Journal of Memory and Language 90 (2016) 161–176

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Recalibration effects in judgments of learning: A signal detection analysis ☆

Katarzyna Zawadzka [a,b,*], Philip A. Higham [a]

[a] Psychology, University of Southampton, UK
[b] School of Psychology, Cardiff University, UK

ARTICLE INFO

Article history:
Received 31 July 2015
revision received 12 April 2016

Keywords:
Judgments of learning
Metacognition
Signal detection theory

ABSTRACT

In this study we investigated the influence of list composition on judgments of learning (JOLs). To this end, we compared JOLs assigned in a multi-cycle procedure to a set of moderately difficult word pairs. Experiment 1 revealed that when difficult new pairs were added to the study list, the mean of JOLs assigned to the moderate pairs increased as compared to the baseline. In Experiment 2, we reversed this pattern by including easy new pairs in the study list. By analyzing metacognitive ROCs (MROCs), we demonstrate that these results were caused by criterion shifts, by which participants adjusted the level of evidence needed to assign particular JOL ratings. Changes in the study list composition led to a recalibration of the JOL scale – i.e. resetting of the criteria – in order to accommodate the addition of new items. We discuss the usefulness of MROCs for detecting criterion shifts in rating tasks.

Introduction

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) judgments, and prospective measures such feeling-of-knowing (FOK) judgments or judgments 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.[1] 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 judgment; correctly recognized vs. not correctly recognized on a recognition test following an FOK judgment, etc.). For resolution, the absolute magnitude of judgments

☆ The authors would like to thank Maciej Hanczakowski and Greg Neil for their helpful comments concerning this research.

* Corresponding author at: Division of Psychology, Nottingham Trent University, Burton Street, Nottingham NG1 4BU, UK.
E-mail address: katarzyna.zawadzka@ntu.ac.uk (K. Zawadzka).

[1] Throughout the paper, scale labels are italicized.

http://dx.doi.org/10.1016/j.jml.2016.04.005
0749-596X/

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 can be calculated from an ordinal scale, and a subjective *1*-to-*6* scale satisfies this requirement.

The same metacognitive ratings can also be elicited on objective scales, such as *0–100%* 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*. 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 *0–9*, *10–19*, ..., *90–99*, *100%* and then ratings and percentage correct are compared at each level), or for the whole test. In both cases, perfect calibration (or realism) requires that the means corresponding to objective performance are equal to mean ratings assigned to the items. 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%* JOL 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%* JOL scale have been formulated in the JOL literature. Recently, Hanczakowski, Zawadzka, Pasek, and Higham (2013; see also Higham, Zawadzka, & Hanczakowski, 2016; Zawadzka & Higham, 2015) cast doubt on the likelihood interpretation of percentage JOLs. Their research concerned the underconfidence-with-practice (UWP) effect (see, e.g., Finn & Metcalfe, 2007, 2008; Koriat, Sheffer, & Ma'ayan, 2002), an impairment of calibration present when the same materials are studied and tested more than once. In a typical UWP experiment, participants first study a list of (typically unrelated) cue–target pairs such as *digit-hunger*. During study, they assign JOLs to each item to indicate how likely it is that they will later remember the target of the pair if provided with the cue on an immediate cued-recall test following study. Following the list, a recall test is administered and performance on this test is compared to JOLs assigned during study. On this first test, participants are typically well calibrated or there is slight overconfidence. Following the first test, the entire procedure is repeated at least once so that the whole experiment consists of two or more identical study-test *cycles*. However, unlike the results from the first cycle, from the second cycle onward, participants are typically underconfident; that is, their JOLs underestimate their actual recall.

Hanczakowski, Zawadzka, et al. (2013) noted that the UWP effect was independent of the instructions given to participants regarding the interpretation of JOLs. In most studies participants were cued at study with a prompt asking them to rate the likelihood of recalling the target at test, such as "With what probability will you remember the target word in about five minutes from now if you see the cue word?" (Rast & Zimprich, 2009). Instructions like these should, at least in theory, convey to participants that the JOL task is in fact a probability rating task, and so the JOL scale is an objective one, with JOL values indicating assessed probability of recall. However, some researchers have used JOL prompts that did not mention the constructs of probability or likelihood at all, and asked instead about confidence (e.g., Scheck & Nelson, 2005; Serra & Dunlosky, 2005). Nevertheless, despite the fact that the likelihood and confidence prompts are profoundly different on a theoretical level, there was no difference in the accuracy (as assessed by calibration) of likelihood- and confidence-prompted JOLs.[2] This led Hanczakowski, Zawadzka et al. to question whether participants in the percentage JOL task were really aiming to maximize calibration. If they were not, this would be consistent with findings from the judgment and decision making literature suggesting that participants do not aim at assessing calibration even if they are provided with direct instructions to do so and examples of what calibration entails (Keren & Teigen, 2001; Lichtenstein & Fischhoff, 1981).

For this reason, Hanczakowski, Zawadzka, et al. (2013) decided to assess the generalizability of the UWP effect to different rating types, such as binary *yes*/*no* JOLs and binary betting decisions.[3] They argued that if the UWP effect was found with ratings other than *0–100%* JOLs, it would be consistent with the claim that this effect reflects inaccurate assessments of the likelihood of future recall. However, what Hanczakowski, Zawadzka et al. found is that, in contrast to the underconfidence observed with the percentage-JOL scale, the proportion of "yes" responses on later cycles with the binary tasks did not differ from the proportion of correctly recalled items, revealing good calibra-

[2] Luna, Higham, and Martín-Luengo (2011) observed similar correspondence between likelihood ratings and RC ratings in a retrospective task.

[3] 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 actually recalled.

tion. Although they did not offer a full explanation of the dissociations observed with the different scales, their findings suggested that participants were assigning low JOLs to items they believed they would ultimately recall (see also Zawadzka & Higham, 2015).

Following up on research by Hanczakowski, Zawadzka, et al. (2013) and Zawadzka and Higham (2015) investigated the assignment of the highest JOLs in a procedure consisting of three study-test cycles. 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, when participants were given a binary betting task instead of the *0–100%* scale-JOL task in Experiment 2 of Zawadzka and Higham, they were able to correctly predict future recall with their bets. This demonstrates that even though participants were aware that recall would 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 percentage JOLs (evinced by the percentage-JOL data).

The findings of Hanczakowski, Zawadzka, et al. (2013) and Zawadzka and Higham (2015) suggested that participants were assigning low JOLs to items they believed they would ultimately recall. This, in turn, suggested that it might be more accurate to interpret JOLs as confidence judgments rather than assessments of likelihood. Confidence judgments differ from likelihood judgments in one important aspect: the scale on which confidence judgments are made may well be subjective, whereas it is not with likelihood judgments. Subjective scales are not conducive to assessment of calibration because the scale values have no absolute meaning; it is not possible to conclude that participants are realistic if items assigned a rating of *40%* have a 40% recall probability any more than if those same items were assigned *4* on a six-point scale.

There are also other reasons suggesting that the *0–100%* JOL scale may not satisfy the objectivity assumption. Dunlosky, Serra, Matvey, and Rawson (2005) reached the same conclusion as Hanczakowski, Zawadzka, et al. (2013) and Zawadzka and Higham (2015) by examining second-order judgments (SOJs) about JOLs. In their study, for each studied word pair participants were asked to assess the likelihood of recalling the target when presented with the cue on a *0–100%* scale, and then, on the same scale, assess their confidence in the accuracy of that JOL – a SOJ. The results revealed markedly lower SOJs for intermediate JOL values than for extremes of the JOL scale. Dunlosky et al. argued that this dissociation between JOLs and SOJs reflects two separate processes that are involved in assigning JOLs. The first process leads to a binary yes/no decision being made regarding the predicted retrieval outcome at test. This yes/no decision is then followed by accumulation of supporting evidence. The more evidence for or against future recall that is gathered during this stage, the higher the final JOL. According to this interpretation, a JOL of *60%* does not reflect a 60% probability of future recall, but rather relatively weak evidence in favor of future recall success.[4]

How, then, should the *0–100%* JOL scale be interpreted? We believe that analyzing JOLs from a signal detection theory (SDT) perspective can be useful in answering this question. Although interpreting metacognitive judgments in terms of SDT is still relatively rare, it is not unheard of (e.g., Ferrell & McGoey, 1980; Hanczakowski, Pasek, Zawadzka, & Mazzoni, 2013; Hanczakowski, Zawadzka, et al., 2013; Higham, 2007, 2011, 2013). In the remainder of this paper, we describe a signal-detection account of the *0–100%* JOL scale (Benjamin & Diaz, 2008; Masson & Rotello, 2009) and assess the consequences of such an approach.

A signal-detection account of JOLs

Fig. 1 shows a signal-detection representation of the *0–100%* JOL task. Two distributions of studied items are positioned on an evidence-for-future-recall dimension.[5] 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 the distribution of later unrecalled items (i.e., further up the evidence dimension). 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 varies 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.

[4] Note that this interpretation of the *0–100%* JOL task could explain why binary *yes/no* judgments fare better at predicting future recall (Hanczakowski, Zawadzka, et al., 2013), as well as reduce dual-task costs stemming from the requirement for concurrent learning and providing monitoring judgments (Mitchum, Kelley, & Fox, 2016), as compared to *0–100%* JOLs.

[5] Here 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).

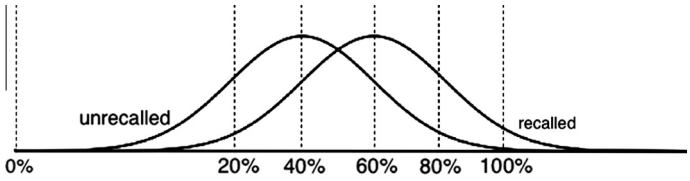

Fig. 1. A signal-detection representation of the JOL task. Two distributions are placed on the evidence-for-future-recall dimension. The distributions on the left versus the right represent items not recalled versus recalled on a given cycle, respectively. Vertical lines denote separate criteria. For 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.

The signal detection approach allows metacognitive receiver operating curves (MROCs) to be plotted. A MROC is a isosensitivity curve (i.e., resolution is the same at all points on the MROC) that displays the relationship between *hit rates* (HRs) and *false alarm rates* (FARs). An example of a MROC is presented in Fig. 2. To generate such an MROC, 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 (HR) is plotted against the proportion of unrecalled items which were assigned a given JOL or higher (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 16 out of 20 correctly recalled items, then the coordinates of the point on the MROC corresponding to the value of *40%* would be (0.4, 0.8). These points on the MROC 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 unit square that falls below the MROC 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 MROC points), the MROC follows the minor diagonal of the plot, and the AUC equals 0.5.

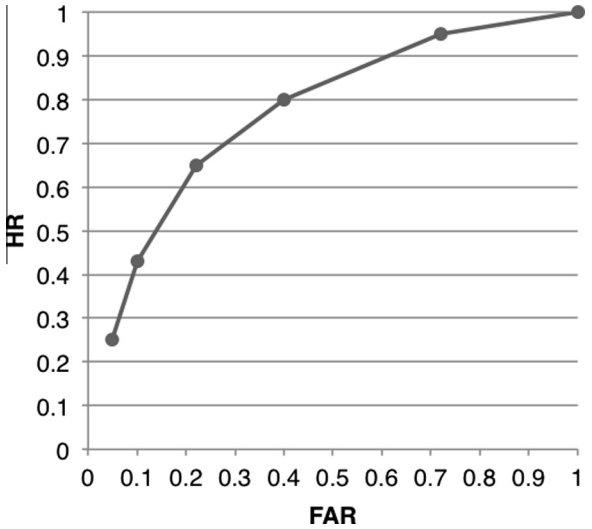

Fig. 2. An example of a metacognitive receiver operating characteristic (MROC) curve. Points on the curve denote separate confidence criteria. The most liberal criterion (*0%*) is placed in the top right corner, while the most conservative criterion (*100%*) is at the end of the curve in the bottom left part of the figure. HR = hit rate; FAR = false alarm rate.

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 map subjective confidence onto the percentage-JOL 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. For example, Susser, Mulligan, & Besken (2013; Experiment 1) followed up on Rhodes and Castel's (2008) research on the effect of font size on JOLs and recall. Previously, Rhodes and Castel 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 was limited to mixed lists that consisted of both items presented in a small and large font. When the list was pure (i.e., consisting solely of items shown in either a large or in a small font), the effect of font size on JOLs disappeared. Similarly, Zawadzka and Higham (2015, Experiment 1) revealed differences in JOLs for previously recalled items despite equated recall performance (see above).

Context effects on JOLs such as those observed by Susser et al. (2013) suggest that the *range of evidence* for future recall, which would have been greater in the mixed-compared to the pure-study-list condition, might be critical to understanding calibration. To illustrate how the SDT model described above might account for such effects using an example more similar to our own research, consider two groups of participants who are presented with a list of word pairs. For the first *moderate-narrow-range* group, the list consists of pairs of moderate difficulty only (henceforth referred to as *critical* pairs). For the sec-

ond *hard-wide-range* group, there are both critical and difficult word pairs on the list. For example, the moderate word pairs might consist of unrelated words (*muscle-rainbow*), whereas the hard word pairs might consist of nonwords (*cament-fissel*). In both cases, participants provide immediate JOLs for each of the word pairs, and later they are given a cued-recall test.

Fig. 3 depicts these two scenarios. In the moderate-narrow-range group (middle panel), two distributions are positioned toward the center of the evidence-for-future-recall dimension with the distribution of unrecalled items on the left, and the distribution of recalled items on the right. In the hard-wide-range group (top panel), there are two distributions of unrecalled items: one for critical pairs, and one for difficult pairs. (For simplification, we assume that none of the targets from 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, because, 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 in the other direction by adding easy new pairs (e.g., related pairs such as *doctor–nurse*) to create an *easy-wide-range* group. This scenario is presented in the bottom panel of Fig. 3. (For simplification, we assume that all easy word pairs were recalled, so there is no distribution for unrecalled easy word pairs.) This time, the new, recalled, easy-item distribution is located to the right of the recalled critical-item distribution, because, on average, evidence is greater for the easy than for the critical pairs. The addition of new, easy items again extends the experienced range of evidence as compared to the pure list of moderate critical word pairs 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. However, rather than *all* the criteria being adjusted, the adjustment is limited to those criteria in the relevant range of evidence. This selective adjustment of criteria is shown in Fig. 3 by comparing the top and middle panels. Note that only the lower criteria (those associated with JOLs of 20%, 40%, and 60%) are decreased to accommodate the low evidence of the hard items. The upper criteria remain static. On the other hand, a comparison of the middle and bottom panels of Fig. 3 shows the opposite criteria adjustment pattern; that is, only the higher criteria (i.e., those associated with 60%, 80%, and 100%) are increased to accommodate the inclusion of easy items, whereas the low criteria remain static. By making selective criterion adjustments in this way, different JOL values can be used to effectively discriminate between items that differ in evidence for their future recall.

Experimental overview

In two experiments, we tested the prediction that the assignment 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 lower 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 higher 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 the control conditions, in which no new items were added, changes in the JOL mean in the experimental group will necessarily affect the magnitude of the difference between the JOL and recall means – 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. On the other hand, in Experiment 2, this apparent underconfidence should increase in the experimental condition compared to the control 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 and tested on the same 60 pairs as on cycle 1. In the experimental condition, only 20 *critical pairs* were the same as those on cycle 1 (and thus

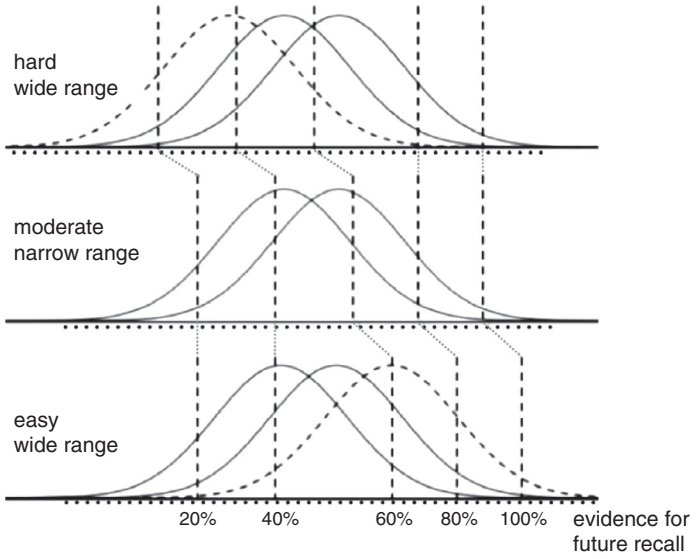

Fig. 3. A graphical illustration of predictions of the bias account. Solid curves represent distributions for critical items, with the unrecalled-item distribution on the left, and the recalled-item distribution on the right. The dotted line below the evidence dimension represents the experienced range of evidence. The middle panel (moderate, narrow range) represents a case in which only critical items of moderate difficulty are studied. The top panel (hard, wide range) includes, in addition to critical items, a dashed leftmost distribution of very difficult new items. The dashed rightmost distribution in the bottom panel (easy, wide range) represents the easy new items distribution. Vertical dashed lines represent confidence criteria. The bias account predicts shifts only of criteria that are relatively 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.

were the same as in the control condition), whereas 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. Before the study phase began, participants were presented with instructions for the JOL task:

> After the presentation of each pair you will be asked to assess on a scale from **0** to **100%** the **probability** that you **will recall** at test the second word from this pair when presented with the first word. For example, if you are certain that you will recall this word, choose '100%'. If you are certain that you will not be able to recall this word, choose '0%'.

At study, all pairs were presented individually 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. They could type in any value from *0%* to *100%*. Time for providing the judgment was not limited.

The test immediately followed the study phase. All pairs studied on a 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 1. Resolution scores are presented in Table 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) × 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.

It was important to establish that performance for the critical pairs, that were the focus on the main analyses reported below, was also comparable between the experimental and control groups in cycle 1. To establish this comparability, we conducted the same 2 × 2 ANOVA on cycle-1 JOL and recall results, only this time restricting the analysis to the critical pairs. 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

Table 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.

Group and pair type	Cycle 1		Cycle 2		Cycle 3	
	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; on cycle 1, all pairs are new.

Table 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)	–

interaction, was significant, both Fs < 1. We also found no difference in cycle-1 resolution (A_g, a nonparametric measure of AUC; Pollack, Norman, & Galanter, 1964) 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 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, $t(29) = 2.276$, $SE = 2.37$, $p = .030$, $d = 0.43$.

All other analyses on the cycle 2 data 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) × 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 1 for the means) – 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 might need 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 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 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) × 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, $F(1,58) = 34.120$, $MSE = 85.70$, $p < .001$, $\eta_p^2 = .370$, 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 × group interaction, $F(1,58) = 5.740$, $MSE = 85.70$, $p = .020$, $\eta_p^2 = .090$: even though recall performance was equated between the control and experimental groups, participants in the experimental condition assigned higher 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 1). The main effect of group was not significant, $F(1,58) = 1.098$, $MSE = 565.22$, $p = .30$, $\eta_p^2 = .019$.

To examine the influence of difficult pairs on JOLs in more detail, we constructed MROC curves for critical pairs (see panel A of Fig. 4). We first compared resolution between the groups. As seen in Fig. 4, the MROC 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 result 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.

Another possible reason for why JOLs differed between the groups is that the manipulation of list context affected the confidence criteria. Specifically, the inclusion of new pairs caused participants to *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 RC 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 MROC in Fig. 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 MROC 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).

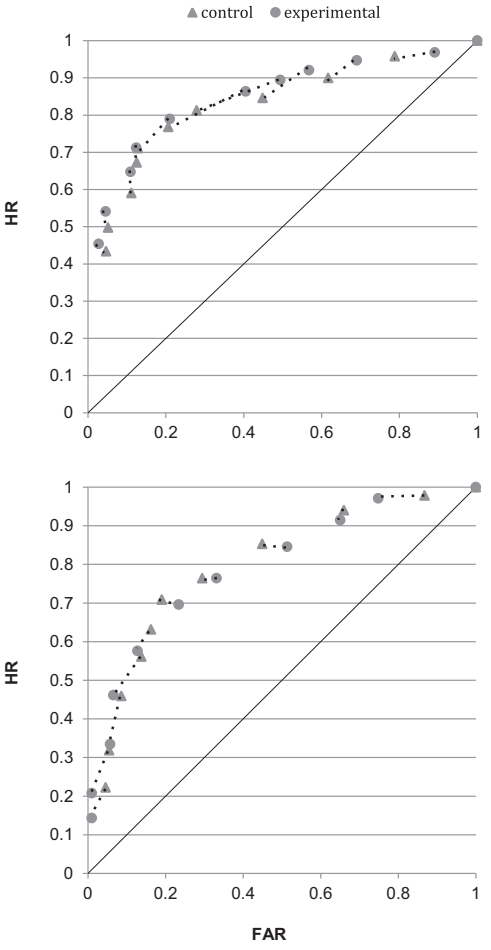

Fig. 4. Metacognitive receiver operating characteristic (MROC) curve for Experiment 1 (top panel) and Experiment 2 (bottom panel). Dashed lines link points denoting the same criteria in the control and experimental group (the bottommost dashed line links the *100%* criteria, the line immediately above links the *90%* criteria, etc.), so that the longer the line, the greater the difference in criterion placement between the groups. 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.

On the other hand, the MROC in the top panel of Fig. 4 suggests criterion placements at the top end of the range (the conservative region) were not affected as much by the addition of new, difficult pairs. Indeed, the placements of the highest confidence criteria (*>60%* – bottom-left area of the MROC) are almost identical between the control and experimental MROCs. This result is understandable 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.

In addition to the MROC 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.[6] The differences in c_1 scores between the experimental and control groups are presented in Fig. 5.[7] Overall, the results are consistent with the selective criterion-shift account outlined 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 JOL mean in the experimental group as compared to the control group was caused mostly by more liberal placement of the criteria associated with *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 (e.g., Pansky & Goldsmith, 2014; see also Hansen & Wänke, 2008, for a related approach, and Criss, 2006, 2010, for a related concept of *differentiation* in recognition memory), the inclusion of new pairs in the experimental group may affect the *perception* of critical pairs: when contrasted 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 MROCs. 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 MROCs reveals, however, that this is not the case; as noted, the high confidence criteria shifted noticeably less than the lower ones.

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 MROCs 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 MROC between the groups. However, a visual inspection of the MROCs shows that this was not the case, as the shapes of 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, MROC 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 interstimulus interval. 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

[6] For technical reasons, we did not statistically analyze 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 equaling either 1 or 0. As such HR and FAR values were common, especially at low- and high-confidence levels, excluding these cases would have led 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.

[7] See Appendix for a full set of c_1 scores across cycles and experiments.

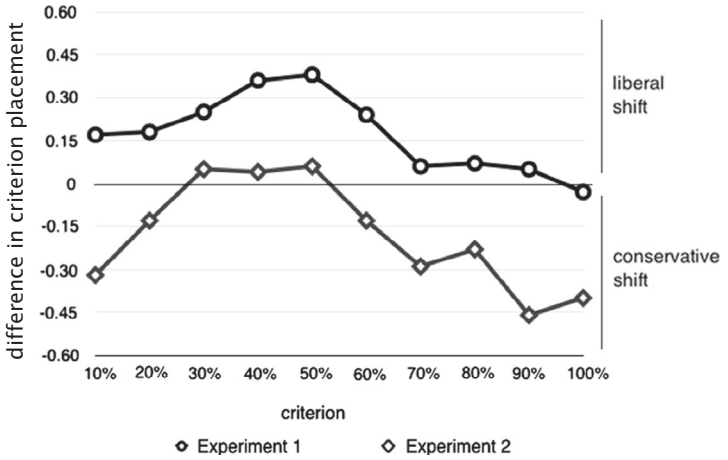

Fig. 5. Differences in c_1 scores between the experimental and control group as a function of criterion in Experiments 1 (upper line) and 2 (bottom line). Values above versus below 0 indicate more liberal versus more conservative responding in the experimental than in the control group, respectively.

Table 3
Means (*SD*s) 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.

Group and pair type	Cycle 1		Cycle 2	
	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.

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 the control and experimental groups. On cycle 2, participants in the control group 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 (not counting the pre-study phase). 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-study phase 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 3. Table 2 presents resolution scores for critical pairs.

Pre-study phase

Only recall performance on the last test of the pre-study phase was analyzed because the format of the last test was identical to that used on the two main study-test cycles of the experiment. 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) × 2 (group: control, experimental) mixed ANOVA, with group as the only between-subjects factor, was conducted on the critical and non-critical pairs pooled together. 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 JOL mean 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 was 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 JOL mean 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 was 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.

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 with 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, $t(32) = 1.721$, $SE = 3.17$, $p = .095$, $d = 0.30$, with higher recall performance for pairs from the pre-study list.

Although the results for new pairs are not the focus of the present study, two interesting aspects of the data for identical pairs are worth noting. First, identical pairs, at least on the surface, should seem easier to learn than pairs from the pre-study phase. However, this was not the case: JOLs for pairs from the pre-study phase exceeded those for identical pairs. 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 Castel et al. (2007), who found that JOLs for identical 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 providing a target that 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 the JOL and recall means 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 MROCs for both groups (see panel B of Fig. 4). However, this time, the data were from cycle 2 rather than cycle 3. 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 MROCs, 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 applied 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 (see also Fig. 5).[8]

The MROCs 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 MROCs. Rather, all points on the MROC 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

[8] Note that no theoretical account would predict the selective misalignment only of the *10%* criteria between the experimental and control groups. Given that no other low-confidence criteria display this trend to a comparable degree, we are inclined to treat this difference in criterion setting as an example of a false positive.

affected and the two MROC 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.[9]

In Experiment 2, 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 MROCs, the context manipulation made the high criteria in the experimental group more conservative. These results support the claim that JOLs are relative in nature, and the mapping between the internal evidence for future recall and JOL values depends on the context in which the judgments are made.

General discussion

In the present study, we employed signal-detection methods to analyze responding in the multi-cycle JOL task. By treating JOL levels as separate confidence criteria, we have demonstrated that the assignment 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 ($\leqslant$*60%*) confidence criteria. In Experiment 2, the range was extended upward by the easy new pairs, consistently affecting the high ($\geqslant$*70%*) confidence criteria. Importantly, these recalibration effects occurred in spite of the lack of differences in resolution, as evidenced by A_g values. The fact that JOL values can be treated as confidence criteria that are malleable and context dependent speaks against the objective interpretation of the *0–100%* JOL scale.

We suggest 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 reasonable 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 the UWP paradigm, recalibration of the percentage JOL scale occurs naturally even when no changes to the procedure are made between the cycles. As items are repeated across the different cycles, there are likely changes in the range of evidence for future recall and these changes could cause recalibration effects even though no new items are added to the list. Here we outline how the UWP effect – the finding of impaired calibration with practice – may at least partially be driven by recalibration.

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 Fig. 1 shift toward the right end of the scale. Moreover, resolution increases (e.g., Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007; Hanczakowski, Zawadzka, et al., 2013; Zawadzka & Higham, 2015), 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.

An example of recalibration at the item level is presented in Fig. 6. The top panel of Fig. 6 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 (see e.g., Wixted, 2007); as a result, the range of evidence gets extended (middle panel). The range of evidence is extended even further in cycle 3 (bottom panel).

In order to accommodate this change in the range of evidence, the rating scale may be recalibrated. Consider items A and B in Fig. 6 taken from a hypothetical study list. As shown 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 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 on cycle 3 is lower than that of item B on the preceding cycle: *80%* compared to *100%*.

[9] It has to be noted, however, that an alternative explanation of our results can be postulated. As our analyses were performed on averaged data, it is viable that a concordant shift of all – rather than selected – criteria might have occurred for some participants, but not for others. In Experiment 1, a subset of more conservative participants might have become more liberal, while in Experiment 2 some more liberal participants might have become more conservative. Although unlikely, this alternative account cannot be excluded on the basis of the current set of data. (We thank David Huber for this suggestion.)

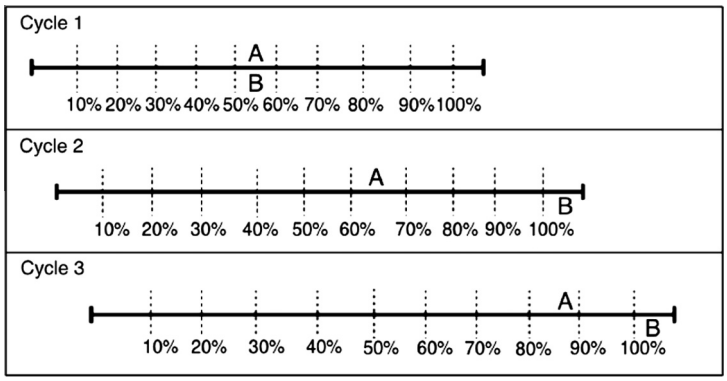

Fig. 6. A graphical presentation of the UWP effect. Three study-test cycles are shown in the three panels: cycle 1 in the top panel, cycle 2 in the middle panel and cycle 3 in the bottom panel. The horizontal lines in each panel represent the range of evidence for the studied items. The dashed vertical lines represent confidence criteria in increments of 10.

This difference exists even though the two items have the same absolute level of evidence in the between-cycle comparison.

Moreover, as the procedure progresses, more items gain more evidence, resulting in a cluster of items with 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 mapping between the scale values and underlying evidence from cycle to cycle is not accompanied by any changes in the perceived likelihood of recalling the studied items. The result is that the JOL mean for items that share the same subjective probability of recall should decrease from cycle to cycle. For example, the mean rating for the subset of items occupying the same location on the evidence dimension could well be *75%* on cycle 1, *70%* on cycle 2 and *65%* on cycle 3. Consequently, if the mean of all JOLs on later cycles is calculated, it falls below that for memory performance. Traditionally, this would be interpreted as underconfidence. However, this result may simply be a result of scale recalibration over cycles and have nothing to do with "true" underconfidence.

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, Zawadzka, et al., 2013; Zawadzka & Higham, 2015), 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, calibration results cannot be meaningfully interpreted as reflecting under- or overconfidence.

Our recalibration account may also provide an explanation for why the UWP effect is found with *0–100%* scale JOLs, but not binary *yes/no* JOLs or binary betting decisions (Hanczakowski, Zawadzka, et al., 2013; Zawadzka & Higham, 2015). Unlike *0–100%* scale JOLs, binary judgments require only one criterion ("yes/no" or "bet/no bet"); shifting this single criterion to a more conservative position to accommodate new learning would result in an unacceptably high metacognitive miss rate (i.e., high proportion of recalled items assigned negative responses). This willingness to shift higher confidence criteria further up the evidence dimension if making *0–100%* scale JOLs coupled with an unwillingness to shift a single criterion upward if making binary judgments results in a scale/binary dissociation (although see Experiment 3 of Zawadzka & Higham, 2015, for an exception to this dissociative pattern).

Implications for interpreting other metacognitive ratings and individual differences

As discussed above, the interpretation of the percentage JOL scale as context dependent can pose problems for experimenters employing *0–100%* JOLs in their research. However, in our view, there is no fundamental difference between the percentage JOL scale and other rating scales that would limit recalibration effects to *0–100%* JOLs. Indeed, past research suggests that recalibration effects can be also found in measures other than JOLs elicited on a percentage scale. One such an example comes from a study by Mickes, Hwe, Wais, and Wixted (2011) who investigated scaling of items high on the evidence dimension in the context of recognition memory judgments. Feedback, which was administered in their Experiment 5, made participants recalibrate their rating scale by making the highest confidence criteria more conservative. Given the ubiquity of feedback manipulations in memory and metamemory studies (e.g., Dunlosky & Rawson, 2012; Koriat, 1997; McGillivray & Castel, 2011; Rhodes &

Tauber, 2011; Verde & Rotello, 2007; Zawadzka, Krogulska, Button, Higham, & Hanczakowski, 2016), its potential for producing recalibration effects certainly warrants future research, with particular focus on the generalizability to other rating types.

Recalibration effects may not only generalize to other rating scales, but may also account for observed differences in realism between populations. For example, compared to younger adults, older adults have been shown to be prone to poor memory accuracy. Their metacognitive ratings, on the other hand, do not seem to reflect this decline, leading to apparent overconfidence (e.g., Connor et al., 1997; Dodson, Bawa, & Krueger, 2007). McDonough, Cervantes, Gray, and Gallo (2014) followed up on these findings with an fMRI study which compared older and younger adults' memory and subjective recollection assessments (made on a *0*-to-*3* scale) for complex pictures. In their experiment, older and younger participants studied complex pictures accompanied by verbal labels. At test, labels were presented one at the time and participants' task was to recollect as many aspects of the picture studied with a given label as possible, and give a rating of the amount of detail recollected on a scale from *0* (no detail) to *3* (high detail). As shown by behavioral data, younger adults outperformed older adults on the memory task. Consistent with the behavioral results, the fMRI data suggested greater perceptual reactivation from memory in younger adults. Despite these objective differences in recollection, subjective ratings of recollected details were comparable between the age groups.

McDonough et al. (2014) presented a recalibration account of their findings, by which older adults adjust their rating scale to the lower quality of information retrieved from memory. In this way, even though older participants recollected less information about the studied pictures, this was not reflected in their subjective ratings of recollection due to their rating criteria being more liberal than those of younger adults. The authors noted that their results "highlight the difficulty of using subjective report as an index of the amount of actual detail retrieved from memory in different groups" (p. 356).

It is therefore clear that the problem with the interpretation of ratings is more general, as it applies not only to the percentage JOL task used in the present study, but to other tasks and rating scales as well. It can also be caused both by experimental manipulations, and by testing groups of participants differing on a particular dimension such as memory capability. We believe that plotting MROCs to corroborate the results may be a good strategy in such cases. As demonstrated in the present study, MROCs 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.

Limitations of the SDT approach

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). Indeed, by analyzing the nature of the MROCs, we were able to successfully eliminated a metacognitive contrast account of our data, an account which would fall in the latter category. However, in this study, 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 MROC. Yet, as noted above, there are other cases that do not allow for such clean conclusions. In theorizing on the usefulness of SDT, it has been noted that it is sometimes 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 MROC – 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 points on the MROC and measures of criterion placement are affected in the same way. However, as noted above, we believe these to be ideal cases that are unlikely to occur often in reality. Rather, it seems more likely that only a portion of the items and/or criteria will be affected by an experimental manipulation as was the case in our data. If so, SDT is a valuable tool for discriminating between real changes in perception of the to-be-rated items on the one hand versus scale recalibration on the other.

A stochastic detection and retrieval model of JOLs

In the present paper, we used a signal detection model to gain insight into the interpretation of scale JOLs. However, it has to be noted that there are alternative models that are capable of capturing the complexities of scale JOL assignment. One such signal detection-like model has been proposed by Jang, Wallsten, and Huber (2012) and dubbed the Stochastic Detection and Retrieval Model (SDRM). In this model, it is assumed that there are two separate samplings from memory for each item. One of these samplings underlies retrieval from memory, while the other one allows a metacognitive rating to be assigned. If the memory sampling returns a strength value above a retrieval threshold, an item is recalled; otherwise, retrieval fails. For the metacognitive sampling, the value returned determines the confidence rating in much the same way that the amount of subjective evidence determines confidence in the SDT model. The correlation between the two samplings constitutes one of the parameters in the SDRM. It can be very high when the overlap between the memory information at retrieval and at rating is also high. It can also be low if, for example, memory deteriorates between the two samplings. As the order of the two samplings is irrelevant, the SDRM can be applied to prospective and retrospective metacognitive judgments alike.

The SDRM model shares some of the qualities with the SDT model of JOLs presented here. Contrary to Jang et al.'s (2012) claim, both SDRM and SDT do not require indepen-

dently defined stimulus categories (such as studied/non-studied items in recognition memory paradigms), but can be applied also to tasks in which stimuli are classified on the basis of participants' responses (such as successful/unsuccessful recall or correct/incorrect answer; see, e.g., Benjamin & Diaz, 2008; Ferrell & McGoey, 1980; Galvin, Podd, Drga, & Whitmore, 2003; Higham, 2007, 2013). Also, applications of SDT to prospective judgments such as JOLs or FOKs (Benjamin & Diaz, 2008; Masson & Rotello, 2009) necessarily require two separate samplings from memory to be made – one at the time of the judgment, and the other at test.

The main difference between the two models lies in the nature of the distributions underlying the memory decisions and metamemory ratings. As shown in Fig. 1, the SDT model of JOLs assumes two separate distributions for recalled and unrecalled items, with the recalled items distribution positioned to the right of the unrecalled items distribution. In the SDRM, on the other hand, a single distribution is postulated. This distribution is further split into two parts by a recall criterion, with recalled items falling to the right of the criterion, and unrecalled items to the left. This assumption makes the SDRM robust, as the model can also be applied to tasks which do not satisfy the requirement of having underlying normal distributions. It is worth noting here, however, that JOL data, when split into recalled/unrecalled categories, tend to be normally distributed (Benjamin & Diaz, 2008), satisfying the normality assumption necessary for calculating SDT measures.[10]

We would argue that both models have the potential to enhance our understanding of the bases of metacognitive judgments above and beyond that which can be gleaned from experimental data. Apart from the similarities discussed above, both the SDRM and the SDT model of JOLs have strengths that the other model does not possess. The SDT model, by the virtue of being based on two underlying distributions, allows for plotting MROCs, which, as shown in the present study, can be a useful tool for theory testing. The SDRM, on the other hand, deals with criterion noise (see, e.g., Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008) – that is, inconsistency in criterion setting across trials – by including a criterion noise parameter that is absent in SDT.[11]

Conclusion

In the present study, we have shown how applying the SDT model to *0–100%* JOL data allows for formulating and testing new predictions regarding the mechanisms responsible for JOL assignment. By plotting MROCs, we have demonstrated that list composition can lead to a selective recalibration of some – but not all – JOL criteria, without affecting the perceived difficulty of the studied items. This recalibration account of JOLs is consistent with previous findings (Dunlosky et al., 2005; Hanczakowski, Zawadzka, et al., 2013; Zawadzka & Higham, 2015) suggesting that JOL assignment does not have to be based on the assessed probability of future recall. We suggest that the SDT approach used in this study can be applied with success to other rating tasks to help distinguish between competing theoretical accounts.

[10] Higham (2007) found that the normality assumption was satisfied with RC ratings as well.

[11] Criterion noise increases with an increase in response scale length (Benjamin, Tullis, & Lee, 2013), which might be especially problematic for scales with multiple response options such as the *0–100%* JOL scale.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2016.04.005.

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. http://dx.doi.org/10.1080/10683160512331329943.

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. http://dx.doi.org/10.3758/s13421-010-0002-y.

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., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*, 84–115. http://dx.doi.org/10.1037/a0014351.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1601–1608. http://dx.doi.org/10.1037/a0031849.

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgments. *Journal of Personality and Social Psychology, 60*, 485–499. http://dx.doi.org/10.1037/0022-3514.60.4.485.

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, and Cognition, 34*, 918–928. http://dx.doi.org/10.1037/0278-7393.34.4.918.

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. http://dx.doi.org/10.3758/BF03194036.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50–71. http://dx.doi.org/10.1037/0882-7974.12.1.50.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength-based mirror effect. *Journal of Memory and Language, 55*, 461–478. http://dx.doi.org/10.1016/j.jml.2006.08.003.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 484–499. http://dx.doi.org/10.1037/a0018435.

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging, 22*, 122–133. http://dx.doi.org/10.1037/0882-7974.22.1.122.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280. http://dx.doi.org/10.1016/j.learninstruc.2011.08.003.

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

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26*, 32–53. http://dx.doi.org/10.1016/0030-5073(80)90045-8.

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. http://dx.doi.org/10.1016/j.jml.2011.01.006.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 238–244. http://dx.doi.org/10.1037/0278-7393.33.1.238.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19–34. http://dx.doi.org/10.1016/j.jml.2007.03.006.

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141*, 124–133. http://dx.doi.org/10.1037/a0024006.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*, 843–876. http://dx.doi.org/10.3758/BF03126546.

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., Pasek, M., Zawadzka, K., & Mazzoni, G. (2013). Cue familiarity and 'don't know' responding in episodic memory tasks. *Journal of Memory and Language, 69*, 368–383. http://dx.doi.org/10.1016/j.jml.2013.04.005.

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. http://dx.doi.org/10.3758/s13423-013-0497-x.

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. http://dx.doi.org/10.1016/j.jml.2013.05.003.

Hansen, J., & Wänke, M. (2008). It's the difference that counts: Expectancy/experience discrepancy moderates the use of ease of retrieval in attitude judgments. *Social Cognition, 26*, 447–468. http://dx.doi.org/10.1521/soco.2008.26.4.447.

Higham, P. A. (2007). No special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General, 136*, 1–22. http://dx.doi.org/10.1037/0096-3445.136.1.1.

Higham, P. A. (2013). Regulating accuracy on university tests with the plurality option. *Learning and Instruction, 24*, 26–36. http://dx.doi.org/10.1016/j.learninstruc.2012.08.001.

Higham, P. A. (2011). Accuracy discrimination and type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In P. A. Higham & 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. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. In J. Dunlosky & S. K. Tauber (Eds.), *Oxford handbook of metamemory*. http://dx.doi.org/10.1093/oxfordhb/9780199336746.013.15 (Advance online publication).

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review, 119*, 186–200. http://dx.doi.org/10.1037/a0025960.

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. http://dx.doi.org/10.3758/BF03196156.

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. http://dx.doi.org/10.1037//0096-3445.126.4.349.

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. http://dx.doi.org/10.1037/0096-3445.133.4.643.

Lichtenstein, S., & Fischhoff, B. (1981). *The effects of gender and instructions on calibration*. Decision research technical report PTR-1092-81-7. Eugene, OR.

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. http://dx.doi.org/10.1080/15248372.2011.577760.

Luna, K., Higham, P. A., & Martín-Luengo, B. (2011). The regulation of memory accuracy with multiple answers: The plurality option. *Journal of Experimental Psychology: Applied, 17*, 148–158. http://dx.doi.org/10.1037/a0023276.

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. http://dx.doi.org/10.1037/a0014876.

McDonough, I. M., Cervantes, S. N., Gray, S. J., & Gallo, D. A. (2014). Memory's aging echo: Age-related decline in neural reactivation of perceptual details during recollection. *NeuroImage, 98*, 346–358. http://dx.doi.org/10.1016/j.neuroimage.2014.05.012.

McGillivray, S., & Castel, A. D. (2011). Betting on memory leads to metacognitive improvement by younger and older adults. *Psychology and Aging, 26*, 137–142. http://dx.doi.org/10.1037/a0022681.

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. http://dx.doi.org/10.1037/a0023007.

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*, 200–219. http://dx.doi.org/10.1037/a0039923.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of Signal Detection Theory. *Psychonomic Bulletin & Review, 15*, 465–494. http://dx.doi.org/10.3758/PBR.15.3.465.

Pansky, A., & Goldsmith, M. (2014). Metacognitive effects of initial question difficulty on subsequent memory performance. *Psychonomic Bulletin & Review, 21*, 1255–1262. http://dx.doi.org/10.3758/s13423-014-0597-2.

Pollack, I., Norman, D., & Galanter, E. (1964). An efficient nonparametric analysis of recognition memory. *Psychonomic Science, 1*, 327–328.

Rast, P., & Zimprich, D. (2009). Age differences in the underconfidence-with-practice effect. *Experimental Aging Research, 35*, 400–431. http://dx.doi.org/10.1080/03610730903175782.

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. http://dx.doi.org/10.1037/a0013684.

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. http://dx.doi.org/10.1080/09658211.2011.613841.

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. http://dx.doi.org/10.1007/s10979-009-9192-x.

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. http://dx.doi.org/10.1037/0096-3445.134.1.124.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1258–1266. http://dx.doi.org/10.1037/0096-3445.134.1.124.

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. http://dx.doi.org/10.3758/s13421-013-0323-8.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory and Cognition, 35*, 254–262. http://dx.doi.org/10.3758/BF03193446.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176. http://dx.doi.org/10.1037/0033-295X.114.1.152.

Zawadzka, K., & Higham, P. A. (2015). Judgments of learning index relative confidence, not subjective probability. *Memory & Cognition, 43*, 1168–1179. http://dx.doi.org/10.3758/s13421-015-0532-4.

Zawadzka, K., Krogulska, A., Button, R., Higham, P. A., & Hanczakowski, M. (2016). Memory, metamemory, and social cues: Between conformity and resistance. *Journal of Experimental Psychology: General, 145*, 181–199. http://dx.doi.org/10.1037/xge0000118.