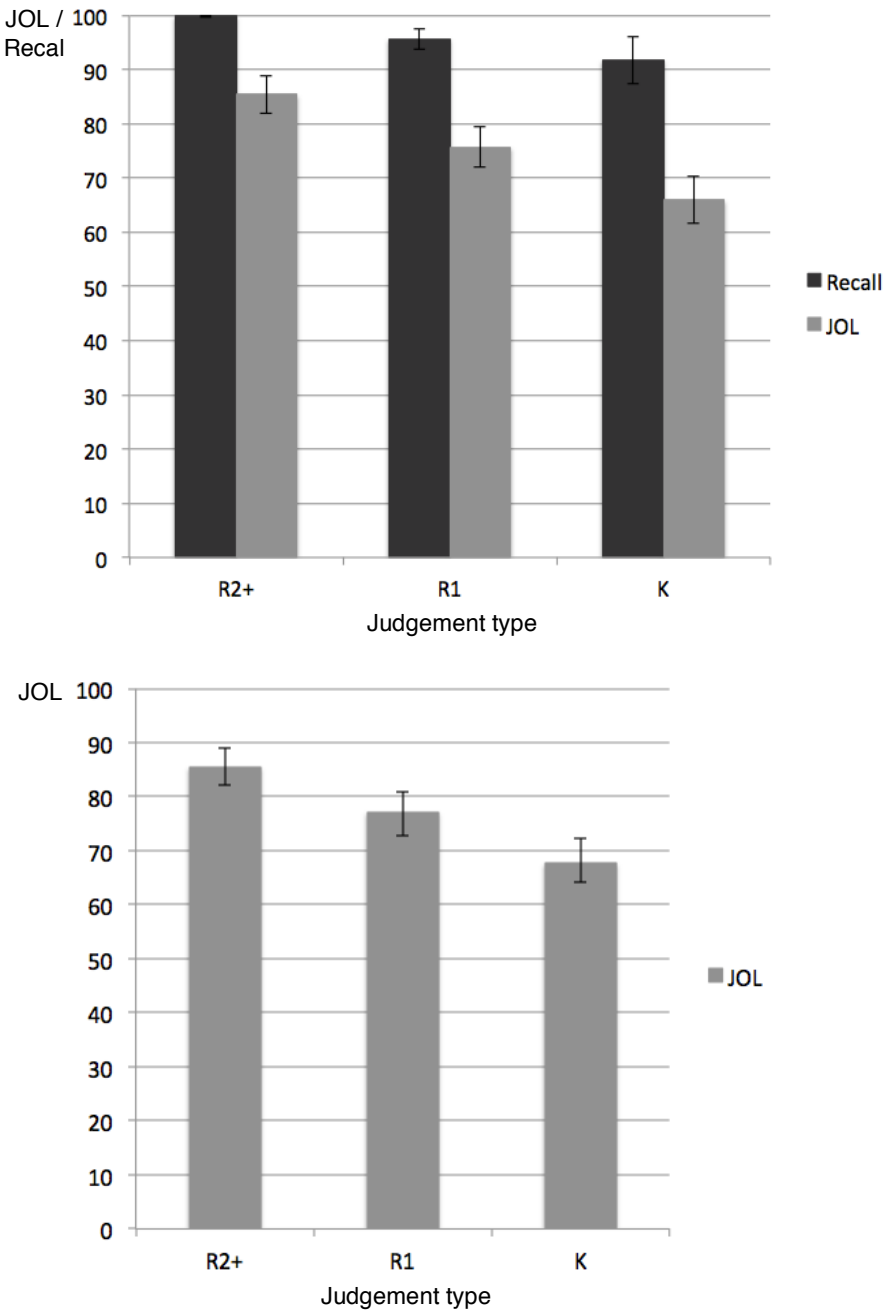

Figure 6.2. Recall performance and JOLs for recalled items assigned Remember ratings with two or more contextual details (R2+), Remember ratings with one contextual detail (R1), and Know (K) ratings. The top panel presents the results for all word pairs. The bottom panel presents JOL data for correctly recalled items only. Error bars denote standard error of the mean.

Experiment 3

In Experiments 1 and 2 Remember and Know judgements were elicited during the last test phase. This was done to ensure that JOLs were not influenced by the previous ratings. If participants could remember whether they chose a “remember” or a “know” option for a given pair, they could later assign higher JOLs to “remembered” items to appear consistent. However, obtaining remember/know judgements on the last cycle limits the conclusions that can be drawn from the results of Experiments 1 and 2. Specifically, it is not known whether the same contextual details that were available at Test 3 were also available during the JOL stage of cycle 3. For example, it is possible that in some cases participants only *identified* at Test 3 new contextual details that were not present at the time of JOL assignment. In this case, these cues could not have influenced cycle 3 JOLs, but nevertheless they could potentially change some of participants’ responses from “know” to “remember”, or from “remember 1” to “remember 2+”. This would have lowered mean JOLs assigned to these classes of items.

To eliminate this possibility, in Experiment 3 we introduced a context rating at Test 1: participants were asked to rate on a scale the quality and quantity of contextual details that they were able to retrieve for each item. We then compared these context ratings from cycle 1 to JOLs elicited on cycle 2. Under these conditions, contextual details encoded during cycle 1 could influence both the context ratings on the cycle-1 test and JOLs made during the study phase on cycle 2.

The use of a scale for the purpose of context rating was supposed to overcome a particular limitation of the remember/know tasks used in the previous experiments. In Experiment 1, the remember/know task allowed only for indicating the presence of contextual information at the time of the rating. In Experiment 2, both the presence and the amount of contextual details could be taken into account. In neither of these experiments, however, the *quality* of contextual information could be rated. It is possible

that some contextual details could be rated as better than others. For example, details that can be related to the self ("The pair *dog-sofa* reminds me of the dog I had when I was little who used to sleep on our sofa") could be rated higher than details relating only to the experimental situation ("I remember imagining a dog sitting on a sofa when I first saw this pair"; see also Castel, Rhodes, & Friedman, 2013). Therefore in Experiment 3 context ratings were provided on a scale, and participants were asked to take into consideration all aspects of contextual details (presence, quantity, quality) when making a rating.

Additionally, we included another type of rating at Test 1. Context ratings were provided only for the half of recalled targets (independent of the correctness of recall). For the other half, participants were asked to provide a confidence rating. We reasoned that querying for context at test could make contextual details more available during the subsequent study/JOL stage of the experiment. This, in turn, should elevate JOLs for pairs for which a context rating was provided above those for which only a confidence rating was required. Such a result would strengthen the conclusion that the availability of cognitive context during the JOL stage of the procedure influences the magnitude of JOLs.

Method

Participants. Twenty-six students of the University of Southampton participated in this study for course credit or monetary compensation.

Materials and Procedure. The same materials were used as in Experiments 1 and 2. This time, however, the procedure consisted of only two study-test cycles instead of three. Both study phases were the same as in previous experiments, with two exceptions: presentation times were increased to 2.5 seconds, and the study list was shortened to 48 word pairs, in order to allow for more successful encoding of both items and contextual information on cycle 1. This was done for two reasons. First, longer presentation times and a shorter list of to-be-encoded pairs were supposed to increase the proportion of items recalled on the cycle-1 test,

for which context of confidence ratings could be elicited. Second, with longer presentation times participants had more time to create cognitive context at study.

Before the first test phase, participants were presented with instructions describing the nature of cognitive context. These instructions were followed by a short test. At this test, participants were presented with the same statements as those presented in Experiments 1 and 2, describing retrieval of targets accompanied and unaccompanied by retrieval of contextual details. As the remember/know task was not used in this experiment, participants were only asked to indicate whether the statements described retrieval of cognitive context ("+ context") or not ("- context").

After the pretest, the cycle-1 test for the word pairs followed. During this test phase, participants were given an additional task. After recalling a word, a prompt appeared on the screen asking participants to provide one of two ratings. The first one was a confidence rating: when presented with the confidence prompt, participants had to rate their confidence that the recalled target was correct on a scale from 0 (no confidence at all) to 3 (high confidence). The second one was a context rating on a scale from "-" to "+++".[12] For this type of rating, participants were instructed that both the amount and quality of available contextual information were to be taken into account. Only one rating was made for each pair and the assignment of word pairs to rating types was random. Rating types alternated between the trials. For half of participants, context rating was the first to be assigned, whereas the other half started with the confidence rating. No rating was elicited for unrecalled pairs.

[12] Different scales were used for both types of ratings in order to ensure that participants did not make the wrong type of rating. For the same reason, the two prompts were presented in different colors and appeared in different places on the screen: the confidence prompt was presented above the pair and in violet, and the context prompt appeared below the pair and in green.

On cycle 2, participants studied and were tested on all word pairs. No additional task was implemented at test 2.

Results and Discussion

Resolution. The means for A_g are presented in Table 6.1. Again, A_g increased from cycle 1 to cycle 2, $t(22) = 3.876$, $SE = 0.033$, $p = .001$, $d = 0.83$.

Calibration. Mean JOLs and recall performance on cycles 1 and 2 are presented in Table 6.1. A 2 (measure: JOL, recall) x 2 (cycle: 1, 2) repeated measures ANOVA revealed a significant main effect of cycle, $F(1, 25) = 423.735$, $MSE = 20.48$, $p < .001$, $\eta_p^2 = .944$, and a main effect of measure, $F(1, 25) = 45.783$, $MSE = 142.29$, $p < .001$, $\eta_p^2 = .647$. The measure x cycle interaction was, however, not significant, $F(1, 25) = 1.219$, $MSE = 67.18$, $p = .28$, $\eta_p^2 = .047$. No UWP pattern was present in the data: on both cycles JOLs were lower than recall performance, $t(25) = 5.143$, $SE = 3.42$, $p < .001$, $d = 1.02$ for cycle 1 and $t(25) = 6.702$, $SE = 2.10$, $p < .001$, $d = 1.42$ for cycle 2.

This is our first failure to replicate the UWP effect in the present set of experiments. We attribute this lack of UWP to prolonged study times. As demonstrated by Scheck and Nelson (2005), under- or overconfidence depends on recall levels. According to their explanation, JOLs are drawn towards an anchor situated between 30% and 50%. If recall performance is far above the anchor, JOLs will usually be lower than mean recall, whereas if it is below the anchor, JOLs will tend to overestimate it. As in our case mean recall performance at Test 1 was 67%, it is plausible that this high recall level led to worsened calibration on cycle 1.

Context judgements. The purpose of introducing two types of judgements at Test 1 was to influence cycle 2 JOLs. We predicted that making a context judgement would draw participants' attention toward contextual details, making this information readily available during the study phase at cycle 2. As a result, JOLs for items for which context

judgements were made should be higher than JOLs for items accompanied by a confidence judgement. A 2 (judgement type: context, confidence) x 2 (measure: cycle 2 JOL, cycle 2 recall) repeated-measures ANOVA revealed only a main effect of measure, $F(1, 25) = 40.125$, $MSE = 150.49$, $p < .001$, $\eta_p^2 = .616$, with JOLs lower than recall performance. The main effect of judgement type and the interaction were not significant, both Fs < 1 (see Table 6.3).

Table 6.3

Means (SDs) for Recall Performance and JOLs as a Function of Rating in Experiment 3.

	Measure	
Rating	Recall	JOL
Confidence	96.08 (6.60)	81.32 (13.07)
Context	97.12 (4.41)	81.38 (11.57)

The absence of a difference between the magnitude of JOLs on cycle 2 for items accompanied at Test 1 by context versus confidence judgements was inconsistent with our predictions. One potential explanation is that processes engaged in a task remain in operation for some time even after a task change, interfering with the completion of a new task (e.g., Kiesel et al., 2010). As in the present experiment the tasks alternated, there are two potential consequences of such an explanation. First, context retrieval might have been enhanced even for items for which a confidence question was asked, leading to an increase in JOLs for these items as well. Second, answering the confidence question might have interfered with context retrieval for items accompanied by the context prompt. In total, this would lead to similar memorial information being

retrieved regardless of the prompt displayed on a particular test trial, leading in turn to equated JOLs.

We further concentrated on items for which context judgements were elicited. These items were split into four categories according to the rating that was assigned. Items assigned a "-" rating were supposed to be based on similar information as Know items in Experiments 1 and 2, as for both types of items no contextual information was retrieved at test. Items with ratings from "+" to "+++" were supposed to be similar to Remember items. However, in contrast to Experiment 2, in the present experiment participants were asked to incorporate not only the amount of contextual details, but also their quality into their judgements. Therefore no direct comparison can be made between R1 and R2+ items from Experiment 2 and the ratings assigned in this experiment.

The results for items assigned context ratings are presented in the top panel of Figure 6.3. A repeated-measures ANOVA performed on JOLs for all items assigned a context rating (-, +, ++, +++) revealed a significant effect, $F(3, 63) = 12.524$, $MSE = 200.45$, $p < .001$, $\eta_p^2 = .374$. The linear trend was significant as well, $F(1, 21) = 26.559$, $MSE = 268.63$, $p < .001$, $\eta_p^2 = .557$, while the quadratic trend was not, $F < 1$, showing that JOLs increased steadily with increasing ratings. For recall, the same ANOVA also revealed a significant effect, $F(3, 63) = 4.088$, $MSE = 128.61$, $p = .01$, $\eta_p^2 = .170$ The linear trend was only marginally significant, $F(1, 21) = 3.765$, $MSE = 280.52$, $p = .067$, $\eta_p^2 = .158$. This time, however, the quadratic component was significant, $F(1, 21) = 7.132$, $MSE = 68.11$, $p = .015$, $\eta_p^2 = .263$, as recall was at ceiling for all items rated "+" or above.

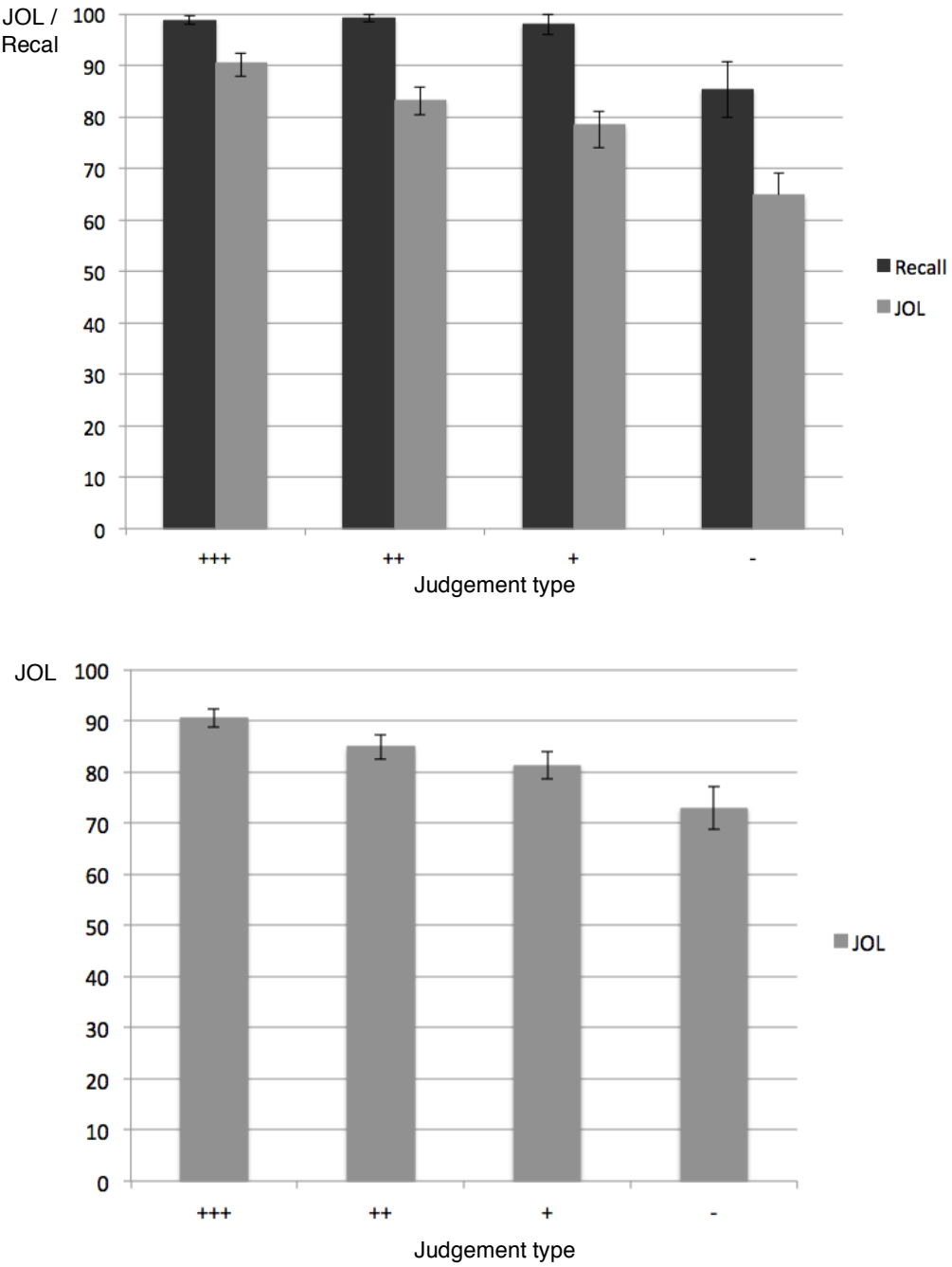

Figure 6.3. Recall performance and JOLs for recalled items assigned context ratings in Experiment 3. The top panel presents the results for all word pairs. The bottom panel presents JOL data for correctly recalled items only. Error bars denote standard error of the mean.

As in Experiment 2, subsequent analyses concentrated on correctly recalled items only (see the bottom panel of Figure 6.3). A repeated-measures ANOVA with rating (-, +, ++, +++) as a factor revealed a significant effect, $F(3, 42) = 11.657$, $MSE = 75.461$, $p < .001$, $\eta_p^2 = .454$. The linear trend was significant, $F(1, 14) = 20.141$, $MSE = 126.35$, $p < .001$, $\eta_p^2 = .590$, demonstrating that JOLs increased with an increase in the amount and quality of recollected contextual details. The quadratic component was not significant, $F < 1$. In total, the pattern of results for JOLs remained the same even when only a selected subset of the data was analysed.

Confidence judgements. The results for items assigned confidence ratings are not of main interest from the perspective of this study. Nevertheless, we present the data for completeness. Means for JOLs and recall performance for items assigned confidence judgements are presented in Table 6.4. For the confidence task, participants were asked to provide their ratings on a scale from 0 to 3. However, as less than a half of participants used all confidence categories, with even fewer using the lowest confidence rating, the lowest two categories were binned for the subsequent analyses. A repeated-measures ANOVA performed on JOLs with confidence rating (0-1, 2, 3) as a factor revealed a significant effect, $F(2, 42) = 31.245$, $MSE = 144.86$, $p < .001$, $\eta_p^2 = .598$. The linear trend was also significant, $F(1, 21) = 40.670$, $MSE = 202.667$, $p < .001$, $\eta_p^2 = .659$, and so was the quadratic trend, $F(1, 21) = 9.299$, $MSE = 87.045$, $p = .006$, $\eta_p^2 = .307$, as the increase between items assigned the lowest and middle confidence ratings was greater than that between items assigned middle and the highest ratings.

For recall, an analogous ANOVA also revealed a significant effect, $F(2, 42) = 6.398$, $MSE = 279.32$, $p = .004$, $\eta_p^2 = .234$. The linear trend was significant, $F(1, 21) = 6.181$, $MSE = 458.20$, $p = .021$, $\eta_p^2 = .227$, as was the quadratic trend, $F(1, 21) = 7.389$, $MSE = 100.44$, $p = .013$, $\eta_p^2 = .260$: recall performance reached ceiling for items assigned ratings of 2 and 3.

When the analysis on JOLs was performed only on ratings for correctly recalled items, a similar pattern emerged as on the full data set, $F(2, 30) = 31.617$, $MSE = 43.600$, $p < .001$, $\eta_p^2 = .678$ for the ANOVA and $F(1, 15) = 53.716$, $MSE = 50.349$, $p < .001$, $\eta_p^2 = .782$, for the linear trend. The quadratic component this time was not significant, $F(1, 15) = 1.420$, $MSE = 36.847$, $p = .25$, $\eta_p^2 = .086$. In general, cycle 2 JOLs increased with an increase in cycle 1 confidence.

Table 6.4

Means (SDs) for Recall Performance and JOLs as a Function of Item Set and Confidence Rating in Experiment 3.

Item Set and Measure	Confidence Rating		
	0-1	2	3
All items			
Recall	79.88 (34.31)	98.00 (10.00)	98.80 (3.40)
JOL	57.56 (23.13)	81.83 (11.05)	88.78 (8.94)
Correctly recalled			
JOL	66.67 (13.22)	81.74 (10.99)	88.82 (9.00)

In general, the pattern of results found for context ratings in the present experiment was consistent with the findings from the remember/know task used in Experiments 1 and 2. Items for which contextual details were absent at the time of Test 1 received the lowest JOLs on the subsequent cycle. For items for which contextual details were

present at Test 1, as indicated by the context ratings, JOLs increased with an increase in ratings. As the context ratings in this experiment were supposed to reflect both the quality and quantity of contextual details retrieved, this suggests that these factors play a role in assigning JOLs in the multi-cycle procedure.

General Discussion

In three experiments, we have demonstrated that JOLs elicited on later cycles of a multi-cycle procedure are affected by contextual details for the rated pairs. Experiment 1 demonstrated that the mere presence of such contextual details has the potential to increase JOLs, as compared to items for which these details were not available. Experiments 2 and 3 have shown that the quality and quantity of such details matter as well: the stronger the overall representation of context, the higher the JOLs. Several features of these results are worth highlighting, to which we now turn.

First, these results conceptually replicate and extend prior findings showing that JOLs are influenced by the presence of contextual details. As suggested by behavioral (Daniels et al., 2009) and electrophysiological (Skavhaug et al., 2013) measures in single-cycle procedures, higher JOLs are assigned to items for which retrieval of such details occurs at a later test. Our findings resemble those of Daniels et al. and Skavhaug et al. in a sense that, in the multi-cycle procedure, we have also found higher JOLs for items later assigned “remember” judgements or earlier assigned ratings indicating the availability of cognitive context. As indicated by recall performance for items accompanied by cognitive context in the present set of experiments, assigning high JOLs to these items is a beneficial strategy: the proportion of correctly recalled items was consistently higher across all experiments for items judged as “remembered” or assigned ratings from “+” to “+++” than for items merely “known” or assigned a “-” rating.

One interesting difference between the present study and the studies of Daniels et al. (2009) and Skavhaug et al. (2013) is a consequence of the number of study-test cycles that participants underwent. In the previous studies, single-cycle procedures were used, in which JOLs were elicited at the time of participants' first encounter with the studied pairs. As a result, JOLs could only be driven by contextual details being *encoded* at the time of study. In the multi-cycle procedure, on the other hand, on cycle 2 and beyond participants assign JOLs to previously encountered items. This creates another potential way in which cognitive context can affect the ratings: namely, by its *retrieval*. It seems likely that participants sometimes access previously encoded details when represented with a previously studied pair; for example, a person may remember creating a particular mental image linking the cue with the target. It is therefore viable that in the multi-cycle procedure both the encoding and retrieval of cognitive context have the potential to influence JOL assignment.

Second, our hypothesis that the UWP effect would be eliminated for "remembered" items was not supported by the present data. Even though the magnitude of the UWP effect was diminished (as shown in Figure 6.1 and Table 6.2) for items accompanied by contextual details, the apparent underconfidence was still present. This pattern persisted even when instructions were changed to induce conservative responding in the remember/know task, which was supposed to limit Remember responses to those that were supported by the strongest memorial evidence. Therefore it cannot be concluded that the UWP effect is driven solely by items for which no contextual details are present.

The data from Experiments 2 and 3 give insight into why the UWP effect might have been present even for items accompanied by cognitive context. These results suggest that it is not only the presence or absence of contextual details that can influence JOLs: when contextual details are available, their quality or quantity may also be used as a cue for JOLs. This allows for making fine-grained distinctions between items for which

cognitive context is available. Crucially, recall performance for these items is excellent (the lowest mean recall performance for items for which contextual details were retrieved was 96% for R1 items in Experiment 2). This means that the distinctions participants make when assigning 0-100% JOLs do not reflect the probability of recalling items accompanied by cognitive context at an immediate test: if it was recall that JOLs were predicting, JOLs for these items should be equated in spite of differences in the quality and quantity of retrieved contextual details.

In this regard, the present results resemble the findings described in Paper 1, which show that JOLs elicited on a 0-100% scale distinguish between items that share the same level of recall performance. To review, the experiments described in Paper 1 demonstrated that participants' scale JOLs elicited on cycle 3 distinguish between items that differ in terms of the number of successful recall attempts on the preceding cycles: they are higher for items recalled on both preceding cycles than for items recalled on only one cycle. Binary bets, on the other hand, which are akin to yes/no JOLs (see Hanczakowski et al., 2013), accurately predicted no difference in recall performance for these two classes of items, showing that participants are able to correctly predict the probability of future recall when the task they face is appropriate for that purpose. These results are consistent with the idea proposed by Hanczakowski et al. (2013) that JOLs are used by participants not for the purpose of rating the probability of future recall, but as ordinal ratings of confidence in future recall. These two interpretations of JOLs differ significantly from the perspective of calibration research: the probability interpretation allows for calculating calibration measures, while the confidence interpretation does not.

The present results are fully consistent with the confidence interpretation of 0-100% JOLs. The fact that JOLs for context-rich items increased with an increase in the quality and quantity of contextual details, while recall was at ceiling, suggests that JOLs cannot be mere ratings of probability of future recall. Therefore the results of this study join the

results of Hanczakowski et al. (2013) and those reported in Paper 1 in demonstrating that UWP is most likely an artefact of using the percentage scale to assess calibration.

The present study extends the findings reported in Paper 1 by highlighting the importance of a particular cue - contextual details associated with a studied pair - for the assignment of high JOLs on the percentage scale. As the results of Experiment 1 from Paper 1 demonstrate, the number of successful recall attempts on the preceding cycles can influence 0-100% JOLs. However, it was not clear which aspects of the pairs previously recalled once and twice served as cues in the JOL task. One possibility is that participants could have relied on their memory for the number of successful recall attempts, simply assuming that each additional recall attempt warrants an increase in JOLs. Pairs recalled on all cycles are also likely to be subjectively easier than pairs for which one of the recall attempts was unsuccessful. Therefore it was not possible to determine which particular aspects of the pairs participants relied on while assigning their 0-100% JOLs.

The present study, on the other hand, clearly demonstrates that a particular cue - the availability of cognitive context - can be employed by participants in the multi-cycle 0-100% JOL task. Although the generic term “cognitive context” encompasses a broad range of experiences, from the perspective of a participant they seem similar enough: they all are created by participants for themselves. It remains an open question whether externally-generated context, such as that used in research on context effects in memory, would influence the assignment of scale JOLs in the same way as internally-generated cognitive context. This question can be addressed by future research.

In sum, the present study demonstrated the influence of retrieval of cognitive context on the assignment of 0-100% JOLs in the UWP paradigm. JOLs were shown to increase with increasing strength and the level of detail of the context available. This was true even for pair types for

which recall performance was equated and at ceiling. This finding is consistent with the notion that participants use their JOLs to rank order the rated items in terms of memorial information available for each of them. As a result, it supports the interpretation of the UWP effect as an artefact of using percentage scales for eliciting JOLs.

7. Paper 3

JOLs are not what they seem: A signal-detection (re)interpretation of judgements of learning

Rating scales are ubiquitous in psychological research. In general, the scales used by psychologists can roughly be divided into two groups (e.g., Biernat, Manis, & Nelson, 1991; Frederick & Mochon, 2012). *Subjective scales* are characterized as having no predetermined meaning: the interpretation of the points on these scales cannot be inferred *a priori*, without taking into account what the ratings actually refer to. For example, on a scale ranging from *very small* to *very large,* the precise meaning of the labels depends on the range of sizes of to-be-rated items. With such scales, there is no contradiction that a *very small* mammal can still be larger than a *very large* insect. *Objective scales*, on the other hand, have predefined, objective referents. The interpretation of, say, weight in grams should always be the same, independent of whether the animal being weighed is an insect or a mammal.

In memory and metamemory research, researchers commonly use measures such as retrospective confidence (RC) judgements, and prospective measures such feeling-of-knowing (FOK) judgements or judgements-of-learning (JOLs), amongst others, to investigate internal assessments of participants' own knowledge. Often the scales metacognitive theorists use are subjective, such as a *1*-to-*6* scale of RC. Metacognitive studies employing subjective scales are often concerned with *resolution* - that is, the extent to which the assigned scale values discriminate between correct versus incorrect responses on some criterial test (e.g., correctly recalled vs. not correctly recalled on a recall test following a JOL judgement; correctly recognized vs. not correctly recognized on a recognition test following an FOK judgement, etc.). For resolution, the absolute magnitude of judgements is irrelevant, as long the

ratings distinguish correctly between these two types of responses. So, for example, if a person assigned FOK ratings of *6* to all subsequently recognized items, the same perfect resolution would be obtained as long as they assigned any ratings lower than *6*, be it *5* or *1*, to all subsequently unrecognized items. Popular measures of resolution, such as gamma correlations or signal detection measures of *d'*, d_a, or area under the Receiver Operating Characteristic (ROC) curve (*AUC*) can be calculated from an ordinal scale, and a subjective *1*-to-*6* scale satisfy this requirement.

The same metacognitive ratings can also be elicited on objective scales, such as *0* to *100%* scales of subjective probability. In order for this scale to be interpreted as objective, the scale values must have some pre-set referents. It is assumed that they refer to the likelihood of some outcome in the long run (a frequentist approach to probability). In the case of JOLs, a rating of *40%* would mean, then, that a person predicts recalling at a future test 40% of all items assigned this rating.

Objective metacognitive scales have one notable advantage over their subjective counterparts: they allow for an additional measure of metacognitive accuracy to be calculated which reflects the correspondence between ratings and objective performance. *Calibration* can be assessed at separate levels on the rating scale (e.g., percentage correct is calculated separately for all items assigned a rating of *0*, *10*, ..., *100%* and then the two values for each level are compared), or for the whole test. In the latter case, a mean of all metacognitive judgements is calculated and compared to memory performance for the whole list. Perfect calibration (or realism) requires the two means to be equal. On the other hand, a rating mean that is lower than the performance mean is interpreted as underconfidence, whereas the reverse pattern is interpreted as overconfidence. Therefore, it is assumed that by having participants use the objective *0-100%* scale, researchers can gain insight into how good they are at estimating, in objective terms, their overall level of

knowledge. Calibration scores have been used by experimenters to draw conclusions about potential similarities or differences in monitoring abilities in developmental research (e.g., Connor, Dunlosky, & Hertzog, 1997; Lipko, Dunlosky, Lipowski, & Merriman, 2012, Rast & Zimprich, 2009), eyewitness research (e.g., Allwood, Ask, & Granhag, 2005; Sauer, Brewer, Zweck, & Weber, 2010) and educational research (e.g., Butler, Karpicke, & Roediger, 2008; Dunlosky & Rawson, 2012), among many other areas of psychology.

However, some concerns regarding the interpretation of the *0-100%* scale have been formulated in the JOL literature. Recently, Hanczakowski, Zawadzka, Pasek, and Higham (2013) cast doubt on the likelihood interpretation of 0-100% JOLs. Their research concerned the underconfidence-with-practice (UWP) effect - an impairment of calibration present when the same materials are studied and tested more than once. Hanczakowski et al. tested whether the UWP effect found with *0-100%* JOL likelihood scales generalised to other predictions of future memory performance, such as binary (yes/no) JOLs and binary betting decisions.13 They found that, in contrast to the underconfidence observed with *0-100%* scales, the proportion of "yes" responses on the binary tasks did not differ from the proportion of correctly recalled items, revealing good calibration. Thus, if scale JOLs were measuring subjective probability and participants were truly underconfident, the question remains as to why this underconfidence was consistently "repaired" with the binary tasks. The most straightforward explanation is that participants were not truly underconfident at all and that the UWP effect is an artefact of *0-100%* JOL scales.

---

13 With binary tasks, realism would be evident if the percentage of "yes" responses (i.e., binary JOL: "yes, I will remember the item later"; binary betting: "yes, I am willing to bet that I will recall the item later") equaled the percentage of items recalled.

On the basis of these results, the authors suggested that unless it is demonstrated that the JOL scale is indeed used by participants to assess the probability of future recall, it may be safer to interpret JOLs as confidence judgements rather than assessments of likelihood. Confidence judgements differ from likelihood judgements in one important aspect: the scale on which they are made may well be subjective. Subjective scales do not allow calibration to be assessed because the scale values have no absolute meaning; researchers cannot conclude that participants are realistic if items assigned a rating of *40* have a 40% recall probability any more than they can if those same items were assigned *4* on a six-point scale. Importantly, if the typical *0-100%* JOL scale is subjective rather than objective, there would be reason to question the validity of some of the calibration effects found in the metacognitive literature.

Following up on research by Hanczakowski et al. (2013), the experiments described in Paper 1 investigated the assignment of the highest JOLs in a procedure consisting of three study-test cycles, akin to that used in UWP research. The results demonstrated that JOLs made on cycle 3 in this multi-cycle procedure were higher for items previously recalled twice (on both preceding cycles) than for items recalled only once (on one or the other preceding cycles). This difference in JOLs, however, was not accompanied by a difference in recall performance: all previously recalled items were extremely likely (>90%) to be recalled again on cycle 3. Importantly, this effect was not caused by incorrect predictions concerning recallability of items previously recalled once and twice: when participants were given a binary betting task instead of the scale JOL task in their Experiment 2, they were able to correctly predict future recall with their bets. This demonstrates that even though participants were aware that recall will be comparable and at ceiling for both classes of items (evinced by the binary-betting data), discriminations were made between the item classes using their scale JOLs (evinced by the scale JOL data). This finding undermines the common assumption that *0-100%* JOL scales

are objective because the scale values do not represent participants' estimates of probability and hence the values to not correspond to predetermined and fixed entities.

However, there are more reasons why the *0-100%* scale may not satisfy the objectivity assumption. Research in other fields of psychology has questioned the assumed immunity of the values on scales considered to be objective to experimental manipulations. For example, Frederick and Mochon (2012) have demonstrated that objective ratings pertaining to, for example, weight in pounds, are susceptible to context effects. In their study, two groups of participants estimated a critical value (e.g., the weight of an adult giraffe). The control group answered only the critical question whereas the experimental group answered the critical question after making an estimation that was much higher or lower than the critical one and which acted as an anchor (e.g., the weight of a wolf or a blue whale). The results revealed an anchoring effect: the critical estimation in the experimental group was lower versus higher than control if the initial question pertained to wolves versus blue whales. More importantly, the authors convincingly demonstrated that their results were not due to a changed representation of the rated object (e.g., the giraffe seeming heavier than it really is), as the anchoring effect did not generalise to different measures presumed to tap the same representation, like the weight of the giraffe in tons, or the number of lions that could feed on one giraffe. Instead, the authors proposed that it was the *interpretation of the rating scale* that was affected by the anchor: the same representation of weight of the giraffe in pounds was simply conveyed by different values, depending on the presence or absence of an anchor. In other words, the interpretation of values on a seemingly objective scale was shown to be context dependent.

How, then, should the *0-100%* JOL scale be interpreted? We believe that analysing JOLs from a signal detection theory (STD) perspective can be useful in answering this question. Although interpreting prospective

metacognitive judgements in terms of SDT is relatively rare, it is not unheard of (e.g., Benjamin & Diaz, 2008; Hanczakowski et al., 2013; Higham, Zawadzka, & Hanczakowski, 2014; Masson & Rotello, 2009). In the remainder of this paper, we present a signal-detection reinterpretation of the *0-100%* JOL scale and assess the consequences of such an approach.

A Signal-Detection Interpretation of JOLs

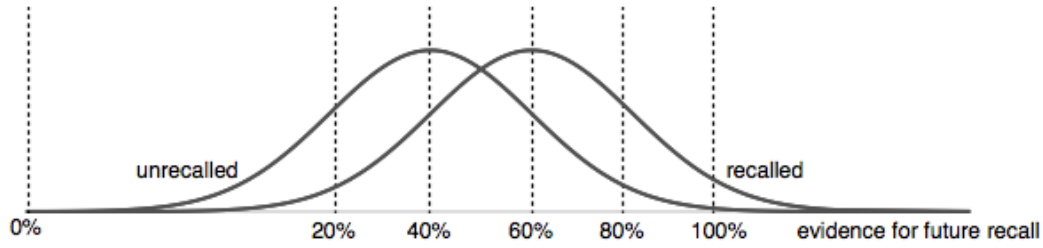

Figure 7.1. A signal-detection representation of the JOL task. Two distributions are placed on the evidence-for-future-recall dimension. The distribution on the left represents items not recalled on a given cycle, while the distribution on the right represents items recalled on a given cycle. Vertical lines denote separate criteria. For a JOL of, say, *20%*, the evidence for that item must exceed the 20% criterion, but fall below the 40% criterion. In order for the highest rating (*100%*) to be assigned, the evidence must exceed the highest criterion.

Figure 7.1 shows a signal-detection representation of the scale-JOL task. Two distributions of studied items are positioned on an evidence-for-future-recall dimension.[14] On average, items that will be recalled at a later point have stronger evidence for future recall than later unrecalled items; therefore the distribution of later recalled items is positioned to the right of

[14] In this paper, we refer to the evidence dimension as representing evidence for future recall, as this is what participants are supposed to rate in the JOL task. However, the exact nature of this internal dimension need not be precisely specified (see e.g., Benjamin & Diaz, 2008).

the distribution of later unrecalled items. The distance between the means of the two distributions shows how well people are able to distinguish between later recalled and unrecalled items - that is, how good their resolution is.

Scale values are treated as separate *criteria* that are malleable and under participants' control. The criteria, denoted by vertical lines in the figure, indicate the minimum amount of evidence that is needed for a given rating to be assigned. The distributions are partitioned by these criteria and each of the criteria is assigned a particular JOL value, in this example in increments of 20. The rule for assigning a JOL to any item sampled from the distributions is straightforward: the item is assigned the JOL value corresponding to the criterion closest to it on the left-hand side (i.e., the nearest criterion with evidence less than or equal to the item's evidence). Thus, an item that falls between the *40%* and *60%* criteria will be assigned *40%*. However, if it has enough evidence to exceed the *60%* criterion as well (but not the *80%* criterion), it will be assigned *60%*. Critically, the positioning of the criteria on the evidence dimension is not static but differs depending on situational context and task demands. For example, if the experimental situation calls for a large amount of evidence before assigning a given rating (i.e., the situation creates a conservative decision strategy), the further to the right the criterion for that rating is located.

The signal detection approach allows ROCs to be plotted. ROCs are isosensitivity curves that display the relationship between hit rates and false alarm rates. An example of a metacognitive ROC is presented in Figure 7.2. To generate such an ROC, for each JOL level (e.g., in increments of 20, i.e. *0, 20, 40, ..., 100%*) the proportion of recalled items which were assigned a given JOL or higher (a *hit rate, HR*) is plotted against the proportion of unrecalled items which were assigned a given JOL or higher (a *false alarm rate, FAR*). For example, if a person assigned a JOL of *40%* or higher to four out of 10 unrecalled (or incorrectly recalled) items, and to eight out of 10 correctly recalled items, then the coordinates

of the point on the ROC corresponding to the value of *40%* would be (0.4, 0.8). These points on the ROC denote separate criteria, showing the minimum amount of evidence for future recall that is necessary for a given JOL value to be assigned. The proportion of the plot that falls below the ROC curve gives a measure of resolution known as the *area under the curve* (*AUC*). The better a person is at using the JOL values to discriminate between items that will and will not be recalled at test, the greater the AUC. If JOLs perfectly discriminate between subsequently recalled and unrecalled items, AUC equals 1.0. Conversely, if JOLs only discriminate at chance levels (i.e., HR = FAR for all ROC points), the ROC follows the minor diagonal of the plot, and the AUC equals 0.5.

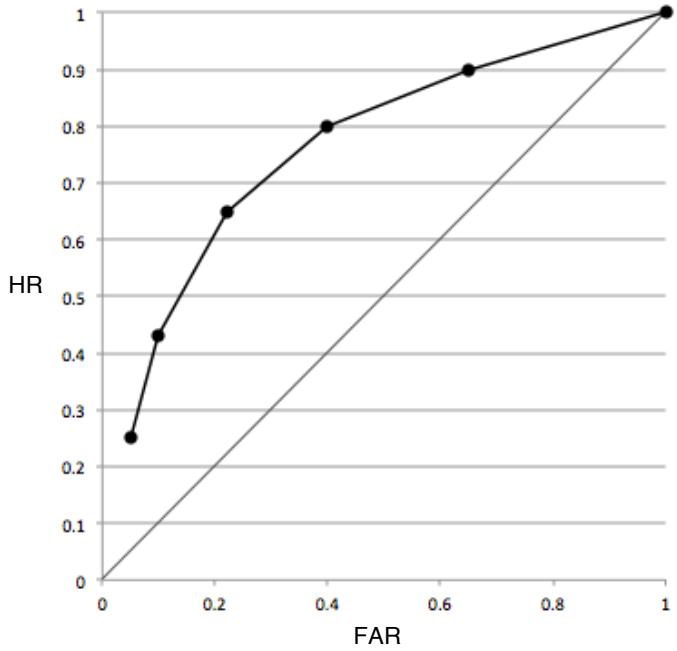

Figure 7.2. An example of a metacognitive ROC curve. Points on the curve denote separate confidence criteria. The most liberal criterion (*0%*) is placed in the top right corner, while the most conservative criterion (*100%*) is at the end of the curve in the bottom left part of the graph.

Context Dependence of JOL values

The signal detection approach can shed new light on the claims concerning the *0-100%* JOL scale. If JOL values are treated as separate criteria, the behavior of these criteria under certain manipulations can be informative of how people employ the percentage scale. Here, we will concentrate on whether the interpretation of *0-100%* JOLs is affected by the context of a study list.

Context dependence of JOLs has been suggested by Koriat (1997), who noted that JOLs are comparative and driven by the relative recallability of items within a list. However, list-context effects have also been documented in studies in which recallability did not differ between the types of items presented within the same list. Recently, Susser, Mulligan, and Besker (2013; Experiment 1) investigated the effect of font size on JOLs and recall. Previously, Rhodes and Castel (2008) demonstrated that JOLs assigned to items presented in a larger font are higher than those assigned to items presented in a smaller font, even though font size has no effect on recall performance. Susser et al. found, however, that this effect is limited to mixed lists, consisting both of items presented in a small and large font. When the list was pure (i.e. consisting either only of items shown in a large or in a small font), the effect of font size on JOLs disappeared. Similarly, the results of Experiment 1 described in Paper 1 revealed differences in JOLs for previously recalled items despite equated recall performance (see above).

Differences in JOLs despite a lack of difference in recall performance can be interpreted in two different ways. One interpretation is to assume that the contrast within the study list makes people truly underestimate their chances of recalling the “weaker” words, and hence JOLs for these items are lowered. The common interpretation of such a pattern would be, then, that for this subset of items participants are underconfident - in other words, they believe their memory for these items to be worse than it really is. This explanation is consistent with the traditional view that, in the JOL

task, people rate the likelihood of future recall, and thus with the objective interpretation of the JOL scale. However, another interpretation is to assume that this contrast influences the *ratings* assigned to the weaker items, but not their *perception*. In other words, although participants are aware that the manipulation does not affect the probability of future recall, they lower their ratings to demonstrate their awareness of the differences between the types of items within the list. As a result, they no longer attempt to rate the likelihood of future recall, and the scale becomes subjective.

Crucially, these two accounts can be distinguished using SDT. The objective, or *metacognitive contrast* account (see e.g., Pansky & Goldsmith, 2014, for an example), predicts that what is affected is the perceived evidence for future recall. Therefore what changes is the placement of the weaker-item distribution on the evidence dimension. The subjective, or *bias* account, on the other hand, predicts that it is the placement of confidence criteria that is affected - the experimental manipulation influences the way in which the criteria are distributed on the evidence dimension.

An experimentally-induced redistribution of confidence criteria would not be a new finding in research employing SDT. One such an example comes from a recognition memory study by Mickes, Hwe, Wais, and Wixted (2011), who investigated the assignment of confidence ratings (on a *1-20* scale) to strong memories. In Experiment 5, the authors gave their participants a recognition memory task in the plurals paradigm (see e.g., Hintzman, Curran, & Oppy, 1992). At test, participants' task was to distinguish between targets and lures using a 20-point confidence scale. The test was split into two halves. After the first half, the experimental group was provided with feedback concerning their responses. Although feedback had no effect on participants' discrimination performance in the second half of the test, it changed their bias: responding became more conservative, and the effect was most pronounced for the high-confidence

ratings. As the ROCs suggested, the manipulation changed the way criteria were distributed: whereas during the first half of the test they were clustered together, after feedback, the distance between the criteria increased.

What manipulation could potentially cause such a redistribution of criteria on the *0-100%* JOL scale? One obvious candidate is the range of evidence for future recall. Suppose that two groups of participants are presented with a list of word pairs. For the first group, the list consists of pairs of moderate difficulty only (henceforth referred to as *critical*). For the second group, there are both critical and difficult word pairs on the list. In both cases, participants provide immediate JOLs for each of the word pairs, and later they are given a cued recall test. Figure 7.3 depicts these two scenarios. In the narrow-range scenario (middle panel), two distributions are positioned on the evidence-for-future-recall dimension: the unrecalled items distribution on the left, and the recalled items distribution on the right. In the wide-range scenario, there are two distributions of unrecalled items: one for critical pairs, and one for difficult pairs. (For simplification, let us assume that none of the targets from very difficult pairs was recalled, so there is no distribution for recalled difficult pairs). The unrecalled difficult-pair distribution is located further to the left of the unrecalled critical-pair distribution, as, on average, the evidence for future recall is weaker for the former pair type. Thus, the range of evidence is greater for the mixed list of critical and difficult pairs compared to the pure list of critical pairs.

Alternatively, the range of evidence can be extended by adding very easy new pairs. This scenario is presented in the bottom panel of Figure 7.3. (Again, let us assume for simplification that all easy word pairs were recalled, so there is no distribution for unrecalled related word pairs). This time, the new easy item distribution is located to the right of the recalled critical item distribution, as, on average, evidence is greater for the easy than for the critical pairs. This scenario is depicted in the bottom panel of

Figure 7.3. The addition of new easy items again extends the experienced range of evidence as compared to the pure list of critical word pairs in the middle panel.

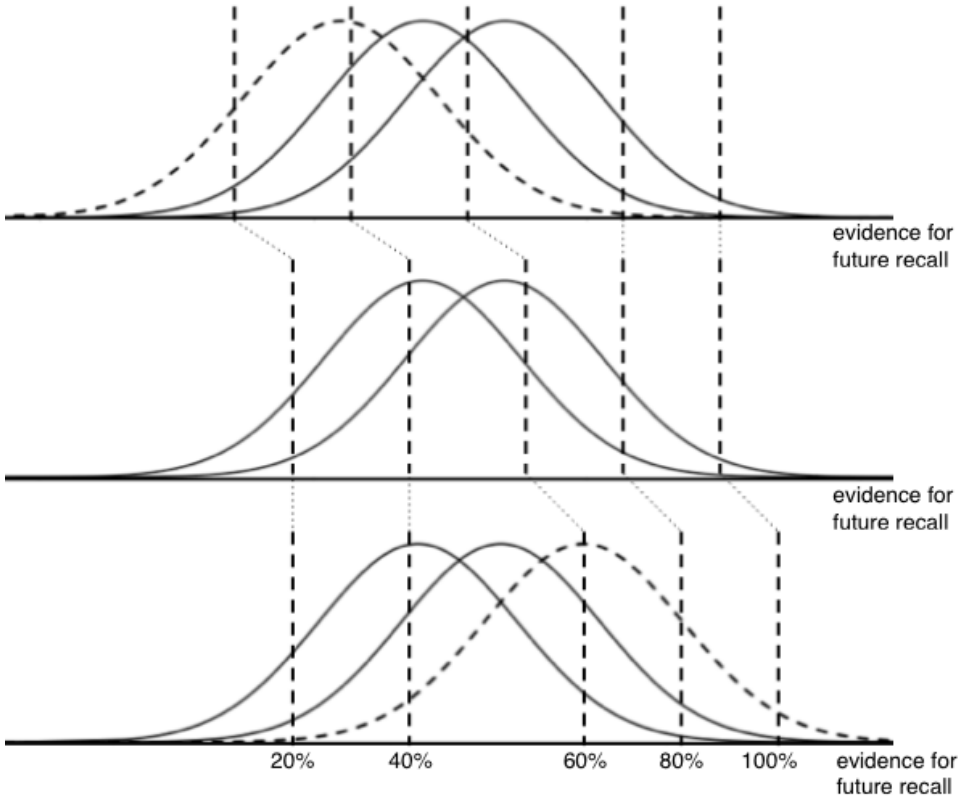

Figure 7.3. A graphical illustration of predictions of the bias account. Solid lines represent distributions for critical items, with the unrecalled-item distribution on the left, and the recalled-item distribution on the right. The middle panel represents a case in which only critical items are studied. The top panel includes, in addition to critical items, a dashed leftmost distribution of very difficult new items. The dashed rightmost distribution in the bottom panel represents the easy new items distribution. Vertical dashed lines represent confidence criteria. The bias account predicts shifts only of those criteria that are close to the new item distribution. In the top panel, this is evidenced by a shift of the lower criteria to the left of the evidence dimension, and in the bottom panel by a shift of the upper criteria to the right of the dimension, as compared to the baseline presented in the middle panel.

How might the difference in range of evidence for future recall influence criterion setting? We propose that participants adjust their criteria to accommodate the range of evidence that they experience. In this way, different JOL values can be used to effectively discriminate between items that differ in evidence for their future recall. This is based on, but goes beyond, rating probability of future recall at an immediate test. As the results Experiment 1 described in Paper 1 demonstrate, participants assigned different ratings to items previously recalled once and twice, even though these items did not differ in terms of their recall at test. However, they likely differed in terms of evidence for future recall: not only were items recalled twice subjectively easier to learn, but also they may well have been more impervious to forgetting (see, e.g., Pyc & Rawson, 2012; Vaughn & Rawson, 2011) in the long run. In any case, there were true differences between the two types of items, and participants' JOLs picked up on these differences. However, those differences did not include the subjective probability of future recall.

Experimental Overview

In two experiments, we tested the prediction that the interpretation of JOL values depends on the range of experienced evidence for future recall. Range of evidence was manipulated by including new items in the study list. In the control conditions of both experiments, participants studied and were tested on the same list of items on all cycles, as it is commonly done in the multi-cycle paradigm. In the experimental conditions, some of the studied items were substituted on cycle 2 with new items in order to extend the range of evidence for future recall for the whole list. In Experiment 1, these new items (new unrelated word pairs, and nonword-word pairs) were more difficult than the *critical items* (i.e. items studied on all cycles) which was intended to extend the range of evidence downward. We predicted that this would affect specifically the low confidence criteria, as the lowest confidence values would be reserved

for the difficult new items. As a result, the placement of these criteria on the evidence dimension should be more liberal (i.e., *less* evidence would be needed for an item to surpass these criteria) in the experimental than in the control condition. In Experiment 2, in order to extend the range upward, the new items (pairs studied repeatedly before the multi-cycle procedure, and pairs in which the cue and the target were the same word) were easier than the critical items. We expected this manipulation to affect the high-confidence criteria, which should become more conservative (i.e., *more* evidence would be needed for an item to surpass these criteria). Note that the addition of new items should have no influence on the evidence for future recall for the critical old items.

If our manipulations were successful, the assignment of JOLs to critical items should be affected, but no effect on recall is anticipated. Compared to control conditions, with no new items included, changes in the JOL mean will necessarily affect the magnitude of the difference between the mean of JOL and mean recall - a common measure of calibration. In Experiment 1, an apparent decrease in underconfidence in the experimental condition, as compared to the control condition, should be found, while in Experiment 2 this apparent underconfidence should increase in the experimental condition.

Experiment 1

Method

Participants. Sixty students of the University of Southampton participated for course credit. Thirty were assigned to the control group, and 30 to the experimental group.

Materials and procedure. The procedure consisted of three study-test cycles. On cycle 1, participants in both groups studied and were tested on the same list of 60 unrelated pairs. The pairs were created from 120 words of medium frequency and ranging from four to eight letters in

length, chosen from the MRC database. On cycles 2 and 3, participants in the control condition were presented with the same 60 pairs as on cycle 1. In the experimental condition, only 20 pairs - henceforth referred to as *critical pairs* - were taken from the initial study list (and thus were the same as in the control condition), and the remaining 40 pairs were new. Twenty of these new pairs consisted of a nonword as a cue and a legal word as a target. The other 20 pairs consisted of two unrelated words not presented before. Different new pairs were presented on the second and third cycles.

The study and test phase procedures were identical for both groups. First, all pairs were presented individually for study for 1.5 s. After the presentation of each pair, the target disappeared from the screen, leaving only the cue. Participants were then asked to rate the likelihood of recalling the target at test when presented with the cue. The rating could be any value from 0% to 100%. Time for providing the judgement was not limited.

The test immediately followed the study phase. All pairs studied on the given cycle were included in the test. On each test trial, participants were presented with a cue and their task was to type in the target that accompanied that cue during the study phase. If they were not able to recall the target, they were asked to press “Continue” to advance to the next cue. The order of presentation of pairs was randomized anew for each participant on each study and test phase.

Results and Discussion

Descriptive statistics for JOLs and recall performance for critical and non-critical pairs are presented in Table 7.1. Resolution scores are presented in Table 7.2.

Cycle 1. On cycle 1, the materials that participants studied and were tested on were the same in both groups. Therefore, no differences between the groups were expected. Nonetheless, we compared cycle-1 performance between the two groups to eliminate the possibility of

sampling error. A 2 (group: control, experimental) x 2 (measure: JOL, recall) mixed Analysis of Variance (ANOVA) conducted on both the critical and non-critical pairs, with group as the only between-subjects variable, revealed only a significant main effect of measure, $F(1, 58) = 10.639$, $MSE = 174.07$, $p = .002$, $\eta_p^2 = .155$. Mean JOLs ($M = 35.89$, $SD = 15.01$) were higher than mean recall performance ($M = 28.03$, $SD = 11.45$). Neither the main effect of group, nor the interaction, was significant, both Fs < 1.

As the analyses performed on cycles 2 and 3 concentrate mostly on critical pairs, we conducted the same 2 x 2 ANOVA on JOL and recall results for these pairs only to ensure that performance for these pairs did not differ between the groups. The pattern of results was identical to that found for the full data set. There was a significant main effect of measure, $F(1, 58) = 8.393$, $MSE = 192.97$, $p = .005$, $\eta_p^2 = .126$, with JOLs ($M = 32.68$, $SD = 16.68$) exceeding recall performance ($M = 25.33$, $SD = 12.91$). Neither the main effect of group, nor the interaction, was significant, both Fs < 1. We also found no difference in resolution (A_g, a nonparametric measure of AUC) between the experimental and control groups, $t < 1$. Taken together, cycle 1 results confirm that baseline performance was equal between the groups.

Cycle 2. First, we checked whether the difficulty manipulation implemented in the experimental group was successful. A repeated-measures ANOVA performed on mean JOLs for three pair types (critical, new word-word, and new nonword-word pairs), was significant, $F(2, 58) = 116.775$, $MSE = 64.377$, $p < .001$, $\eta_p^2 = .801$. JOLs for the critical pairs were higher than those for new word-word pairs, $t(29) = 8.284$, $SE = 1.81$, $p < .001$, $d = 1.53$, which were, in turn, higher than those for new nonword-word pairs, $t(29) = 8.937$, $SE = 1.86$, $p < .001$, $d = 1.76$ (see Table 7.1). This result demonstrates that participants distinguished between these types of pairs using their JOLs. A similar ANOVA performed on the recall data for these pairs was significant as well, $F(2, 58) = 36.840$, $MSE = 88.28$, $p < .001$, $\eta_p^2 = .560$. Recall performance for the three pair types

mirrored the pattern for JOLs, with the critical pairs being recalled more often than the new word-word pairs, $t(29) = 6.321$, $SE = 2.19$, $p < .001$, $d = 1.18$, which were recalled more often than the nonword-word pairs, $t(29) = 2.276$, $SE = 2.37$, $p = .030$, $d = 0.43$.

Table 7.1

Means (SDs) for JOLs and Recall Performance for Critical and Non-Critical Repeated Pairs in the Control and Experimental Groups and New Word-Word and Nonword-Word Pairs in the Experimental Group in Experiment 1.

	Cycle 1		Cycle 2		Cycle 3	
Group and Pair Type	JOL	Recall	JOL	Recall	JOL	Recall
Control						
critical	33.05 (16.00)	25.00 (13.13)	38.32 (15.58)	55.83 (19.17)	59.58 (20.32)	73.50 (17.98)
non-critical repeated	37.97 (15.00)	31.93 (10.77)	43.81 (14.62)	64.27 (15.50)	64.77 (17.97)	80.20 (14.38)
Experimental						
critical	32.32 (17.60)	25.33 (12.93)	46.26 (17.96)	58.00 (19.24)	68.18 (16.35)	74.00 (17.29)
non-critical						
repeated	36.99 (14.91)	27.27 (13.87)	-	-	-	-
new word-word	-	-	31.27 (16.53)	43.00 (16.43)	34.18 (17.43)	41.33 (21.37)
new nonword-word	-	-	14.62 (12.48)	38.00 (13.43)	14.81 (11.45)	7.67 (10.06)

Note: The terms “repeated” and “new” refer to pair status on cycles 2 and 3, as on cycle 1 all pairs are new.

Table 7.2

Means (SDs) for A_g for Critical Pairs in Control and Experimental Groups in Experiment 1 and Experiment 2.

Experiment and Group	Cycle 1	Cycle 2	Cycle 3
Experiment 1			
control	.69 (.11)	.77 (.12)	.85 (.12)
experimental	.70 (.14)	.79 (.11)	.89 (.10)
Experiment 2			
control	.63 (.15)	.86 (.16)	-
experimental	.64 (.16)	.84 (.15)	-

All other analyses on the cycle 2 and 3 data, unless noted otherwise, were performed for the 20 critical pairs only, which were identical for both groups. Cycle 2 JOLs and recall performance for these pairs were subjected to a 2 (group) x 2 (measure) mixed ANOVA that was analogous to the one conducted in cycle 1. The main effect of measure was again significant, $F(1, 58) = 40.431$, $MSE = 158.69$, $p < .001$, $\eta_p^2 = .411$, only this time, mean recall performance exceeded mean JOLs ($M = 56.92$, $SD = 19.06$ and $M = 42.29$, $SD = 17.15$, respectively). Had list context exerted an effect on JOLs in the predicted direction, JOLs should have been higher in the experimental group than the control group whereas recall should have been equated, producing an interaction. However, although the data pattern was in the predicted direction – that is, the mean difference between JOLs and recall performance was numerically greater in the control (17.5%) than in the experimental condition (11.7%; see Table 7.1) – neither the main effect of group nor the interaction was significant, $F(1, 58) = 1.555$, $MSE = 492.78$, $p = .22$, $\eta_p^2 = .026$, and $F(1, 58) = 1.576$, $MSE = 158.69$, $p = .21$, $\eta_p^2 = .026$, respectively.

One potential reason that our between-group manipulation of list composition did not exert a significant interactive pattern on cycle 2 performance is that in order for the new pairs to be perceived as difficult, the level of performance for old, critical pairs probably needs to be high enough for participants to consider these pairs as easy. Only then would experimental participants be inclined to adjust their confidence criteria relative to the control group. Although recall performance on cycle 2 for these pairs was better than cycle 1, and better than for the new, non-critical pairs introduced on cycle 2, it may not have been high enough to warrant a criterion shift. However, cycle 3 performance should meet these requirements, to which we now turn.

There was no between-group difference in resolution (A_g) for critical pairs, $t < 1$ (see Table 7.2).

Cycle 3. As on cycle 2, a repeated-measures ANOVA performed on mean JOLs for three pair types (critical, new word-word, and new nonword-word pairs) studied in the experimental group, was significant, $F(2, 58) = 197.613$, $MSE = 110.79$, $p < .001$, $\eta_p^2 = .872$. JOLs for the critical pairs were higher than those for new word-word pairs, $t(29) = 11.384$, $SE = 2.99$, $p < .001$, $d = 2.08$, which were, in turn, higher than those for new nonword-word pairs, $t(29) = 9.172$, $SE = 2.11$, $p < .001$, $d = 1.91$ (see Table 7.1). The same ANOVA performed on the recall data for these pairs was also significant, $F(2, 58) = 270.494$, $MSE = 122.01$, $p < .001$, $\eta_p^2 = .957$. Recall performance again was the highest for the critical pairs, which were recalled more often than the new word-word pairs, $t(29) = 11.611$, $SE = 2.81$, $p < .001$, $d = 2.19$. New word-word pairs were, in turn, recalled more often than the nonword-word pairs, $t(29) = 10.869$, $SE = 3.09$, $p < .001$, $d = 2.48$.

Cycle 3 JOLs and recall performance for critical pairs were subjected to a 2 (group) x 2 (measure) mixed ANOVA that was analogous to the one conducted in cycles 1 and 2. As with cycle 2, it revealed a significant main effect of measure, $F(1, 58) = 34.120$, $MSE = 85.70$, $p <$

.001, η_p^2 = .370, again caused by the mean of JOLs being lower than recall performance (*M* = 63.88, *SD* = 18.79 and *M* = 73.75, *SD* = 17.48, respectively). However, unlike the cycle 2 analysis, the main effect was qualified by a significant measure x group interaction, *F*(1, 58) = 5.740, *MSE* = 85.70, *p* = .020, η_p^2 = .090: even though recall performance was equated between the control and experimental groups, participants in the experimental condition assigned lower JOLs to the critical items than participants in the control condition, decreasing the discrepancy between the two measures (5.82% vs 13.92%; see Table 7.1). The main effect of group was not significant, *F*(1, 58) = 1.098, *MSE* = 565.22, *p* = .30, η_p^2 = .019.

To examine the influence of difficult pairs on JOLs in more detail, we constructed ROC curves for critical pairs (see panel A of Figure 7.4). We first compared resolution between the groups. As seen in Figure 7.4, the ROC curves for the experimental and control groups overlap, which suggests comparable levels of resolution. To confirm that, we calculated A_g, which did not differ between the conditions, $t < 1$. This shows that our manipulation of list difficulty did not impair participants' ability to discriminate between subsequently recalled and unrecalled critical items on cycle 3. Thus, neither resolution nor recall performance differed between the groups, so neither variable is able to explain the difference in JOLs found in cycle 3.

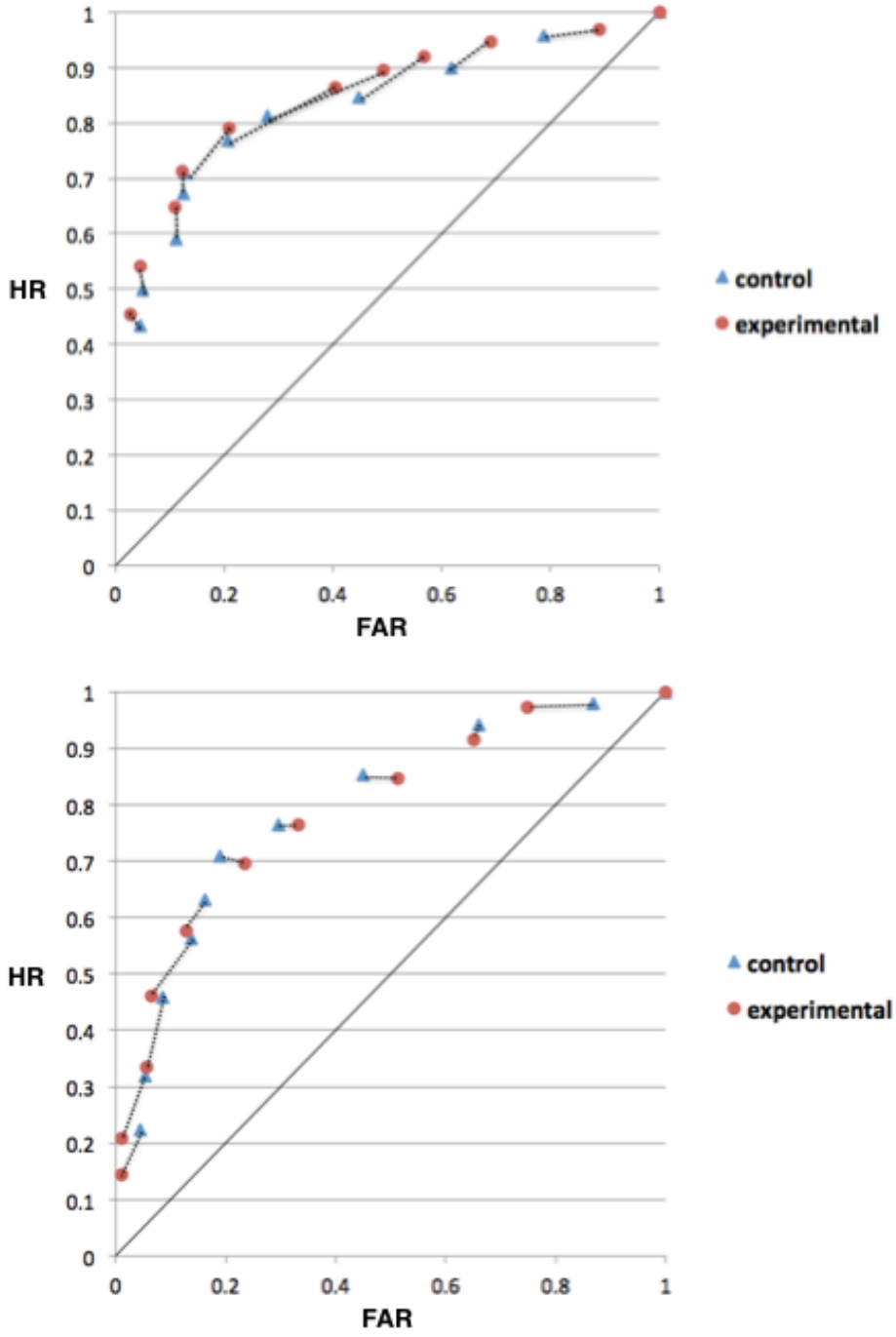

Figure 7.4. ROCs for Experiment 1 (top panel) and Experiment 2 (bottom panel). The overlap between the curves presented in each panel suggests comparable resolution in both groups. The selective misalignment of points on the two curves (from the top right corner to the middle of the curve in the top panel, and in the bottom left corner in the bottom panel) suggests criterion shifts.

Another possible reason that JOLs differed between the groups is that the manipulation of list context affected the confidence criteria. Specifically, the inclusion of new pairs made participants *recalibrate* their confidence scale such that the amount of evidence for future recall warranting the assignment of particular JOL values was adjusted (see Hanczakowski, Zawadzka, & Higham, 2014, for similar considerations regarding retrospective confidence ratings). Specifically, according to the SDT model, new, difficult pairs would be located at a new, low end of the evidence dimension, extending the total range of evidence downward. To accommodate these pairs, participants would have shifted their lower confidence criteria downward as well. This shifting is evidenced in the ROC in Figure 7.4 by the liberal (top-right) points being offset between the groups, with the points in the experimental group being further to the top-right of the ROC space (i.e., more liberal) than those in the control group. The consequence of the lower-criteria shifts was an increase in JOLs assigned to difficult *critical* items (which are of only moderate difficulty in the experimental group, occupying the middle of the evidence range).

On the other hand, the ROC in the top panel of Figure 7.4 suggests criterion placements at the top end of the range (the conservative region) were not affected by the addition of new, difficult pairs. Indeed, the placements of the highest confidence criteria (> *60%* – bottom-left area of the ROC) are almost identical between the control and experimental ROCs. This result is sensible because criteria in this region of the evidence dimension are far away from the region occupied by the new, difficult items. Consequently, they do not need to be adjusted to accommodate them and JOLs assigned to the easiest critical items remain unchanged.

Table 7.3

Mean c_1 Across Confidence Levels in the Control and Experimental Groups on Cycle 3 of Experiment 1 and On Cycle 2 in Experiment 2. The Lower the c_1 Value, the More Liberal the Criterion.

	Confidence Level									
Experiment and Group	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*	*80%*	*90%*	*100%*
Experiment 1										
control	-1.05	-0.65	-0.37	-0.13	0.04	0.24	0.29	0.41	0.68	0.77
experimental	-1.22	-0.83	-0.62	-0.49	-0.34	0.00	0.23	0.34	0.63	0.80
Experiment 2										
control	-1.50	-0.94	-0.44	-0.09	0.16	0.31	0.45	0.70	0.99	1.17
experimental	-1.18	-0.81	-0.49	-0.13	0.10	0.44	0.74	0.93	1.45	1.57

In addition to the ROC analysis, c_1, a criterion measure suitable for cases in which the underlying distributions have unequal variance (see Macmillan & Creelman, 2005), was calculated from group data for each criterion level.[15] The means for c_1 are presented in Table 7.3. Overall, the results are consistent with the selective criterion shift account outlined

[15] For technical reasons, we did not analyse statistically between-group differences in criterion setting, as it was not possible to calculate measures of criterion setting for each participant. Calculating measures such as c_1 requires converting HRs and FARs to *z* scores. This, however, cannot be done for HRs and FARs equalling either 1 or 0. As such HR and FAR values were common especially at the low and high confidence levels, excluding these cases would lead to substantial data loss. When corrections were used to convert HRs and FARs equaling 0 or 1 to values that would allow calculation of *z* scores, the resulting corrected data violated the normality assumption, also precluding calculation of c_1.

above: in the experimental group, the lowest and middle criteria were shifted to a greater extent than the high criteria. Therefore it can be concluded that the increase of the mean of JOLs in the experimental group as compared to the control group was caused mostly by the more liberal use of the ratings of *60*% and below.

There is, however, more than one other mechanism that could be responsible for the observed differences in JOLs between the groups. According to the *metacognitive contrast* explanation, the inclusion of new pairs in the experimental group may affect the *perception* of critical pairs: when compared to new, difficult pairs, the critical pairs seem easier than they really are. This effect would be represented in the SDT model as a distribution shift rather than a criterion shift; that is, the inclusion of new difficult pairs in the experimental group would cause the distributions of critical items to increase. As the perceived amount of critical-item evidence for future recall increases, higher confidence criteria are surpassed and thus higher JOL values are assigned. The fundamental difference between the criterion shift and metacognitive contrast accounts lies therefore in the accuracy of assessments that participants make. Whereas in the former case participants still can accurately assess the amount of evidence for future recall, in the latter case this assessment is distorted.

Crucially, it is possible to distinguish between the two accounts by investigating the ROCs. If it is metacognitive contrast that produces the difference in JOLs between the experimental and control conditions on cycle 3 – that is, in the experimental condition both unrecalled and recalled critical items indeed seemed easier than they really were, producing a distribution shift – the placement of *all* confidence criteria should differ between the conditions. An inspection of the ROCs reveals, however, that this is not the case; as noted, the high (≥*70*%) confidence criteria shifted noticeably less than those below *70*%.

However, a more complex version of the metacognitive-contrast account might be postulated. In principle, it is conceivable that only the

perception of pairs characterized by a relatively low level of evidence for future recall would be affected by the inclusion of difficult new pairs, as these two types of pairs would be close to each other on the evidence dimension. If this were true, primarily the items at the bottom end of the unrecalled item distribution would shift upward. The items at the top end of this distribution, as well as items within the recalled-item distribution (i.e., items with more evidence that are not as close to the new, difficult pairs) would remain static. Such a selective shift upward would effectively reduce the variance of the unrecalled item distribution in the experimental group compared to the control group. Because ROCs are sensitive to the ratio of the variances of the evidence distributions, the net result of this account, therefore, would be a difference in the shape of the ROC between the groups. However, a visual inspection of the ROCs shows that this was not the case, as both curves are virtually identical. We conclude, therefore, that the metacognitive contrast account is not a viable explanation for the present set of results.

Experiment 2

Experiment 1 was successful at demonstrating that if new, difficult items were introduced on later cycles of the multi-cycle paradigm, participants adjusted their confidence criteria to accommodate them, which increased mean JOLs assigned to critical items. The purpose of Experiment 2 was to experimentally demonstrate that if non-critical items set an easy (rather than hard) context, participants' attempts to accommodate these items will result in lowered JOLs to critical items relative to the control condition in which easy, non-critical items are absent. Furthermore, ROC analysis should demonstrate that the reason for this effect is shifting of the upper confidence criteria to higher placements on the recall evidence dimension.

Method

Participants. Sixty-six students of the University of Southampton and Cardiff University participated for course credit or payment. Thirty-three were assigned to the control group, and 33 to the experimental group.

Materials and procedure. The procedure consisted of a pre-study phase and two study-test cycles. The materials were the same as in Experiment 1. Out of 60 word pairs used on cycle 1 of that experiment, 15 were assigned to the pre-study condition, and the remaining 45 were used in the multi-cycle procedure. During the pre-study phase, participants studied and were tested four times on a list of 15 unrelated cue-target pairs. Repeated study was implemented so that these items would be well learned. The study phases were the same with each pair presented for 1.5 s with a 500 ms ISI. The tests, however, were simple initially but then gradually became more difficult to facilitate learning (e.g., Finley, Benjamin, Hays, Bjork, & Kornell 2011). The first was a recognition test, where the cue was presented and participants were supposed to choose the target from among three alternatives, two of which were new. The second test was cued-recall during which the cue was presented along with the first letter of the target and participants were expected to type in the target. The third and fourth tests were also cued recall, but only the cue was presented with no target letter. These latter tests were the same as those in the JOL phase of the experiment.

After the pre-study phase, 45 unrelated pairs were studied and tested in two cycles. On cycle 1, the same 45 pairs were used in both groups. On cycle 2, participants in the control condition studied and were tested on the same 45 pairs as on cycle 1. In the experimental condition, 15 critical pairs were taken from the list studied on cycle 1, and the remaining 30 pairs were new. Fifteen of these new pairs were taken from the pre-study phase of the experiment, and hence were highly familiar. The remaining 15 new pairs consisted of cues and targets that were *identical* (e.g., *grass-*

grass; Castel, McCabe, & Roediger, 2007). These new pairs were expected to elicit high JOLs and set an easy context.

The procedure within each cycle was the same as in Experiment 1, although note that there were only two rather than three study-test cycles in this experiment compared to the last. The reduction in the number of cycles and in the number of pairs studied on each cycle was implemented to limit fatigue effects that might otherwise have arisen with the pre-test that was added in this experiment. The order of presentation of pairs within each cycle was randomized anew for each participant on each study and test phase, including the pre-study phase.

Results and Discussion

Descriptive statistics for mean JOLs and recall performance are presented in Table 7.4. Table 7.2 presents resolution scores for critical pairs.

Pre-study phase. In this phase only recall performance on the last test, identical in format to that used on the two main study-test cycles in the latter phase of the experiment, was measured. On average, participants recalled correctly 13.2 (88%) out of the 15 tested items in the control group (SD = 2.65), and 13.5 (90%) in the experimental group (SD = 2.66), $t < 1$. Thus, our pre-test procedure was successful at producing excellent learning of the items, meaning that introducing these items in second study-test cycle in main experiment should create an easy context.

Cycle 1. As in Experiment 1, a 2 (measure: JOL, recall performance) x 2 (group: control, experimental) mixed ANOVA, with group as the only between-subjects factor, was conducted on both the critical and non-critical pairs. It revealed only a main effect of measure, $F(1, 64)$ = 24.373, MSE = 215.468, $p < .001$, $\eta_p^2 = .276$: on the first cycle the mean of JOLs exceeded mean recall performance (M = 43.42, SD = 15.37 vs M = 30.80, SD = 17.70). Neither the main effect of group nor the interaction were significant, both Fs < 1. The same ANOVA conducted on the data for

critical pairs only produced similar results. Only the main effect of measure was significant, $F(1, 64) = 5.136$, $MSE = 314.64$, $p = .027$, $\eta_p^2 = .074$, with the mean of JOLs exceeding mean recall performance ($M = 44.27$, $SD = 16.65$ vs $M = 37.27$, $SD = 21.02$). Neither the main effect of group nor the interaction were significant, both Fs < 1. Resolution (A_g) also did not differ between the groups, $t < 1$. These results demonstrate that the level of performance before the introduction of the experimental manipulation was equated between the groups.

Table 7.4

Means (SDs) for JOLs and Recall Performance for Critical and Non-Critical Repeated Pairs in the Control and Experimental Groups and Non-Critical New Studied and Identical Pairs in the Experimental Groups in Experiment 2.

	Cycle 1		Cycle 2	
Group and Pair Type	JOL	Recall	JOL	Recall
Control				
critical	44.66 (15.77)	35.56 (21.51)	51.90 (20.36)	60.81 (26.07)
non-critical repeated	43.42 (14.86)	29.52 (18.93)	46.14 (19.25)	56.64 (24.31)
Experimental				
critical	43.88 (17.72)	38.99 (20.69)	47.21 (20.07)	64.44 (22.95)
non-critical				
repeated	42.57 (15.62)	25.97 (16.99)	-	-
new studied	-	-	77.10 (20.26)	85.00 (19.08)
new identical	-	-	66.84 (20.07)	79.59 (18.78)

Note: The terms “repeated” and “new” refer to pair status on cycle 2, as on cycle 1 all pairs are new.

Cycle 2. To confirm that the "easy" pairs in the experimental group were indeed perceived as easier than the critical pairs, a one-way ANOVA was performed on mean JOLs for the three pair types: critical, identical, and studied. The ANOVA revealed a significant effect, $F(2, 64) = 35.144$, $MSE = 216.59$, $p < .001$, $\eta_p^2 = .523$. Mean JOLs for critical pairs were lower than for identical pairs, $t(32) = 4.771$, $SE = 4.11$, $p < .001$, $d = 0.83$, which were, in turn, lower than those assigned to pairs taken from the pre-study phase, $t(32) = 2.701$, $SE = 3.80$, $p = .011$, $d = 0.47$. The same ANOVA conducted on recall data also revealed a significant effect, $F(2, 64) = 20.045$, $MSE = 187.532$, $p < .001$, $\eta_p^2 = .385$. As for JOLs, recall was lower for critical than for identical pairs, $t(32) = 3.887$, $SE = 3.89$, $p < .001$, $d = 0.68$. The difference in recall performance between the identical pairs and pairs from the pre-study phase was marginally significant, $t(32) = 1.721$, $SE = 3.17$, $p = .095$, $d = 0.30$.

Although the results for new pairs are not the focus of the present study, two interesting aspects of the data for identical pairs have to be noted. First, JOLs for pairs from the pre-study phase exceeded those for identical pairs which, at least on the surface, should seem easier to learn. The most parsimonious explanation of that result is that the previously studied pairs were learned so well that at this stage of the experiment they simply did not require additional learning. Words constituting identical pairs, on the other hand, had not been encountered before in the course of the experiment. Hence, these pairs required encoding on cycle 2. This is consistent with the recall results: the difference in recall performance between pre-studied and identical pairs was in the same direction as the difference in JOLs and marginally significant. Second, in contrast to McCabe et al. (2007), who found that JOLs for these pairs overestimated recall performance, in our data we found a 14% *under*estimation, as participants were able to recall correctly almost 80% of targets after a single presentation. A potential explanation of the excellent recall performance is that identical pairs stood out during the test phase: these

were the only pairs in which cues (and identical targets) were not highly familiar. Both the critical and repeated pairs had been encountered before - either during the pre-study phase, or on cycle 1 - while the identical pairs were new to participants. Therefore, participants might have simply adopted the strategy of indicating that the target was the same as the cue whenever they encountered a relatively unfamiliar cue at test.

The remaining analyses on cycle-2 data were performed on the 15 critical pairs only. The same measure x group ANOVA as on cycle 1 was performed on cycle 2 JOL and recall data for critical pairs. Again, the main effect of measure was significant, $F(1, 64) = 49.009$, $MSE = 115.06$, $p < .001$, $\eta_p^2 = .434$, although this time the mean of JOLs underestimated mean recall performance ($M = 49.55$, $SD = 20.27$ vs $M = 62.63$, $SD = 24.44$). Crucially, the interaction was significant as well, $F(1, 64) = 4.965$, $MSE = 115.06$, $p = .029$, $\eta_p^2 = .072$: the difference between means of JOLs and recall performance increased with the inclusion of new, easy pairs in the experimental group (17.23%) compared to the control group (8.91%). The main effect of group was not significant, $F < 1$.

As in Experiment 1, we plotted and compared cycle 2 ROCs for both groups (see panel B of Figure 7.4). Again, the two curves were similar, suggesting comparable resolution. This was confirmed by the comparison of A_g, which did not differ between the groups, $t < 1$. As in Experiment 1, selective criterion shifts seem to be the only viable explanation of our results. As evidenced by the ROCs, the placement of the criteria between *70* and *100%* (bottom-left corner) consistently differed between the groups by one criterion: the amount of evidence needed for a rating of *70%* in the experimental group warranted a rating of *80%* in the control group, and the same applies to the other, higher criteria up to the end of the scale. The lower criteria, on the other hand, mostly overlap between the groups, the only exception being the *10%* criterion.

The ROCs are again not consistent with the metacognitive contrast account for the same reasons as in Experiment 1. Specifically, if the entire

distribution of items was shifted by the presence of the easy items (i.e., critical items had less subjective evidence of later recall in the experimental group compared to the control group), then there would not be selective misalignment of only the conservative points on the ROCs. Rather, all points on the ROC would be misaligned. Conversely, if only the critical items high on the dimension were shifted to a lesser point on the dimension, then the ratio of variances would be affected and the two ROC curves would not overlap. Overall, the results of Experiment 2 confirm the finding of Experiment 1 that manipulating context with non-critical items influences certain criterion settings in the JOL task.

In this experiment we reversed the pattern obtained in Experiment 1. By introducing new, easy pairs, we increased, rather than decreased, the discrepancy between the means of JOLs and recall performance. As evidenced by the ROCs, the context manipulation made the high criteria in the experimental group more conservative. These results supports the claim that JOLs are relative in nature, and the meaning of JOL values depends on the context in which the judgements are made.

General Discussion

In the present study, we employed signal-detection methods to analyse responding in the multi-cycle JOL task. By treating JOL levels as separate confidence criteria, we have demonstrated that the meaning of particular JOL values is context dependent, and it is influenced by the range of evidence for future recall for all items on the study list. In Experiment 1, the inclusion of difficult, new pairs in the experimental group extended the range downward, compared to the control group, affecting the positioning of the low and middle (≤*60%*) confidence criteria. In Experiment 2, the range was extended upward by the easy new pairs, consistently affecting the high (≥*70*%) confidence criteria. The fact that JOL values can be treated as confidence criteria, malleable and context dependent, suggests that the *0-100%* JOL scale is subjective in nature.

As mentioned in the Introduction, in order for scale ratings to be interpreted as reflecting probability, a ratio scale with predetermined referents is required. As our results suggest, the *0-100%* JOL scale does not satisfy this requirement, as the meaning of the JOL values depends on the experimental context. Therefore, this scale cannot be used for the purpose of rating *probability* of future recall. Consequently, the common interpretation of JOL values as reflecting probability levels is likely incorrect. It does not mean, though, that the ratings made on this scale cannot be meaningfully interpreted. We propose instead a more parsimonious explanation that JOLs represent the *ranking* of the items within the list in terms of evidence for future recall. For such an interpretation, only one assumption concerning the subjective rating scale is necessary: the order of confidence criteria on the dimension should be impervious to experimental manipulations (i.e. the rating of *40%* should always be higher than *30%* and lower than *50%*, etc.). We suspect that this assumption is satisfied in a great majority of cases, which makes interpreting *0-100%* JOLs as relative measures of confidence a safe option.

Recalibration and the UWP Effect

In the present study, we used the multi-cycle procedure to create baseline conditions on cycle 1, and then demonstrated that when an experimental manipulation is introduced, the placement of certain JOL criteria on the subsequent cycles can be affected. It seems viable, though, that in this paradigm such a recalibration of the JOL scale due to changes in the range of evidence for future recall occurs naturally even when no changes to the procedure are made between the cycles. We believe that the UWP effect - the finding of impaired calibration with practice - may be one of the manifestations of this process.

Consider the multi-cycle paradigm from the perspective of SDT. On all cycles, participants study and are tested on the same list of word pairs. As the procedure progresses from one cycle to the next, memory

performance for the study list improves. As a result, the two distributions presented in Figure 7.1 shift toward the right end of the scale. Moreover, resolution increases (e.g., Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007; Hanczakowski et al., 2013), which is represented in the SDT model as a gradual decrease in the degree of overlap of the distributions from cycle to cycle, as the distribution of recalled items separates from the unrecalled items distribution. This extends the range of evidence for future recall for the items populating these distributions. As a result, it creates space for the recalibration effects to occur. This is akin to Experiment 2 from our study, inasmuch as the range of evidence is extended upward from cycle to cycle.

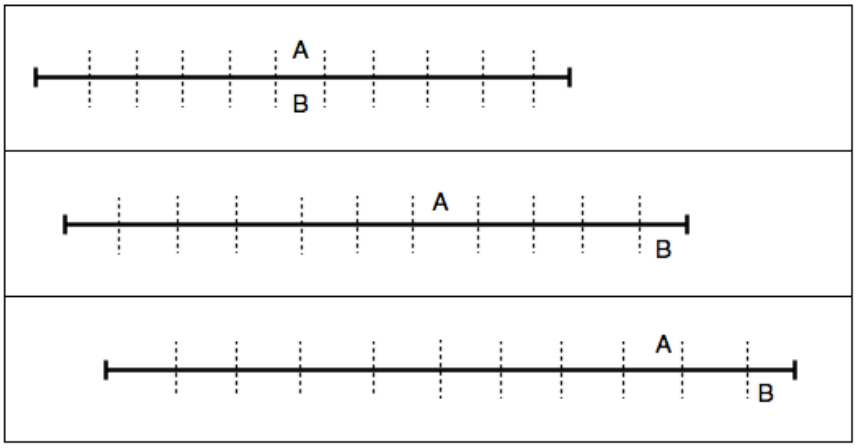

Figure 7.5. A graphical presentation of the UWP effect. Three study-test cycles are shown in the three panels: cycle 1 in the top panel, cycle 2 in the middle panel and cycle 3 in the bottom panel. The horizontal lines in each panel represent the range of evidence for the studied items. The dashed vertical lines represent confidence criteria in increments of 10 (JOL values are not presented for the purpose of clarity).

An example of recalibration at the item level is presented in Figure 7.5. The top line of Figure 7.5 represents the range of evidence for future recall on cycle 1, which can be thought of as baseline. On cycle 2, most items gain evidence for future recall compared to cycle 1. However, this

gain can be greater for some of the studied items; as a result, the range of evidence gets extended (middle line). The same is true for cycle 3 (bottom line).

In order to accommodate this change in the range of evidence, the rating scale may be recalibrated. Consider again items A and B in Figure 7.5 taken from a hypothetical study list. As demonstrated in the top panel, during the first study/JOL phase, the evidence for future recall is comparable for items A and B, so both items get the same rating of *50%*. The evidence for both items increases from cycle 1 to cycle 2, although not to the same extent. Item A gains less than item B, and therefore their ratings diverge: item A gets a rating of *60%*, while item B is now assigned a *100%* rating. As the procedure progresses to the next cycle, evidence for these items changes again. Item B gets strengthened even more, and retains the highest rating of *100%*. Item A this time also gains considerable evidence for future recall, and now the evidence available for this item is comparable to that of item B on the preceding cycle. However, as the range of evidence increased between cycles 2 and 3, more items surpass the evidence for item A on cycle 3 than item B on cycle 2, rendering item A weaker than B between cycles given the changing context of the study list. As a result, the rating assigned to item A is lower than that of item B on the preceding cycle: *80%* compared to *100%*.

Moreover, as the procedure progresses, the number of items occupying the top end of the evidence dimension increases, producing a narrow range of high evidence. If participants want to rank order these items in terms of their evidence for future recall, the more strong items there are, the more fine-grained the distinctions between them need to be. Consequently, criteria for the highest JOL values are drawn toward the top end of the dimension. This may lead to some items with high (but not the highest) levels of evidence being assigned relatively low JOLs, as the higher ratings are reserved for items positioned even further to the right of the dimension.

The key to interpreting the UWP effect in terms of recalibration is to realize that the change in the meaning of scale values from cycle to cycle is not accompanied by any changes in the perceived likelihood of recalling the studied items. In the example above, a participant may be perfectly aware that the evidence for future recall for item A on cycle 3 is identical to that of recalling item B on cycle 2, yet the ratings she uses for these items differ. The result is that the mean of JOLs for items which share the same level of probability of recall should decrease from cycle to cycle. For example, the mean rating for the subset of items the participant correctly thinks are 75% likely to be recalled (assuming for a moment that participants indeed assess the studied items in these terms) could well be *75%* on cycle 1, *70%* on cycle 2 and *65%* on cycle 3. Consequently, if the mean of all judgements is calculated, it falls below that for memory performance. Traditionally, this would be interpreted as underconfidence on cycles 2 and 3. However, this result may simply be an artefact produced by misinterpreting the way in which the rating scale is used.

Note that the recalibration account can potentially explain the differential effects of repeated practice on resolution and calibration. Recall that in the UWP paradigm, resolution improves from cycle to cycle (e.g., Finn & Metcalfe, 2007, 2008; Hanczakowski et al., 2013), while calibration worsens. For calculating resolution, a subjective scale is sufficient. As long as selective criterion shifts do not lead to changes in the ordering of the criteria on the evidence dimension, the measure of resolution should not be affected by the changes in the range of evidence for the studied items. Calibration, on the other hand, requires an objective scale, a requirement that is likely not met, therefore the calibration results cannot be meaningfully interpreted as reflecting under- or overconfidence.

Context effects of this sort may also provide an explanation for why the UWP effect is found with *0-100%* scale JOLs, but not binary JOLs or binary betting decisions (Hanczakowski et al., 2013; Paper 1, this thesis). Specifically, as participants proceed through the cycles, many items

become very well learned, assigned high JOLs, and recalled with near 100% accuracy. In other words, these well-learned, easy items may set a context for judgements assigned to other items in the list in much the same way that the new, easy items in Experiment 2 of the present study set the context for the critical items. That is, the presence of easy, well-learned items in the later cycles of the UWP paradigm may be accommodated by participants shifting their *higher* confidence criteria further *up* the dimension, *lowering* mean JOLs to other items in the list, and producing the UWP effect (i.e., overall mean JOLs assigned to all items in the list would be lowered by the presence of well-earned items on later cycles that have set an easy context). With binary judgements, on the other hand, accommodating the easy items in this manner would be less likely because there is only one criterion ("yes/no" or "bet/no bet"), hence the scale/binary dissociation.

The recalibration account of the UWP effect is also fully compatible with the dominant memory-for-past-test (MPT) account (Finn & Metcalfe, 2007, 2008). Successful recall of an item at the preceding test(s) is undoubtedly a good predictor of future recall (see e.g., Koriat et al., 2002; Paper 1, this thesis), and therefore can strongly contribute to the overall level of evidence for future recall for that item. Previously recalled items should, on average, surpass previously unrecalled items in terms of the strength of this evidence. As the ratings reflect the relative ranking of items within a list in terms of evidence for future recall, the ratings for previously unrecalled items are lower than it would follow from their probability of recall.

The relationship between the recalibration and anchoring (England & Serra, 2012; Scheck & Nelson, 2005) accounts of the UWP effect is more problematic. It is certainly possible to envisage a version of the anchoring account that would predict shifts of only a subset of the criteria and relative stability of the remaining ones, Nevertheless, none of the current formulations of the anchoring account would predict such a pattern of

results. This does not mean, though, that the present results completely dismiss the possibility of an anchor influencing the assignment of JOLs in the multi-cycle procedure, as it is still viable that recalibration and anchoring separately contribute to the magnitude of JOLs.

Implications for Metacognitive Research

The interpretation of the JOL scale as context dependent poses several problems for experimenters employing JOLs in their research. One such problem has already been noted by Hanczakowski et al. (2013). If the rating scale is not objective, calculating calibration becomes a methodological error. As a result, the mere comparison of mean JOLs and recall does not provide meaningful information about how accurate people are in assessing their knowledge. The UWP effect discussed above is one example of such a misinterpretation of the results of a JOL task.

The other problem is more general, as it applies not only to the *0-100%* rating scale used in the present study, but to other rating scales such as *1*-to-*6* as well. As we have demonstrated, an experimental manipulation that changes the range of the rated items has the potential to change the interpretation of the JOL rating scale. In our experiments, it was a substitution of a subset of items within the study list with another subset differing in difficulty, but most likely it is only one of many manipulations capable of exerting such effects. In such a situation, a difference in ratings between groups, which is sometimes calculated (e.g., Connor et al., 1997; Rhodes & Tauber, 2011), may not be indicative of any difference in the overall level of internal confidence. Instead, it may simply be a result of differences in the placement of the criteria. Moreover, sometimes the ranges can differ between tested groups of people even when the procedure they complete is exactly the same. This could stem, for example, from differences in memory capability between groups of participants; age differences or cognitive impairments can potentially serve as examples.

We believe that plotting ROCs to corroborate the results may be a good strategy in such cases. As demonstrated in the present study, ROCs can help distinguish between effects caused by selective criterion shifts and actual changes in internal assessments. In this way, making spurious interpretations of ratings data can potentially be avoided.

It remains an open question whether recalibration effects can be found in other measures used in metacognitive research, such as FOK judgements or judgements of RC, and, if they do, how these effects operate. Recently, recalibration has been proposed as one of potential explanations of the dud-alternative effect on RC. The dud-alternative effect is a finding of inflated confidence in alternatives on a multiple-choice test when an additional, improbable alternative (a *dud*) is presented (Charman, Wells, & Joy, 2011; Hanczakowski et al., 2014; Windschitl & Chambers, 2004). According to the recalibration account of this effect, the presence of an easily rejectable dud alternative on a test trial recalibrates the confidence scale for that trial: low confidence ratings are reserved for the dud alternative, which inflates the ratings for the remaining alternatives compared to a dud-absent condition. As with the recalibration interpretation of JOLs, the recalibration account of the dud-alternative effect assumes that it is the extension of the evidence scale that is responsible for the changes in ratings. Crucially, according to this explanation there is no change in the internal assessment of the non-dud alternatives; the effect is purely attributable to different interpretations of the response scale. Note, however, that at present our understanding of the mechanisms underlying the dud-alternative effect is relatively poor and there are other viable accounts of this effect, some of them assuming a true change in representation of the non-dud alternatives when a dud is presented (e.g., the contrast account; Hanczakowski et al., 2014; Windschitl & Chambers, 2004).

In our view, there is no fundamental difference between the JOL scale and other scales used in metacognitive tasks that would allow for the

recalibration effects to occur only for JOLs, but not for other metacognitive measures. Unless research specifically directed at other task types is conducted, it may be safer to assume that the meaning of FOK or RC scale judgements can similarly be susceptible to experimental manipulations.

Limitations

As we have shown, the signal-detection approach can be a useful tool for distinguishing between differences in ratings stemming from criterion shifts (traditionally thought of as a form of metacognitive control) and changes in perceived level of evidence for future recall of the rated items (reflecting metacognitive monitoring). However, in this paper, we have considered only a case where the placement of a subset of the criteria is influenced by a manipulation, while the remaining criteria remain unaffected, as shown on an ROC. Yet there are other cases that do not allow for such clean conclusions. In theorising on the usefulness of SDT, it has been noted that it is often not possible to distinguish between criterion shifts and concordant distribution shifts (e.g., Goldsmith, 2011; Higham, 2011). A concordant distribution shift requires the two distributions to move in lockstep, preserving the distance between the means. In this way, discrimination - and, consequently, the shape of the ROC - is unaffected. As the placement of the criterion is measured *relative to the distributions*, it does not matter whether it is the criterion or the distributions that change their position on the dimension: in both cases, the measure of criterion placement is affected.

Although the outcome of distribution versus confidence criteria shifts result in the same outcome for bias measures, the two scenarios differ on a fundamental level: one assumes a true difference in perception of the rated items, while the other one does not. Yet these two cases would be undistinguishable on an ROC: both would present as a shift in placement of all criteria. In such a situation, SDT would not provide any additional information regarding the processes underlying the observed pattern of

results. The only reasonable recommendation would be, then, to exert caution while interpreting the data and seek corroborative evidence (e.g., from other tasks assumed to tap the same process) before formulating a strong conclusion.

8. Conclusions

8.1 The interpretation of 0-100% JOLs

The experiments described in this thesis were aimed at establishing the mechanism that drives the discrepancy between the mean of 0-100% JOLs and recall performance - the UWP effect - which is commonly found in procedures consisting of multiple study-test cycles. This was achieved by investigating how 0-100% JOLs assigned on cycles 2 and 3 should be interpreted.

Three interpretations of 0-100% JOLs were proposed. The probability interpretation, assumed by the proponents of the *UWP as a manifestation of true underconfidence* account, assumes that JOLs directly reflect the assessed probability of future recall. Any discrepancy between the mean of JOLs and recall performance stems therefore from imperfect probability assessments. The distorted-rating and ranking interpretations are consistent with the *UWP as an artefact* account. The former interpretation assumes that even though people indeed aim at rating probability in the UWP paradigm, the translation of their internal probability assessments to 0-100% ratings is distorted. Finally, the latter interpretation assumes that, in the UWP paradigm, 0-100% JOLs are not meant as probability ratings; instead, they are used by participants for the purpose of ranking the studied items in terms of evidence for future recall performance.

In the present set of experiments, three different methods were used to investigate the interpretation of 0-100% JOLs in the UWP paradigm. First, in Paper 1, the binary betting task was used in Experiment 2 to establish whether participants are able to track their recall performance with betting decisions both on the test level, and on the level of subsets of items differing in their past recall performance. The results replicated and extended the previous findings of Hanczakowski et al. (2013): betting decisions tracked recall performance very closely for all item types. This good calibration found for binary bets strongly suggests that internal probability assessments are not impaired in the multi-cycle procedure.

This result was therefore inconsistent with the probability interpretation of 0-100% JOLs, according to which repeated study-test phases cause participants to underestimate their future memory performance. If this was true, the UWP pattern should be found not only in the 0-100% JOL task, but in the betting task as well.

The results for the betting task can be accommodated by the distorted-rating and ranking interpretations of 0-100% JOLs. The distorted-rating interpretation would posit that although both binary bets and 0-100% JOLs are supposed to reflect probability of future recall, betting decisions are impervious to some cues that distort scale ratings. For example, anchoring effects, which are the most common artefactual explanation of the UWP effect (e.g., England & Serra, 2012; Scheck & Nelson, 2005), require a scale in order for the anchor to be set.[16] For this reason, anchoring would not be possible for binary bet / no bet decisions. The ranking interpretation of 0-100% JOLs would posit that a scale is necessary in order for fine-grained distinctions between the to-be-rated items to be made. Binary decisions, therefore, are of limited use from the ranking perspective. For this reason, they do not reveal the UWP pattern found for 0-100% JOLs.

The second method used to gain insight into the meaning of 0-100% JOLs was to concentrate on a subset of items for which future recall is easily predictable for participants: namely, items for which recall is at ceiling. In Paper 1, cycle-3 recall exceeded 90% for items recalled on the preceding cycles. In Paper 2, final-cycle recall was close to 100% for items for which contextual details were retrieved. Importantly, these highly

[16] Note that the distorted-rating interpretation of 0-100% JOLs is not the only interpretation which can accommodate anchoring effects. Anchoring can also have the potential to affect the assignment of JOLs if the ranking interpretation is adopted. In that case, the ranking of items according to their evidence for future recall would not be distorted: although the ratings would be drawn toward the anchor, the ordering of items in terms of their evidence should not be affected.

recallable items could further be split into subgroups. In Paper 1, the basis for this split was the number of successful recall attempts on the preceding cycles: some items were recalled on both cycles, while others were recalled only once. In Paper 2, items were split into subgroups on the basis of the number of retrieved contextual details (Experiment 2) or the rated quality and quantity of these details (Experiment 3). Crucially, none of the differences between these subgroups was based on recall performance.

The results of the experiments presented in Papers 1 and 2 consistently showed that participants distinguished with their 0-100% JOLs between item types that were equated in terms of recall performance, but differed in other aspects. This is again not consistent with the interpretation of 0-100% JOLs as probability ratings: if participants are aware of the lack of differences in recallability between these items, JOLs should not differ as well. This time, the distorted-rating interpretation is also unable to accommodate these results. As the probability of future recall does not differ between these items, if factors such as the presence of an anchor distort the ratings so that they no longer match the assessed probability of future recall, they should do so to the same extent regardless of any differences between the to-be-rated items. As a result, 0-100% JOLs for these items should also be equated. This, however, was not the case.

The only interpretation of 0-100% JOLs that can accommodate these findings is the ranking interpretation. According to the ranking account, 0-100% JOLs are used to distinguish between items having different characteristics. Even if recall performance is equated, this ranking can be performed as long as to-be-rated items differ in terms of evidence for future recall. In the present set of experiments, items recalled twice on the preceding cycles or those for which several contextual details were retrieved are likely to have more evidence for future recall than items recalled only once or with only one contextual detail. Therefore the ranking

interpretation would predict higher JOLs assigned to the former than the latter subsets of items - the pattern that was consistently found across the experiments presented in Papers 1 and 2.

Finally, the ROC analysis presented in Paper 3 demonstrated that, in the experimental groups in both experiments described in that paper, the placement of the JOL criteria differed compared to the control groups. This difference in criterion placement was, however, not accompanied by any difference in the perception of the rated items, suggesting that it was the meaning of JOL values that was affected by the experimental manipulation. This is again not consistent with the probability interpretation of 0-100% JOLs, which does not permit any changes in the meaning of JOL values. If JOLs were ratings of probability of future recall, the meaning of these values should by definition always be the same: it should reflect the assessed probability of future recall.

The distorted-rating interpretation of 0-100% JOLs formally does not allow for changes in the meaning of JOL values for the same reason as the probability interpretation, as according to this interpretation participants in a JOL task also aim to assess probability. The question remains whether the systematic distortion that affects the translation of these probability assessments to JOL values can produce effects like those found in the present data. To answer that question, first it is necessary to define the anchoring account in terms of SDT. The basic assumption of this account is that when an anchor is present, the ratings are drawn toward it. This requires repositioning of the JOL criteria. Counterintuitively, in order for the ratings to be *closer* to the anchor value, the criteria need to be placed *further* from the anchor: this means that they become more liberal for values below the anchor, and more conservative for values above the anchor (see Figure 8.1). Assume that the anchor is set around 50%. According to SDT, less evidence would be needed in order to assign values below the anchor - between 10 and 40% - therefore ratings for items falling below the anchor point should increase as compared to a no-

anchor scenario. Similarly, as more evidence would be needed in order for values between 60 and 100% to be assigned, the ratings for items above the anchor point should decrease.

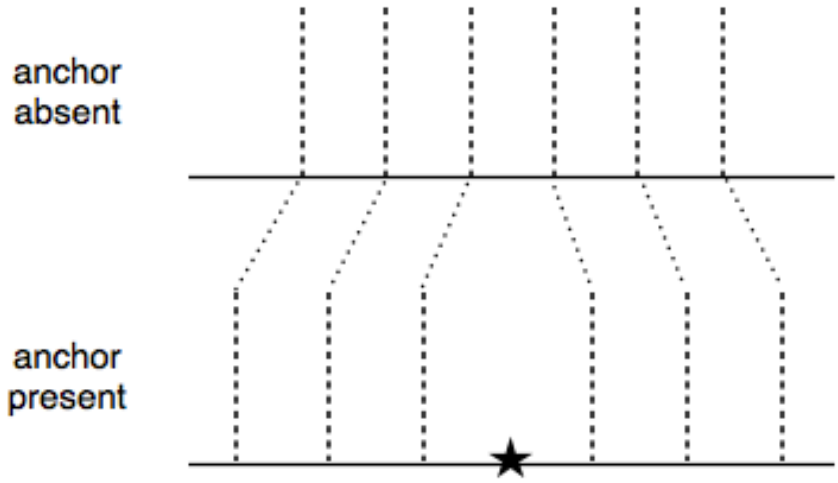

Figure 8.1. A signal-detection representation of anchoring. The top panel presents the positioning of JOL criteria at the evidence-in-future-recall dimension when no anchor is present. The bottom panel presents the same criteria when an anchor (denoted by a star) is set.

As discussed in Paper 3, the results from Experiments 1 and 2 described in that paper are not consistent with the metacognitive contrast account. For this reason, it can be assumed that the assessments of probability of future recall for the critical items did not differ between the experimental and control groups in these experiments. What differed, however, was the range of evidence for the to-be-rated items. Could it be, therefore, that the extent to which criteria are affected by the anchor depends on the experienced range of evidence? It could be postulated that when the range of experienced evidence for future recall increases - for example, with additional learning - the criteria may be pushed by the

anchor even more to the right end of the dimension, as compared to a no-anchor scenario. This would result in precisely the pattern found in Experiment 2 of Paper 3, with the highest criteria becoming more conservative when the range of evidence is greater.

However, the current formulations of the anchoring account of the UWP effect do not postulate any mechanism that would predict such selective criterion shifts when the range of evidence for the to-be-rated items is extended. England and Serra (2012) assumed that the anchor is set on cycle 1, and is equivalent to the mean of JOLs assigned on that cycle. What follows from their assumption is that the value of the anchor across cycles should be relatively stable, even when the range of evidence for to-be-rated items changes between cycles (England and Serra estimated it to be between 20 and 30% in their set of experiments). If this is true, this value should also remain unaffected by experimentally-induced changes in the range of evidence. As a result, there is no reason to predict that the ratings would be distorted by the presence of the anchor to a greater extent with an increase in the range of evidence. This makes the distorted-rating interpretation of 0-100% JOLs not likely.

The ROC results are again consistent with the ranking interpretation of 0-100% JOLs. For this interpretation, the exact meaning of the JOL values is irrelevant: for one person, a value of, say, 60% can serve as a high-confidence rating, while for another it can mean relatively low confidence in future recall. There is only one assumption that needs to be met: that the ordering of the values on the scale does not change. According to the ROCs, this assumption holds in the current data. Therefore, from the theoretical point of view, the results of the experiments described in Paper 3 can be accommodated by the ranking interpretation of JOLs.

Taken together, the present data clearly eliminate the probability interpretation of 0-100% JOLs in the UWP paradigm. This has important consequences for the interpretation of the UWP effect itself. Although both

the probability and the distorted-rating interpretations assume that JOLs are based on estimates of probability of future recall, only the probability interpretation posits that JOLs made by participants directly reflect their probability assessments. If 0-100% JOLs cannot be interpreted as such, then the mean of JOLs that is calculated in UWP studies also cannot be thought of as reflecting the proportion of items predicted to be recalled at a future test. As a result, under- or overconfidence cannot be inferred from the comparison of the mean of JOLs to recall performance. The results are therefore consistent with the accounts of the UWP effect which postulate that this effect is an artefact produced by the use of the percentage scale to elicit JOLs.

Moreover, if the probability interpretation of 0-100% JOLs is dismissed, even calculating the mean of JOLs is incorrect from a methodological point of view. The reason for that is that the probability interpretation is the only one that predicts that the judgements are made on a ratio scale. In the case of the distorted-rating and the ranking interpretations, the 0-100% scale becomes an ordinal one. As a result, only the probability interpretation allows for meaningfully comparing means of 0-100% JOLs between experimental conditions or experiments.

The distorted-rating and ranking interpretations of 0-100% JOLs both assume that the UWP effect should be thought of as an artefact. Of the two, the distorted-rating interpretation seems less likely, as it is clearly inconsistent with the results of Papers 1 and 2 which demonstrated that people distinguish with their JOLs between items sharing the same levels of recall probability, and the interpretation of the results of Paper 3 in terms of this interpretation lack any theoretical underpinning. The ranking interpretation received much stronger support in the present data. All three methods used in the experiments comprising this thesis to establish the meaning of 0-100% JOLs in the UWP paradigm produced results consistent with this interpretation. It can be safely assumed, therefore, that 0-100% JOLs in the multi-cycle procedure are used to distinguish between

items that have different levels of evidence for future recall. As the findings reported in this thesis demonstrate, the criterial amount of evidence needed for the assignment of particular JOL values depends on the context of the study list. The results presented in Paper 3 suggest that the range of the to-be-rated items is one factor that plays a role in how the JOL criteria are distributed. Extending the range of evidence toward one end of the evidence dimension causes the neighbouring criteria to disperse. Another potential factor is also related to the distribution of items on the evidence dimension. If many items share similar levels of evidence for future recall, more fine-grained distinctions may be necessary. In this case, a subset of JOL criteria may be clustered close to each other to enable making these subtle distinctions.

8.2 Recalibration in the UWP paradigm

In this thesis, a new, recalibration account of the UWP effect was proposed. The recalibration account belongs to the group of artefactual accounts of the UWP effect. It sees this effect as stemming from changes in meaning of JOL values that take place from cycle to cycle, resulting from changes in the context of the study list. This occurs for two reasons. First, the range of evidence for the to-be-rated items increases from one cycle to the next, as some items gain in strength more than the rest. Second, the number of items with strong evidence increases with repeated studying and testing. This forces the JOL criteria to be readjusted in order for participants to be able to successfully rank order the items in terms of evidence for future recall. In simple terms, according to the recalibration account, 0-100% JOL criteria are placed where they are needed the most.

The recalibration account is able to account for several notable findings present in the UWP literature, including those seemingly inconsistent with other accounts of the UWP effect. For instance, it can be used to explain the results of Koriat (1997; see section 2.1.1) who demonstrated the UWP effect in a one-cycle procedure by manipulating

list composition. In Experiment 3, he manipulated the number of presentations of each item: items were presented once, twice, or three times within the study list. This resulted in the UWP pattern for items presented twice and thrice, while for items presented only once recall and JOLs matched. In Experiment 4, Koriat varied presentation times, with items presented for two, four or eight seconds. This manipulation produced the UWP effect only for items with longer presentation times (4 s and 8 s). An interesting aspect of these findings is that they cannot be accommodated by the MPT account of the UWP effect (Finn & Metcalfe, 2007, 2008), as there is no past test performance to be used as a cue for JOLs.

From the perspective of the recalibration account, manipulations such as those used by Koriat (1997) produce a wide range of evidence for future recall. As the JOL criteria have to be set in a way that would allow for ranking items on all levels of evidence, the high criteria need to be placed relatively far up the evidence dimension. Moreover, as two thirds of the studied items become strongly encoded (either by repetition or by long presentation times), there is greater need for making fine-grained distinctions between these strong items than between weakly encoded items placed lower at the dimension. As demonstrated in Experiment 2 from Paper 3, such an extension of the range of the dimension covered by the criteria to the right and the clustering of high criteria at the far end of this range are able to increase the discrepancy between the mean of JOLs and recall performance. In Koriat's data, this might have produced the UWP effect for the items that were repeated or presented for longer.

Similarly, recalibration can explain findings traditionally attributed to anchoring. As discussed in detail in section 2.2.1, Scheck and Nelson (2005) demonstrated that in the immediate JOL task, while the UWP effect was present for easy items, it disappeared for difficult items. The easy and difficult items were taken from the Nelson and Dunlosky (1994) norms. According to these norms, mean performance for easy items was 21% on

cycle 1 and 53% on cycle 2, while difficult items had mean performance of 4% on cycle 1 and 21% on cycle 2. Therefore the increase in recall levels from cycle to cycle was supposed to be greater for easy than for difficult pairs (31 versus 17 percentage points). In the actual data this difference was in the same direction, albeit slightly lower (approximately 35 versus 25 percentage points). This likely increased the range of experienced evidence for future recall from cycle 1 to cycle 2, creating space for recalibration effects. To accommodate the whole range of evidence, high criteria might have been made more conservative, lowering the mean of JOLs for easy items.

The interpretation of the results from past UWP studies presented above in terms of recalibration is, of course, based on the generalisation of the present data and as yet untested. Nevertheless, it generates predictions that can be verified with the help of signal detection methods, which can be done in future studies.

The next step is to establish the relationship between the recalibration account and other artefactual accounts of the UWP effect. According to Hanczakowski et al. (2013), the MPT account can be interpreted as supporting the UWP-as-an-artefact view. The authors called it the *confidence interpretation* of MPT. This interpretation assumes that participants use information about their past test performance as a cue for 0-100% JOLs. However, the particular values assigned to previously recalled and unrecalled do not play a role: the only requirement is that previously recalled items should get higher ratings than previously unrecalled items. It is therefore consistent with the ranking interpretation of 0-100% JOLs promoted in this thesis.

As mentioned in the General Discussion of Paper 3, the recalibration account of the UWP effect can easily accommodate the use of the MPT heuristic. As suggested by the memory for past performance results described in Paper 1, when queried, people can access the information regarding not only their performance on the last test, but even on the test

before the last. If this information is retrieved at the time of making a JOL, it can add to the overall volume of evidence for future recall. It can therefore be useful for ranking items in terms of this evidence, especially when there are many high-evidence items and thus more fine-grained distinctions need to be made.

Hanczakowski et al. (2013) also distinguished between two versions of the anchoring account of the UWP effect, both of which are consistent with the UWP-as-an-artefact view. The *probability* interpretation of anchoring requires an assumption that participants in the multi-cycle JOL task attempt to rate the probability of future recall, but the anchor distorts the ratings, producing the UWP pattern. Note that this interpretation of anchoring is compatible with the distorted-rating interpretation of 0-100% JOLs presented in this thesis. As the present results do not support this interpretation of JOLs, the probability version of the anchoring account will not be further discussed. The *confidence* interpretation of anchoring, on the other hand, assumes only that participants in the JOL task attempt to rank the items, and the particular values they use in order to achieve this goal are affected by the presence of an anchor. It is therefore compatible with the ranking interpretation of 0-100% JOLs.

Could the recalibration results be thought of as a mere instantiation of anchoring under its confidence interpretation? Certainly it is possible to put forward such an anchoring explanation of the results presented in Paper 3. However, it would suffer from the same limitation as the distorted-rating interpretation of these results described above. Namely, adopting this explanation would require an assumption that the extent to which the anchor affects the ratings depends on the range of experienced evidence for future recall. So far, none of the existing versions of the anchoring account would predict such a result.

Although the anchoring account, at least in its current formulation, cannot explain by itself the data presented in Paper 3, it may still be possible to reconcile the recalibration and anchoring explanations of the

UWP effect. It is viable that both effects could operate at the same time in the multi-cycle paradigm. Anchoring would affect the overall setting of the JOL criteria, by pushing them away from the anchor. Recalibration would be responsible for the setting the criteria in such a way that would allow for ranking the to-be-rated items in terms of their evidence for future recall, for example by adjusting them to the range of these items. When applied to the results of Experiment 2 from Paper 3 (in which the magnitude of the UWP effect was greater in the experimental than in the control group), both anchoring and recalibration would be responsible for the presence of the UWP effect in both groups, while recalibration would further increase the discrepancy between the mean of JOLs and recall performance between the groups.

However, before adopting the anchoring-and-recalibration account of the UWP effect, it has to be taken into account that the recalibration account of UWP has one substantial advantage over the anchoring explanation: it can be experimentally verified. To recapitulate the arguments presented in section 2.2.1, the main weakness of the anchoring account lies in its circularity. In the past studies concerned with anchoring, the existence and placement of the anchor was inferred from the obtained pattern of results. It is therefore not surprising that the same pattern of results can be, in turn, explained by assuming that the anchor is present. This is linked to another serious drawback of this approach: in JOL research, there seems to be no objective way of assessing if the ratings are truly influenced by an anchor. The recalibration account, on the other hand, generates predictions that can be verified by using the signal-detection approach. Moreover, the rating interpretation of 0-100% JOLs, which is assumed by the recalibration account, has been confirmed experimentally in the experiments described in this thesis. For these reasons, it is safer to subscribe to the purely recalibration account, which can explain the same data as the anchoring account without making unverifiable assumptions.

8.3 Summary

The present thesis introduces a novel account of the UWP effect - the recalibration account. This new account is based on the assumption that it is confidence, not probability, that people rate in the multi-cycle 0-100% JOL task. By rejecting the probability interpretation of 0-100% JOLs, the present findings also reject the accounts of UWP that interpret the UWP pattern as a manifestation of psychological underconfidence (Finn & Metcalfe, 2007, 2008; Koriat, 1997; Koriat et al., 2002; Koriat et al., 2006). The rejection of the probability interpretation of JOLs does not rule out automatically a possibility that people become underconfident with practice. However, even if this was true, the UWP pattern does not require true psychological underconfidence in order to be found in the data.

If the ranking interpretation of 0-100% JOLs, consistently favoured by the present data, is adopted, comparing the mean of JOLs to recall, aside from being methodologically incorrect, does not give any insight into the accuracy of metacognitive assessments. From the perspective of ranking to-be-rated items in terms of *evidence* for future recall, it does not matter how good people are in assessing the *probability* of future recall. Overall overconfidence, underconfidence or realism of these probability assessments should make no difference for the ranking of items. In other words, 0-100% JOLs can be informative when it comes to assessing resolution (which is based on rank-ordering the to-be-rated items), but are not suitable for assessing calibration. Consequently, the UWP effect, which is based on calibration data, should be interpreted simply a misinterpretation of experimental results caused by improper interpretation of the rating scale.

9. Appendices

APPENDIX A

Unrelated word pairs used in all experiments

CUE	TARGET	CUE	TARGET	CUE	TARGET
AGENT	SILENCE	EMOTION	CRIPPLE	SENATOR	SWING
AMOUNT	REVENGE	GLIMPSE	MARBLE	SHARE	CAMERA
ANKLE	MANAGER	GRAPE	COMFOR	SHELL	STOUT
APPLE	ELEMENT	GRAVY	SERVANT	SHOWER	MATCH
AWARD	GOSSIP	GRUDGE	PULSE	SKETCH	RIVER
BATTERY	ADVICE	ILLNESS	GUILT	SPEECH	CELLAR
BLANKET	COUNCIL	INCOME	REVIEW	STABLE	ADDRESS
BUBBLE	SCENT	MUSCLE	PERCH	STATUE	BREAD
CARGO	ISLAND	NEEDLE	FUNERAL	STEER	OUTLINE
CEILING	ESCAPE	NURSE	MUSEUM	STREAM	WEAPON
CELERY	FUTURE	PARENT	SCOUT	STRIPE	SPORT
CHECK	GLOVE	PASSAGE	HOBBY	TEMPER	ANCHOR
CHILL	FLOOD	PATRON	ELDER	THUNDER	PASTRY
CLINIC	BUTCHER	PITCHER	MODEL	TOURIST	SHAPE
COLONY	ORDEAL	POLICE	VERSE	TRACTOR	WOUND
COMMENT	SMOKE	PROFIT	SPACE	TREATY	PLEDGE
DELAY	CRUSH	PUDDING	SUBWAY	VILLA	POWER
DENTIST	TORCH	RABBIT	METHOD	WAGON	PILLAR
DEPOSIT	DRAPERY	REPAIR	CROOK	WARRAN	NATURE
DESPISE	BLADE	RESERVE	LOVER	WRENCH	INSECT

APPENDIX B

New word-word and nonword-word pairs used in Experiment 1 (Paper 3)

Cycle 2		Cycle 3	
CUE	TARGET	CUE	TARGET
ALARM	VISION	JEWEL	CHAMBER
BALANCE	JUNGLE	MISSION	OPINION
BOILER	SEASON	NOVEL	FELLOW
BUTTON	SHELTER	PACKAGE	CONTACT
CHARM	CRACKER	PARADE	SHEEP
CLUSTER	SHOCK	PIGEON	LAYER
FERRY	SHRUB	POTATO	CORNER
FLAME	COMPANY	RIVAL	MARGIN
GALLERY	MOUTH	EAGLE	WEATHER
MONKEY	LEGEND	MEETING	VANILLA
RECEIPT	BARRIER	QUEST	PLANT
SIRUP	COLLAR	SIREN	CREAM
STAIN	TENNIS	SYSTEM	RAILWAY
SUGAR	CHANT	TRUTH	SISTER
TOWEL	GARAGE	COFFIN	CEMENT
TRAGEDY	RADIO	JACKET	COPPER
TRUNK	MASTER	LIQUID	TERRACE
VICTORY	MANKIND	LUGGAGE	STATUS
WAITER	DAMAGE	MARKET	PENSION
WINNER	ENGINE	ORCHARD	BRIDGE

Cycle 2		Cycle 3	
CUE	TARGET	CUE	TARGET
ARMAN	CONCERT	BINICAL	TRAIL
BLISSEN	CLOCK	CRABLE	FACTOR
CALIDON	SPEAKER	PRAMIS	ANGER
CAMENT	CREDIT	ROGATION	CABINET
FISSEL	SLOPE	TRESPAT	SWAMP
GARDER	CHASE	BELLAND	SCIENCE
HALBERT	SALMON	BRENDER	CRIME
HENSION	CHARITY	CORBIT	EMPIRE
MANIPER	CRYSTAL	DELICON	LEADER
MESTIC	FLESH	GRAMEN	FAIRY
PASSET	STRAP	TAMID	FINDING
PLANDER	IVORY	WAVEN	TORTURE
POTIMER	ACCORD	BECKLE	CRAFT
PURDEN	MEASURE	BLINDEN	STROLL
SCULLET	LIQUOR	CLORAL	EDITION
SUBBEN	BUTLER	FLEMIN	LIMIT
TARRION	BASKET	LOMAND	INSULT
TUMMEL	FICTION	SENDAL	CARPET
WIDICOM	STORM	SONDER	POUND
WIMBER	ROUTINE	VISARY	STACK

References

Allwood, C. M., Ask, K., & Granhag, P. A. (2005). The cognitive interview: Effects on the realism in witnesses' confidence in their free recall. Psychology, Crime & Law, 11, 183–198.

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39,* 171-184.

Ariel, R., Hines, J. C., & Hertzog, C. (2014). Test framing generates a stability bias for predictions of learning by causing people to discount their learning beliefs. *Journal of Memory and Language, 75,* 181-198.

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative mnemonic accuracy. In J. Dunlosky, R. A. Bjork (Eds.). *Handbook of Memory and Metamemory* (pp. 73-94). New York: Psychology Press.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55-68.

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgments. *Journal of Personality and Social Psychology, 60,* 485-499.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34,* 918-928.

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review, 14,* 107-111.

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition, 41,* 28-35.

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law & Human Behavior, 35,* 479–500.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12,* 50–71.

Daniels, K. A., Toth, J. P., & Hertzog, C. (2009). Aging and recollection in the accuracy of judgments of learning. *Psychology and Aging, 24,* 494-500.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2012). Adaptation to cognitive context and item information in the medial temporal lobes. *Neuropsychologia, 50,* 3062-3069.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2013). Parahippocampal cortex activation during context reinstatement predicts item recollection. *Journal of Experimental Psychology: General, 142,* 1287-1297.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523-533.

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33,* 1096-1115.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22,* 271-280.

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review, 19,* 715-722.

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language, 64*, 289-298.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 33,* 238-244.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58,* 19-34.

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141,* 124-133.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition,* 16**,** 309-313.

Goldsmith, M. (2011). Quantity-Accuracy Profiles or type-2 signal detection measures? Similar methods towards a common goal. In P. A. Higham, J. P. Leboe (Eds.) *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea* (pp. 128-136). Basingstoke: Palgrave MacMillan.

Hanczakowski, M., Zawadzka, K., & Higham, P. A. (2014). The dud-alternative effect in memory for associations: Putting confidence into local context. *Psychonomic Bulletin & Review, 21,* 543-548.

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language, 69,* 429-444.

Hertzog, C., Saylor, L. L., Fleece, A. M., & Dixon, R. A. (1994). Meta-memory and aging: Relations between predicted, actual and perceived memory task performance. *Aging and Cognition, 1,* 203-237.

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson & Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30,* 67-80.

Higham, P. A. (2011). Accuracy discrimination and type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In P. A.

Higham and J. P. Leboe (Eds.), *Constructions of Remembering and Metacognition*: *Essays in Honour of Bruce Whittlesea* (pp. 109-127). Basingstoke: Palgrave MacMillan.

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (in press). *Internal mapping and its impact on measures of absolute and relative metacognitive accuracy*. Chapter to appear in J. Dunlosky and S. K. Tauber (Eds.) *Oxford Handbook of Metamemory*.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology Learning, Memory, and Cognition, 18,* 667-680.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3-28.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469-486.

Keren, G., & Teigen, K. H. (2001). Why is p = .90 better than p = .70? Preference for definitive predictions by lay consumers of probability judgments. *Psychonomic Bulletin & Review, 8*, 191–202.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., et al. (2010). Control and interference in task switching - A review. *Psychological Bulletin, 136*, 849-874.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119,* 80-113.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 187–194.

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval fluency. *Memory & Cognition, 34,* 959-972.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643-656.

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 595-608.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131,* 147–162.

Kornell, N., Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138,* 449-468.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22,* 787-794.

Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young children are not underconfident with practice: the benefit of ignoring a fallible memory heuristic. *Journal of Cognition and Development, 13,* 174-188.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman– Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 35, 509–527.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 1263-1274.

McCabe, D. P., Roediger, H. L., & Karpicke, J. D. (2011). Automatic processing influences free recall: Converging evidence from the process dissociation procedure and remember-know judgments. *Memory & Cognition, 39,* 389-402.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113,* 123–132.

Metcalfe, J., Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: General, 134,* 1084-1097.

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239-257.

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,* 267-270.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

Pansky, A., & Goldsmith, M. (2014). Metacognitive effects of initial question difficulty on subsequent memory performance. *Psychonomic Bulletin & Review.* Advance online publication.

Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86,* 777-803.

Pyc, M. A., & Rawson, K. A. (2012). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition, 40,* 976-988.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory* & *Cognition,* 21, 89-102.

Rast, P., & Zimprich, D. (2009). Age differences in the underconfidence-with-practice effect. *International Aging Research: An International Journal Devoted to the Scientific Study of the Aging Process, 35,* 400-431.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology. General, 137*, 615–625.

Rhodes, M. G., & Tauber, S. K. (2011). Monitoring memory errors: The influence of the veracity of retrieved information on the accuracy of judgments of learning. *Memory, 19,* 853-870.

Roediger, H. L., & Karpicke, J. D. (2006). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Sci*ence, 17, 249–255.

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember–know judgments using signal detection and threshold models. *Memory & Cognition, 34,* 1598–1614.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review, 12,* 865-873.

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct–perceptual competence. *Psychological Review, 107,* 525–555.

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness identification. *Law & Human Behavior, 34,* 337–347.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51,* 71–79.

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134,* 124–128.

Serra, M., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1258–1266.

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgments of remembering and forgetting. *The Quarterly Journal of Experimental Psychology, 65,* 2231-2257.

Skavhaug, I.-M., Wilding, E. L., & Donaldson, D. I. (2013). Immediate judgments of learning predict subsequent recollection evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* 159-166.

Susser, J. A., Mulligan, N. W., & Besken, M. (2013). The effects of list composition and perceptual fluency on judgments of learning. *Memory & Cognition, 41,* 1000-1011.

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases of younger and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27,* 474-483.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185,* 1124-1130.

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*, 1127-1131.

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 198–215.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025-1054.

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General,* 130, 361-379.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441-517.