Article:
Zawadzka, K., Higham, P.A. and Hanczakowski, M. (2017) Confidence in Forced-Choice Recognition: What Underlies the Ratings? Journal of Experimental Psychology: Learning, Memory, and Cognition , 43 (4). pp. 552-564. ISSN: 0278-7393

https://doi.org/10.1037/xlm0000321

Confidence in forced-choice recognition: What underlies the ratings?

Katarzyna Zawadzka[1], Philip A. Higham[2], and Maciej Hanczakowski[3]

[1]Nottingham Trent University

[2]University of Southampton

[3]Cardiff University

Word count: 10,691 (excluding references)

Author Note

Katarzyna Zawadzka, Division of Psychology, Nottingham Trent University; Philip A. Higham, Psychology, University of Southampton; Maciej Hanczakowski, School of Psychology, Cardiff University.

The authors would like to thank Giuliana Mazzoni for helpful discussions concerning this project.

Correspondence concerning this article should be addressed to Katarzyna Zawadzka, Division of Psychology, Nottingham Trent University, Burton Street, Nottingham NG1 4BU, UK, email: katarzyna.zawadzka@ntu.ac.uk, or Maciej Hanczakowski, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, UK, email: hanczakowskim@cardiff.ac.uk.

Article published in Journal of Experimental Psychology: Learning, Memory, and Cognition (http://dx.doi.org/10.1037/xlm0000321).

Abstract

Two-alternative forced-choice recognition tests are commonly used to assess recognition accuracy that is uncontaminated by changes in bias. In such tests, participants are asked to endorse the studied item out of two presented alternatives. Participants may be further asked to provide confidence judgments for their recognition decisions. It is often assumed that both recognition decisions and confidence judgments in two-alternative forced-choice recognition tests depend on participants' assessments of a difference in strength of memory evidence supporting the two alternatives – the relative account. In the present study we focus on the basis of confidence judgments and we assess the relative account of confidence against the absolute account of confidence, by which in assigning confidence participants consider only strength of memory evidence supporting the chosen alternative. The results of the study show that confidence in two-alternative forced-choice recognition decisions is higher when memory evidence is stronger for the chosen alternative and also when memory evidence is stronger for the unchosen alternative. These patterns of results are consistent with the absolute account of confidence in two-alternative forced-choice recognition but they are inconsistent with the relative account.

Keywords: Confidence; Recognition; Two-alternative forced-choice test

Confidence in forced-choice recognition: What underlies the ratings?

In the recognition literature, two ways of assessing recognition performance are commonly used. In standard single-item recognition tests, decisions as to whether single items were studied earlier or not are required. In two-alternative forced-choice (2AFC) recognition tests, decisions regarding which of two presented items was studied are required. For both types of tests, a recognition question is often accompanied by a question concerning one's confidence in the recognition decision. Confidence in single-item recognition has been widely studied in the recognition literature, which has led to the development of various models to account for differences in the distributions of confidence ratings (Wixted, 2007; Yonelinas, 1994; Yonelinas & Parks, 2007). However, confidence in 2AFC recognition has received much less scrutiny. The purpose of the present study is thus to elucidate the basis of confidence in 2AFC recognition decisions.

In a 2AFC recognition test there are two possible sources of memory information: a target and a lure. It is commonly assumed that performance in such a task depends on combining evidence from these two sources. Specifically, models of recognition memory assume that evidence for the two alternatives is compared and the alternative supported by more evidence is endorsed as the target (e.g., Jang, Wixted, & Huber, 2009; Macmillan & Creelman, 2005). For recognition decisions, the magnitude of the difference is unimportant; all that is needed is for one alternative to be supported by more evidence than the other and it will be chosen. However, the same need not be true for *confidence* in the recognition decision. According to the *relative account* of confidence, the larger the difference in evidence supporting one alternative over the other, the stronger the confidence expressed in the recognition endorsement of this alternative (e.g., Clark, 1997).

One strand of research is particularly informative in relation to the relative account of confidence in 2AFC recognition decisions. A number of studies have examined confidence in 2AFC recognition decisions when the similarity between targets and lures is manipulated (Tulving, 1981; Dobbins, Kroll, & Liu, 1998; Heathcote, Bora, & Freeman, 2010; Heathcote, Freeman, Etherington,

Tonkin, & Bora, 2009). Tulving (1981) developed a paradigm in which participants studied picture halves and were later given a 2AFC test in which studied halves were paired with three types of lures: similar lures which were other, non-presented halves of the target picture (A – A' pairs), dissimilar lures which were non-presented halves of other studied pictures (A – B' pairs), and novel lures which were halves corresponding to non-studied pictures (A – X' pairs). Comparison of A – A' and A – B' pairs revealed a cross-over dissociation of accuracy and confidence; that is, accuracy was higher for A – A' pairs but confidence was higher for A – B' pairs. As shown by Clark (1997), this inversion can be accounted for by the relative account if it is assumed that the distribution of differences in mean evidence supporting the two alternatives is characterized by lower variance for A – A' pairs than for A – B' pairs.[1]

However, important insights into the relative account of confidence can also be gleaned from a comparison of correct responses to A – B' and A – X' pairs. For these pairs, the target strength is held constant and only the strength of the lure is varied by the virtue of B' being similar to one of the studied items and X' being dissimilar from all studied items. If the A alternative is correctly endorsed, the relative account of confidence predicts that confidence should be higher for A – X' pairs than for A – B' pairs. Indeed, Heathcote et al. (2009) observed such a pattern in a study in which they examined overall confidence scores in a recognition test for faces. However, support for the relative account has not been unequivocal. For example, Tulving (1981) and Dobbins et al. (1998) presented confidence results separately for correct and incorrect recognition decisions and showed a pattern clearly inconsistent with the relative account prediction. For both studies, when participants chose the B' alternative, they were more confident than when they chose X' and when participants chose correctly the A alternative, the strength of the lure did not matter at all.

[1] It is assumed here that the means of two distributions of differences in evidence are equal but the variance is smaller for the A – A' distribution than the A – B' distribution due to shared evidence between alternatives for the former pairs. A smaller variance means that a greater proportion of differences in evidence are concentrated near the mean, which favors correct identifications of a target. On the other hand, smaller variance means also that a smaller proportion of pairs lie in the tails of the distributions where the magnitude of the differences is largest. As a result, confidence is lower for the A – A' pairs even if accuracy is greater.

An alternative model of confidence in recognition decisions can be derived from the literature concerned with metacognitive judgments. Confidence in one's recognition decisions is in essence a metacognitive judgment concerning one's own memory processes. Whereas the recognition literature often looks at how the same information (i.e., memory evidence) shapes both accuracy and confidence, as in the relative account of confidence in 2AFC recognition, the metacognitive perspective acknowledges that metacognitive judgments may be shaped by factors different from those which support memory performance (Busey, Tunnicliff, Loftus, & Loftus, 2000; Reinitz, Peria, Séguin, & Loftus, 2011; Reinitz, Séguin, Peria, Loftus, 2012). One prominent strand of theorizing in the metacognitive literature is the accessibility theory developed by Koriat (1993, 1995; see also Hanczakowski, Zawadzka, Collie, & Macken, 2016). In the accessibility theory, it is assumed that people's perceptions concerning what they know are shaped by the overall volume of memory information about the target that can be accessed at retrieval, independently of whether this information is partial or complete, correct or incorrect or even whether it is relevant or irrelevant to the memory question that is being asked (Brewer, Marsh, Foos, & Meeks, 2010). This accessibility framework suggests an alternative account of confidence in 2AFC recognition. According to the *absolute account*, the primary factor behind confidence in the 2AFC task would be information accessed for a chosen alternative. Such an account seems at first blush better suited for describing the pattern of data observed for A – B' and A – X' pairs in the paradigm developed by Tulving (1981) than the relative account. When participants choose the target (A), evidence supporting the unchosen lure is discounted; instead, their confidence should primarily depend on memory evidence gathered for this target, which may be comparable for both types of pairs. When participants choose a lure (either B' or X'), their confidence should primarily depend on memory evidence gathered for this lure, which on average should be stronger for B' lures similar to studied items than for X' lures that are not similar to any of the studied items.

The present study was designed to rigorously test the relative and absolute accounts of confidence in 2AFC recognition tests. In order to provide such a test, it is necessary to look

independently at the role of both the strength of evidence supporting the chosen alternative and the difference in evidence supporting both alternatives. As already described, the strength of the chosen alternative and the difference in strength of both alternatives can be disentangled using Tulving's (1981) paradigm. However, the paradigm is limited in that it only allows for assessment of the role of the strength of the chosen alternative when participants endorse a lure (as strength differs between B' and X' lures) and the role of the difference in strengths of both alternatives when participants endorse a target (difference is strength is then larger for A – X' than for A – B' pairs). It is unclear whether the same conclusions would hold if the role of the strength of the chosen alternative was assessed for targets and the role of the difference in strengths of both alternatives was assessed when participants endorse a lure. Accordingly, in the present study we provided a new test of the relative and absolute accounts of confidence by relying on two different methods.

First, we independently varied evidence supporting both targets and lures. Thus we created pairs in which both targets and lures were supported by weak or strong memory evidence and also pairs in which either only a target or only a lure was supported by strong memory evidence. This design allowed us to look both at confidence when the strength of the chosen alternative is varied and at confidence when the strength of the unchosen alternative is varied, with all comparisons conducted separately for correct and incorrect choices. The relative account of confidence straightforwardly predicts confidence to increase as the chosen alternative gains in strength and/or as the unchosen alternative decreases in strength. The absolute account of confidence also predicts confidence to increase as the chosen alternative gains in strength, but its predictions differ depending on the strength of the unchosen alternative. When unchosen alternatives are weak, the strength of the chosen alternative can range from relatively weak (only marginally stronger than the rejected alternative) to very strong. However, if unchosen alternatives are strong, the set of chosen alternatives is restricted to the strongest alternatives. Given that the average strength of the chosen alternatives should thus be higher when the unchosen alternatives are strong rather than weak, the

absolute account predicts increased confidence in a recognition choice with increased strength of the unchosen alternative, a prediction that is directly opposite to that of the relative account.[2]

Second, we included in our design pairs that were composed from either two targets or two lures. The strength of evidence supporting two alternatives in these pairs was varied across pairs but held constant within pairs. Given that for these pairs, at all levels of strength, the mean difference in evidence between the two alternatives should be zero, the relative account predicts no difference in confidence between strong and weak pairs. By contrast, the absolute account, which argues for the role of strength of the chosen alternative only, makes a clear prediction that confidence should be higher for pairs consisting of two equally strong alternatives than for pairs consisting of two equally weak alternatives.

In order to contrast targets and lures supported by strong and weak memory evidence we adapted the plurals paradigm (Hintzman & Curran, 1994) for the purpose of the present study. In the plurals paradigm, participants study various concrete nouns in their singular or plural form. On the final recognition test, participants are presented with studied words and lures which are words for which the plurality form has been changed from encoding. Thus, if participants study *frog* and *tables*, later they might be given the test pair *frog* – *table* for which *frog* is the target and *table* is the lure. The plurals paradigm allows the strength of evidence supporting the lure to be manipulated by varying the number of presentations of its parent word. Thus, multiple presentations of *tables* should increase memory support for the word 'table' by virtue of their similarity.

An additional factor needs to be considered in relation to the plurals paradigm, however, which is recall-to-reject process (see Rotello, Macmillan, & Van Tassel, 2000, for evidence of recall-to-reject in the plurals paradigm; see also Rotello & Heit, 1999, for a discussion of failures to detect the influence of recall-to-reject in this paradigm). Participants may be able to reject a lure related to

[2] Note that the aforementioned patterns observed by Tulving (1981) and Dobbins et al. (1998) of no difference in confidence as a function of strength of the unchosen alternative are actually inconsistent with both accounts. We return to this issue in the General Discussion.

a strong study item by virtue of recalling the original study word and inferring that a presented alternative should be rejected given its different form. Successful recall-to-reject may serve to increase confidence when the rejected lure is similar to a strong study item, thus reducing a possible impact of the difference in strength that is predicted by the relative account of confidence (i.e., under conditions of greater similarity, this account would predict low confidence, which may be increased if participants are able to confidently reject a lure based on recall-to-reject). However, it is vital to note that recall-to-reject by definition occurs only for lures, and not for targets, so its influence on the pattern of confidence can be inferred if asymmetries occur between various conditions depending on whether the endorsed item is a target or a lure. This issue will be discussed further after we report the results of each of the experiments.

Experiment 1

The present experiment looked at confidence in 2AFC recognition decisions in a paradigm that varied orthogonally the strength of evidence supporting targets and lures in the plurals paradigm. The strength of the alternatives (both targets and lures) was varied by the number of presentations of their parent words: more presentations of parents words should induce a greater similarity of the recognition alternative to memory records, thus increasing the strength of memory evidence for this alternative (save for the recall-to-reject process for lures, which is addressed later). The final recognition test included also what will be referred to as null trials: target-target pairs for one group of participants and lure-lure pairs for the other group. For null trials, both alternatives were on average of equal strength, either high or low, which was also manipulated by varying the number of presentations of their parent words. The main focus of the present experiment was on confidence when: a) the strength of the chosen alternative varies, b) the strength of the unchosen alternative varies, and c) the strength of both alternatives is equal. All comparisons were conducted while accounting for the correctness of the recognition decisions.

Method

Participants. Fifty-two undergraduate students of Cardiff University participated in this experiment for course credit or monetary compensation. Twenty-six participants were assigned to each of the experimental groups.

Materials and design. A total of 160 singular nouns with length ranging from four to eight letters were chosen from the MRC database. A set of 160 plural nouns was created by adding the letter "-s" to each singular noun; this resulted in the length range for plural nouns of five to nine letters. For study, a list of 140 words was used. Half of the words within the study list were presented in singular form (e.g., *frog*), while the other half were presented in plural form (e.g., *tables*). For counterbalancing purposes, two versions of the study list were created so that half of the participants saw the words *frog* and *tables*, while the other half were presented with *frogs* and *table*. In addition to varying word forms, word strength was manipulated by varying the number of presentations at study. Non-strengthened words were presented only once, while strengthened words were presented three times.

At test, 160 words were presented, 80 in singular and 80 in plural form. Three different item categories were used. Words presented in the same form at study and at test (either singular-to-singular or plural-to-plural) served as targets at test. Half of the targets were non-strengthened (studied once), and half were strengthened (studied three times). Words presented in different form at study and at test (singular-to-plural or plural-to-singular) served as lures, with half of them being non-strengthened (i.e., presented at study once in a different form) and half being strengthened (i.e., presented at study three times in a different form). New words, not presented before, were also included in the lure category. New words were included for consistency with previous studies using the plurals paradigm (e.g., Rotello & Heit, 2000) and results from trials containing such new words were not analyzed.

There were two versions of the test list, each reflecting the assignment of words to singular or plural form at study. The only between-participants factor in this experiment was the type of null

trials at test, with one group being presented with target-target trials, and the other with lure-lure trials. Figure 1 presents the assignment of word types to test trials in both experimental groups.

Procedure. The procedure was the same for both experimental groups. Before the study phase began, participants were provided with instructions for the encoding task:

> You will now be presented with study words. Try to memorize as many words as possible.
> Each word will appear on a computer screen for a brief period of time. Some of these words will be presented more than once. Also, some of the words will be presented in a singular form, whereas other words will be presented in plural. Try to memorize the form in which the words are presented. To make the task easier, you may pronounce each word quietly to yourself.
> Note that the words will be presented at a very fast rate. However, please try to concentrate on memorizing the words during the whole presentation.

At study, each word was presented for 700 ms, with a 300 ms inter-stimulus interval. The order of presentation was randomized anew for each participant. After the presentation of the whole study list, instructions for the retrieval task were displayed:

> Your memory for the words presented earlier will now be tested. You will be presented with pairs of words. Your task is to choose ONE word from each pair that was presented before. Some of the words may be presented in a different form than the one in which they appeared in the study list. Such a word should be considered an INCORRECT answer. For example, imagine that the word "BANANA" appeared on the study list. In this case, "BANANA" would be the correct answer, whereas "BANANAS" should be considered incorrect.

> After choosing the word that was presented before, rate your confidence that your answer is correct. Type in "1" if you are guessing and "6" if you are sure that this is the correct answer. Use numbers 2-5 to indicate intermediate levels of confidence.

On each test trial, test words were aligned horizontally. Participants could not advance to the next trial unless an answer and a confidence rating were provided.

Results

The main purpose of the present experiment was to examine confidence in 2AFC recognition as a function of the strength of the chosen and unchosen alternatives. Before presenting the confidence data, however, we discuss how manipulations of strength of the alternatives affected accuracy of recognition decisions.

Accuracy. The means for the accuracy data can be found in Table 1. The null trials (two lures or two targets) are not discussed here as, by definition, all responses on these trials were either correct (two targets) or incorrect (two lures), but given that the type of null trials defined our two experimental groups, all analyses are performed with the group as a between-participants factor. A 2 (target strength: strong vs. weak) x 2 (lure strength: strong vs. weak) x 2 (null trials group: two targets vs. two lures) mixed Analysis of Variance (ANOVA) on hit rates yielded a significant main effect of target strength, $F(1, 50) = 35.10$, $MSE = .03$, $p < .001$, $\eta_p^2 = .41$, and a significant main effect of lure strength, $F(1, 50) = 5.05$, $MSE = .03$, $p = .029$, $\eta_p^2 = .09$. No other effect was significant, largest $F(1, 50) = 3.00$, $p = .089$, for the target strength x group interaction. These results indicate that, predictably, 2AFC recognition performance improved if recognition targets were strong, $M = .67$, $SD = .14$, rather than weak, $M = .53$, $SD = .15$, and if recognition lures were weak, $M = .63$, $SD = .15$, rather than strong, $M = .57$, $SD = .15$.

Confidence. The means for the confidence data for standard pairs can be found in Table 1. For confidence data, results were analyzed separately for trials on which a correct and incorrect alternative was endorsed. For correct trials on which a target was endorsed, a 2 (target strength) x 2 (lure strength) x 2 (group) mixed ANOVA on mean of confidence judgments yielded significant main effects of target strength, $F(1, 50) = 34.74$, $MSE = 0.56$, $p < .001$, $\eta_p^2 = .41$, and lure strength, $F(1, 50) = 5.26$, $MSE = 0.54$, $p = .026$, $\eta_p^2 = .10$. No other effect was significant, all Fs < 1. The main effect of target strength shows that participants were more confident in their correct recognition decisions when endorsed targets were strong, $M = 4.06$, $SD = 0.89$, rather than weak, $M = 3.44$, $SD = 0.86$. The main effect of lure strength shows that participants were also more confident in their correct recognition decisions when the unchosen lures were strong, $M = 3.87$, $SD = 0.94$, rather than weak, $M = 3.63$, $SD = 0.79$.

For the analysis of incorrect trials on which lures rather than targets were endorsed, eight participants needed to be excluded due to missing cells (i.e., perfect recognition performance in one of the conditions). A 2 (target strength) x 2 (lure strength) x 2 (group) mixed ANOVA on mean confidence judgments from the remaining 44 participants yielded significant main effects of target strength, $F(1, 42) = 5.57$, $MSE = 0.64$, $p = .023$, $\eta_p^2 = .12$, and lure strength, $F(1, 42) = 22.84$, $MSE = 0.83$, $p < .001$, $\eta_p^2 = .35$. No other effect was significant, largest $F(1, 42) = 1.21$, $p = .28$. The main effect of lure strength shows again that participants were more confident in their incorrect recognition decisions when the endorsed alternatives (i.e., lures for incorrectly answered trials) were strong, $M = 3.87$, $SD = 0.74$, versus weak, $M = 3.22$, $SD = 1.00$. The main effect of target strength shows again that participants were also more confident in their incorrect recognition responses when the unchosen alternatives (i.e., targets for incorrectly answered trials) were strong, $M = 3.69$, $SD = 0.90$, rather than weak, $M = 3.40$, $SD = 0.81$.

Mean confidence ratings assigned to null pairs can be found in Table 2. Confidence for null trials was analyzed with a 2 (group: target-target vs. lure-lure) x 2 (strength of alternatives: strong vs.

weak) mixed ANOVA. This yielded a significant main effect of strength of alternatives, $F(1, 50) = 51.20$, $MSE = 0.28$, $p < .001$, $\eta_p^2 = .51$ and a marginal effect of group, $F(1, 50) = 3.75$, $MSE = 1.03$, $p = .058$, $\eta_p^2 = .07$. However, both effects were qualified by a significant interaction, $F(1, 50) = 7.61$, $MSE = 0.28$, $p = .008$, $\eta_p^2 = 13$. The interaction arose because the effect of strength on confidence was larger for null trials consisting of two targets rather than two lures, although it was reliable for both comparisons, $t(25) = 7.32$, $SE = 0.14$, $p < .001$, and $t(25) = 2.99$, $SE = 0.15$, $p = .006$, respectively.

Discussion

The results of the present experiment can be summarized as four main points. First, the accuracy of 2AFC recognition decisions was affected in a predictable way by our strength manipulations. Participants were more accurate when targets were strengthened by multiple presentations at study but they were also less accurate when lures closely resembled words that were strengthened at study. Second, confidence in recognition decisions was affected by the strength of the chosen alternative so that participants were more confident when they were endorsing the alternative that was strong rather than weak. This effect occurred for both chosen targets and lures. Third, confidence in recognition decisions was also affected by the strength of the unchosen alternative so that participants were more confident when the unchosen alternative was strong rather than weak. Again, this effect occurred for both unchosen targets and lures. These two strength effects, one for the chosen alternatives and the other for the unchosen alternatives, were independent. Fourth, confidence in decisions made for null trials consisting of two targets or two lures was also affected by the strength of these alternatives, with greater confidence when both alternatives were strong rather than weak. This last effect was stronger for pairs composed of two targets but was present also for pairs composed of two lures.

We outlined earlier two hypotheses of how confidence in 2AFC may depend on strength of the two alternatives: the relative account and the absolute account. The relative account fails to provide an explanation for the majority of our findings. This account stipulates that in assigning

confidence in 2AFC recognition decisions, participants weigh two alternatives and give higher confidence judgments when the difference in evidence supporting each alternative is larger. The relative account predicts confidence to be higher for stronger chosen alternatives, as observed in our data, but it also clearly predicts confidence to be higher when the unchosen alternative is weaker. The fact that confidence increased with increasing strength of the unchosen alternative is not predicted by the relative account in its simplest formulation. Also, the pure version of the relative account predicts no strength effect for the null trials in which strength of support for all alternatives should, on average, be equal and thus confidence should be generally low. In the present study, not only was confidence for target – target null trials clearly high in general, but an effect of strength was clearly detected for both types of null trials.

The results, by contrast, are fully consistent with the absolute account by which the strength of the chosen alternative is fundamental for the magnitude of the confidence judgment. This account also clearly predicts the main effect of strength of the chosen alternative that was observed in the present experiment for both correct and incorrect recognition trials. It also provides a straightforward account for the results of null trials, for which the strength of the chosen alternative is the same as the manipulated strength of the overall pair.

Most importantly, the absolute account predicts the counterintuitive finding of increased confidence with increased strength of the unchosen alternative. Consider first the correct trials for which targets are endorsed. In the presence of weak lures, many targets will have enough strength to exceed that of those lures, even if that evidence is relatively weak. However, when lures become strong by virtue of their similarity to repeated parent words, only a few targets supported by relatively strong memory evidence will be correctly endorsed. Because confidence according to the absolute account is based solely on the strength of the chosen item, this mean that strong lures will increase confidence in chosen targets compared to weak lures. The same logic applies to incorrect trials in which the average strength of the chosen lure should be higher when these lures are chosen

in the presence of strong rather than weak targets. Indeed, this finding was also observed in Experiment 1.

As signalled earlier, one issue that needs to be discussed in relation to the present findings is the problem of recall-to-reject. Some studies have documented the process of recall-to-reject operating in the plurals paradigm (Rotello & Heit, 2000), although other studies have shown that its role may be small unless explicitly encouraged by the instructions (Rotello & Heit, 1999). Needless to say, our instructions were mute on this issue of recall-to-reject but we still cannot exclude the possibility that participants used this strategy in the present task. What consequences would this have for our results? Recall-to-reject should be more successful if the parent word is strengthened by multiple presentations. The consequences for performance of this strengthening are unclear because stronger parent words should simultaneously increase the strength of lures, facilitating incorrect responding, and increase recall-to-reject, reducing incorrect responding. Several studies used an old/new associative recognition task in which pairs of two re-paired words were tested that were characterized by increased familiarity and increased effectiveness of recall-to-reject due to multiple presentations of the original pairs containing re-paired words. The balance of these two effects commonly led to the manipulation of strength having no effect on performance (Buchler, Faunce, Light, Gottfredson, & Reder, 2011; Kelley & Wixted, 2001), although an increase in incorrect responding has also been observed (Malmberg & Xu, 2007). It thus remains possible that increased strength of lures whose parent words were repeated masked the increased effectiveness of recall-to-reject in our data, still leading to the observed pattern of better recognition performance for weaker lures.

Given that the possibility of participants using of recall-to-reject strategy in our study, we need to consider the consequences recall-to-reject could have for the confidence patterns. If strengthening parent words increases the effectiveness of recall-to-reject, then this could translate into increased confidence despite similarity of these lures to a strongly encoded parent word. This

additional confidence gained from effective recall-to-reject could account for the pattern of higher confidence when the unchosen lure is strong as well as the pattern of increased confidence for strong rather than weak lure-lure pairs – two patterns that are not handled by the relative account of confidence in its basic form but that are predicted by the absolute account. However, it needs to be stressed that recall-to-reject is a monitoring strategy that operates on lures and should not affect the processing of targets. Thus, the recall-to-reject augmentation of the relative account of confidence can help to explain the pattern of confidence when lures are rejected – either in standard or null pairs – but it does not change the predictions of this account when targets are rejected – either in standard or null pairs. A vital thing to note, however, is that patterns observed in our study were the same whether considered for unchosen lures or for unchosen targets. Thus, the relative account with the additional component of recall-to-reject cannot handle the pattern of increased confidence on incorrect trials when the strength of the unchosen target is varied and neither can it handle the pattern observed for null target-target pairs. Thus, although the recall-to-reject account, mostly due to its flexibility inherent in the idea of two opposing effects producing an unpredictable balance, could possibly explain some of our results, it is in fact not able to provide an overarching account of the data.

Experiment 2

Experiment 2 was conducted to extend the results of Experiment 1 to a different type of confidence judgment. We used exactly the same design and materials, varying independently the strength of the targets and lures in the plurals paradigm and including two groups that were tested with null trials composed of two targets or two lures. The only thing that was changed in the present study was the format of responding in a 2AFC recognition test. Experiment 1 asked participants to first endorse the target and then rate confidence in their decision. This is a metacognitive confidence scale by which participants rate the correctness of their responses. It is possible that at least some results observed in Experiment 1 were due to this metacognitive nature of the scale that puts a

chosen alternative in the focus of attention required to make a metacognitive judgment. It is of interest whether the same patterns would be observed with a scale that deemphasizes the metacognitive focus on the decision concerning the chosen alternative and instead requires greater consideration of differences between two alternatives. In Experiment 2 we thus asked participants to respond by using a one-step confidence scale by which they judged how likely it was that an item presented on the left/right was studied.

Method

Participants. Fifty-two undergraduate students of Cardiff University participated in this experiment for course credit or monetary compensation, with 26 assigned to each of the two experimental groups. Results of one person from the lure-lure group were lost due to an experimenter's error.

Materials and procedure. Experiment 2 utilized the same materials as Experiment 1. The procedure was also the same as in Experiment 1, with only one exception: instead of a two-step procedure, with participants first choosing one word and indicating confidence that this choice was accurate, a one-step procedure was implemented. This meant that the choice of the studied word was made using the confidence scale. The final part of test instructions was adapted to convey this change:

> To indicate your answer, use the confidence scale. Type in "1" if you are sure that the word on the left is the correct answer and type in "6" if you are sure that the word on the right is the correct answer. Use numbers 2-5 to indicate intermediate levels of confidence.

Results

Accuracy. The accuracy scores for the present experiment were derived from participants' confidence judgments. Thus, confidence judgments 1-3 were re-coded as endorsements of the

alternative presented on the left and confidence judgments 4-6 were re-coded as endorsements of the alternative presented on the right. The means of the derived accuracy scores are presented in Table 1. A 2 (target strength) x 2 (lure strength) x 2 (group) mixed ANOVA on the derived hit rate for standard pairs yielded a significant main effect of target strength, $F(1, 49) = 16.07$, $MSE = .04$, $p < .001$, $\eta_p^2 = .25$. This effect confirms that, predictably, recognition performance was better for strong targets, $M = .67$, $SD = .16$, versus weak targets, $M = .56$, $SD = .13$. Contrary to the results of Experiment 1, the main effect of lure strength was not significant, $F(1, 49) = 3.06$, $MSE = .03$, $p = .086$, $\eta_p^2 = .06$, although the numerical trend was in the same direction as the effect observed in Experiment 1, with better performance for trials with weak lures, $M = .64$, $SD = .11$, versus strong lures, $M = .60$, $SD = .15$. No other effect was significant, largest $F(1, 49) = 1.39$, $p = .24$.

Confidence. The confidence data were also re-coded in order to analyze the effects of the strength of the chosen and unchosen alternatives in the same way as in Experiment 1. Ratings 1 and 6 were re-coded as reflecting highest confidence (value of 3) for the 'chosen' alternative, ratings 2 and 5 were re-coded as reflecting moderate level of confidence (value of 2) and ratings 3 and 4 were re-coded as reflecting lowest confidence (value of 1). These derived means for the confidence data can be found in Table 1. They were again analyzed separately for trials in which a correct and incorrect alternative were chosen. For correct trials on which targets were chosen, a 2 (target strength) x 2 (lure strength) x 2 (group) mixed ANOVA yielded a significant main effect of target strength, $F(1, 49) = 60.74$, $MSE = 0.11$, $p < .001$, $\eta_p^2 = .55$, and a significant main effect of lure strength, $F(1, 49) = 8.64$, $MSE = 0.14$, $p = .005$, $\eta_p^2 = .15$. No other effect was significant, largest $F(1, 49) = 3.14$, $p = .083$, for the target strength x lure strength interaction. These results replicate the results of Experiment 1. The main effect of target strength shows that participants were more confident in their correct recognition decisions when endorsed targets were strong, $M = 2.37$, $SD = 0.33$, rather than weak, $M = 2.02$, $SD = 0.37$. The main effect of lure strength shows that participants were also more confident in their correct recognition decisions when the unchosen lures were strong, $M = 2.27$, $SD = 0.37$, rather than weak, $M = 2.12$, $SD = 0.36$.

For the analysis of incorrect trials on which lures rather than targets were chosen five participants needed to be excluded due to missing cells (i.e., perfect recognition performance in one of the conditions). A 2 (target strength) x 2 (lure strength) x 2 (group) mixed ANOVA on the means of derived confidence judgments from the remaining 46 participants yielded a significant main effect of lure strength, $F(1, 44) = 5.28$, $MSE = 0.18$, $p = .026$, $\eta_p^2 = .11$, again showing that confidence was higher when the chosen alternative was strong, $M = 2.01$, $SD = 0.44$, rather than weak, $M = 1.87$, $SD = 0.45$. The main effect of target strength was marginal, $F(1, 44) = 3.76$, $MSE = 0.13$, $p = .059$, $\eta_p^2 = .08$, but a clear numerical trend suggested higher confidence when the unchosen target was strong, $M = 2.00$, $SD = 0.44$, rather than weak, $M = 1.89$, $SD = 0.42$, consistent with the results of Experiment 1. No other effect was significant, largest $F(1, 44) = 2.67$, $p = .11$.

The derived confidence means for null pairs can be found in Table 2. Derived confidence judgments for null trials were analyzed with a 2 (group: target-target vs. lure-lure) x 2 (strength of alternatives: strong vs. weak) mixed ANOVA. This yielded a significant main effect of strength of alternatives, $F(1, 49) = 18.82$, $MSE = 0.08$, $p < .001$, $\eta_p^2 = .28$. The main effect of group was not significant, $F(1, 49) = 3.95$, $MSE = 0.20$, $p = .053$, $\eta_p^2 = .07$, and neither was the interaction, $F < 1$. The numerical trend that was present in the analysis of the type of trials suggested slightly higher confidence for target-target pairs, $M = 2.16$, $SD = 0.28$, than for lure-lure pairs, $M = 1.98$, $SD = 0.35$. These results differ from the results of Experiment 1 inasmuch as a significant interaction that was obtained there pointed to a larger effect of strength for two-target pairs compared to two-lure pairs. The main observation remains, however, that a clear effect of strength was observed, with higher confidence for strong pairs, $M = 2.20$, $SD = 0.25$, than for weak pairs, $M = 1.96$, $SD = 0.30$, for trials for which an average difference in strength for two alternatives should be null.

Discussion

The results of Experiment 2 largely replicate the results of Experiment 1. Two effects that were significant in Experiment 1 – the effect of lure strength on accuracy and the effect of target

strength on confidence when the lure was endorsed – were not significant here, although clear trends were observed that were consistent with these previous results. One effect that was not replicated here was an interaction of strength and type of the null trial. It is unclear why this effect was not present here; in any case, this remains of secondary importance for the purpose of the present study.

The fact that patterns of data observed in Experiment 1 were generally replicated indicates that variations in the format of responding do not fundamentally change the processes by which confidence judgments are made, at least under present conditions. We speculated that a metacognitive framework of confidence by which participants are asked to focus on their responses rather than the status of the tested items could have reduced the importance of the relative strategy of assigning confidence in Experiment 1. This hypothesis is refuted by the current results which again revealed patterns that are not predicted by the relative account of confidence in 2AFC recognition. Of particular interest is again the counterintuitive finding that confidence was greater when the unchosen alternative was strong rather than weak. Again, the absolute account, but not the relative account, is well-posed to explain this finding as well as the overall pattern of results.

The relative account of confidence neither predicts nor explains the patterns observed here. However, this account could gain additional flexibility if another factor was to be taken into account. As described earlier (see footnote 1), patterns of accuracy-confidence dissociations can be handled by this account if variances of the distributions of differences in evidence strength are considered. This raises a question of whether our manipulation of strength might have led to changes in variance of memory evidence associated with targets and lures and consequently to changes in differences in evidence supporting two alternatives. Any such difference in variances could have consequences for interpreting our results. The null trials could serve as an example of such interpretative consequences. If strengthening by repetition increases variances of target and lure distributions positioned on the dimension of memory evidence, then it also leads to an increase in variance of a

distribution of pairs positioned on the dimension of within-pair differences in this evidence. As argued for the case of choice similarity effects described by Tulving (1981), under the relative account greater variance of this distribution translates into higher average confidence (Clark, 1997). This is because more pairs are then characterized by a large difference in evidence, which, according to the relative account, should serve as a basis for high confidence ratings. One could thus argue that the effect of strength on null trials reflected this increased variance rather than a simple strategy of basing confidence on the strength of the chosen alternative without considering the difference in strength between alternatives, as supposed by the absolute account.

The question of changes in variance in memory evidence due to repetition is an empirical one. Previous investigations into this issue suggested that the repetition manipulation is not associated with any changes in variance of memory evidence (e.g., Ratcliff, Sheu, & Gronlund, 1992; Starns, 2014; Starns & Ratcliff, 2014). However, these previous studies used standard materials employed in recognition studies – unrelated words – and it is possible that these effects are different in the plurals task. For this reason, in Experiment 3 we addressed the issue of whether repetition manipulation affects variances of memory evidence for both targets and lures in the plurals paradigm.

Experiment 3

In the present experiment we assessed how the repetition of parent words affects variance of memory evidence associated with targets and lures in the plurals paradigm. To remain consistent with Experiments 1 and 2, we used exactly the same materials and design, again manipulating the strength of the tested items and including null pairs. Variances for different item types – in our case, strong and weak items – can be compared by plotting z-receiver operating characteristic (z-ROC) curves that can be derived from confidence judgments (a proxy for a bias manipulation) in old/new recognition. The slope of the z-ROC is determined by the ratio of variances for the item distributions, so that, assuming underlying Gaussian distributions, a slope of 1 reveals that the variances are equal,

and deviations from this value suggest unequal variances. Thus, in Experiment 3 we changed the format of responding, this time asking participants to rate confidence that the item is new or old on a 1-6 scale, separately for each word in the pair. Hence, we effectively changed our procedure to old/new recognition, while preserving the display format of pairs for the purpose of consistency across experiments.

Despite the similarity of our procedure to standard old/new recognition, our z-ROC analysis was somewhat unconventional. Specifically, rather than using confidence to plot the z-scores for hit rates versus false alarm rates to determine the ratio of variances for target and lure distributions, we plotted z-scores for the endorsement rates of strong versus weak targets on one z-ROC, and z-scores for the endorsement rates of strong versus weak lures on the other. We did this to investigate how strength affected the target variance and the lure variance in the 2AFC task used in Experiments 1 and 2. Target-target and lure-lure z-ROCs allow us to assess separately the role of strength on targets and lures, respectively. This, in turn, allows us to determine whether the variance of the distribution of evidence differences in 2AFC was affected by strength, which is important in evaluating the viability of the relative account.

Method

Participants. 68 Cardiff University students participated in this experiment for course credit or monetary compensation, with 34 participants assigned to each of the experimental groups, defined again by the presence of either target-target or lure-lure pairs.

Materials and procedure. The materials were the same as in Experiments 1 and 2. The procedure was also the same as in previous experiments, save for the format of responding at test. Participants were instructed ‘to decide for EACH word whether it was presented before, using a scale from 1 to 6, where 1 means "certain new" and 6 means "certain old"’. They were also explicitly

informed that two old or two new words could be presented at the same trial, and that in such a case they should give both of these words high (or low) ratings.

Results

Figure 2 shows z-ROCs constructed from data pooled across participants for targets (panel A) and lures (panel B). The slopes calculated from group data using Principal Component Analysis (Vokey, 2016) equalled 0.967 for targets, and 0.794 for lures. We further calculated two z-ROC slopes for each participant – one for targets, and one for lures. When averaged across participants, the slope for targets equalled 0.959 (*SD* = 0.37). A Bayesian t-test conducted using the JASP software (JASP Team, 2016) showed that the null hypothesis of the mean being not different from 1 was 5.05 times more likely than the alternative hypothesis, consistent with variances for strengthened and non-strengthened targets being comparable. The slope for lures equalled 0.806 (*SD* = 0.31), which was 9,152 times as likely under the alternative hypothesis as compared to the null. This analysis confirmed that strengthened lures had a greater variance than non-strengthened lures.

Discussion

In the present experiment we assessed how the repetition manipulation of parent words affects the variance of the distribution of memory evidence associated with targets and lures in the plurals paradigm. Replicating previous research that looked at a similar issue in a standard recognition test (e.g., Starns, 2014; Starns & Ratcliff, 2014), we found the repetition manipulation to have little effect on variance for targets. Going beyond these results, we also found that repetitions of parent words affected the variance of the distribution of memory evidence for lures derived from these parent words. The variance of memory evidence for strong lures was larger when the parent word was repeated rather than presented only once.

The main question is whether the variance effects of repetitions revealed in the present study can account for the patterns of confidence described in Experiments 1 and 2. We argue that

they cannot. The effects of repetition were observed for lures which could potentially affect confidence patterns in all cases in which comparisons are conducted across different levels of lure strength: the effects of lure strength on confidence for the lure-lure null trials, the effect of the unchosen alternative when the target is chosen on standard trials, and the effect of the strength of the chosen lure itself on standard trials. In these cases, greater variance for stronger lures may translate into greater variance in the distribution of differences in evidence and thus greater confidence. The reason why this does not account for the full pattern of results is again the fact that all results documented in Experiments 1 and 2 were largely independent of whether they were considered specifically for lures or targets. No effect of repetition was observed on the variance of evidence for targets. Thus, comparisons that examined how the target strength affected confidence – specifically null-trial comparisons involving target – target pairs and standard-trial comparisons if the lure was chosen – were unaffected by any changes in variance.[3]

General Discussion

The present study examined the basis of confidence judgments in 2AFC recognition tests. We employed the plurals paradigm, which allowed us to manipulate the strength of targets and lures by varying the number of presentations of their parent words. We then assessed the influence of strength of both chosen and unchosen alternatives on confidence in standard pairs as well as the role of strength in confidence for null trials composed of either two lures or two targets. The most important results, documented in Experiments 1 and 2, indicated that confidence in 2AFC recognition was dependent both on the strength of the chosen alternative and the strength of the unchosen alternative but not on the difference in strength between these alternatives. First, we found that the stronger the chosen alternative, the higher the confidence. This relationship held regardless of whether the chosen alternative was a lure or a target. Second, for standard target-lure pairs, confidence was higher when the unchosen alternative was strong rather than weak and this

[3] Note that the effect of the strength of the unchosen target did not interact with the effect of strength of the chosen lure.

difference also occurred independently of whether the chosen alternative was a target or a lure. Third, for null trials consisting of two targets or two lures, confidence was higher for two strong rather than two weak alternatives.

The results summarized here were used to test two competing accounts of confidence in 2AFC recognition test. According to the relative account, confidence depends on the difference in strength of evidence supporting two alternatives on each recognition trial. Several findings argue against this account. First, the relative account predicts that for null trials – trials consisting of two items of the same average strength and thus preserving the average difference in evidence at zero – the strength of individual alternatives should not matter for confidence. The relative account thus fails to explain why confidence was consistently higher for pairs composed of two strong targets rather than two weak targets and also for pairs composed of two strong lures rather than two weak lures. Second, the relative account predicts that increasing the strength of the unchosen alternative should reduce confidence. In a stark contrast to this prediction, the results revealed greater confidence whenever recognition endorsements were made in the presence of strong rather than weak alternatives.

Although the relative account did not fare well in its purest form, we also discussed two mechanisms by which the relative account could be modified to explain these results. First, confidence in the presence of a strong unchosen alternative might have increased because of a recall-to-reject mechanism. Given the nature of our experimental task, it is likely that the recall-to-reject strategy was employed on some of the test trials. This mechanism would also be consistent with the difference in variances of strong and weak lure distributions (see Experiment 3): recall-to-reject should be more likely for strengthened than for non-strengthened lures, decreasing the evidence for some of these lures and thus increasing the variance of the whole distribution. However, the recall-to-reject mechanism cannot fully account for the pattern of results found in our data. Recall-to-reject can operate only for lures and thus it does not explain why confidence in the

presence of a stronger unchosen alternative was increased both when this unchosen alternative was a lure and when it was a target. Similarly, this mechanism cannot account for the pattern observed for target-target pairs for which there should be no recall-to-reject. Second, confidence might have been affected by increased variance of evidence supporting strong compared to weak alternatives. In Experiment 3 we directly assessed whether repetitions of parent words affect variances of evidence supporting targets and lures. Although the results of this experiment revealed increased variance of evidence for lures corresponding to repeated parent words, no such increase was observed for targets. In effect, the variance argument suffers from the same shortcoming as the recall-to-reject argument which is its failure to account for the symmetry of the results across the lure/target status.

The second hypothesis under consideration was the absolute account by which confidence in 2AFC recognition depends primarily on the strength of just the chosen alternative. This hypothesis provides a good account of our results. It is straightforwardly consistent with the observation that confidence depends on the strength of the chosen alternative for both standard and null pairs. More interestingly, it also predicts the counterintuitive finding that confidence is increased in the presence of a strong unchosen alternative. This happens because the presence of strong unchosen alternatives has implications for the strength of the chosen alternative: Alternatives that are chosen from pairs with strong unchosen alternatives are necessarily stronger on average than alternatives that are chosen from pairs with weak unchosen alternatives. This filtering out of weaker alternatives by strong unchosen alternatives translates into higher confidence.

The fact that confidence in 2AFC recognition seems at least sometimes to depend on the strength of the chosen alternative has important consequences for our understanding of a confidence-accuracy relationship. There are various ways of asking the question of whether confidence tracks accuracy (see Roediger, Wixted, & DeSoto, 2012, for a detailed discussion). One can be interested whether people are more confident in their more accurate responses or whether

people who express greater confidence in their responses are actually more accurate. However, from the perspective of studies into basic memory processes, perhaps the most interesting question is whether conditions that lead to more accurate responding lead also to higher confidence being expressed for these responses. Roediger et al. have recently summarized research on this topic by arguing that "the exceptions are sufficiently few that we can safely conclude that when an independent variable affects accuracy of memory reports, subjects' confidence in those reports will virtually always be affected the same way (however, see Tulving, 1981, for a somewhat different case)" (p. 98). However, as noted by Busey et al. (2000), confidence is likely to track differences in performance only if confidence is shaped by exactly the same factors as those that determine the accuracy of memory responses. The accuracy of responses in 2AFC recognition test depends on the difference in evidence supporting two alternatives. The larger this difference is, the more accurate participants' responses can potentially be. If confidence does not depend on this difference, but instead depends only on the strength of the chosen response, then this creates a fertile ground for confidence-accuracy dissociations. Indeed, our results point to a situation in which increasing the strength of the lure reduces the accuracy of recognition decisions (a significant effect in Experiment 1 and a marginal effect, p = .086, in Experiment 2), yet it increases confidence in these decisions.[4]

The current results provide a particularly striking example of confidence-accuracy dissociation where lower accuracy is accompanied by higher confidence. We argue that this happens because confidence is based on different factors than those that determine recognition accuracy (cf. Busey et al., 2000). To be sure, both confidence and accuracy can depend on the same type of memory information that we refer to simply as strength, without postulating the necessity of

[4] Across experiments we analyzed confidence in correct and incorrect responses separately. The negative relationship between accuracy and confidence across experimental manipulations is thus evident when accuracy is contrasted with confidence in correct responses. We also reanalyzed confidence results collapsing across correct and incorrect recognition decisions. This collapsed analysis bolsters our conclusions based on confidence in correct responses as in both Experiment 1 and 2 overall confidence was reliably higher when lures were stronger [Experiment 1: $F(1, 50) = 28.62$, $MSE = 0.24$, $p < .001$, $\eta_p^2 = .40$, for the main effect of lure strength on overall confidence, with higher confidence when lures were strong, $M = 3.86$, $SD = 0.74$, rather than weak, $M = 3.50$, $SD = 0.69$; Experiment 2: $F(1, 49) = 13.33$, $MSE = 0.08$, $p = .001$, $\eta_p^2 = .21$, with higher confidence when lures were strong, $M = 2.20$, $SD = 2.06$, rather than weak, $M = 2.06$, $SD = .032$].

considering unique contributions of familiarity and recollection (see Dobbins et al., 1998; Heathcote et al., 2010; Migo, Montaldi, Norman, Quamme, & Mayes, 2009; for arguments that this distinction is necessary to account for other findings obtained with 2AFC recognition tests). However, the important point is that this memory information contributes differentially to accuracy and confidence. Whereas accuracy is necessarily dependent on a difference in strength of two alternatives which sets the upper boundary on how accurate recognition decisions can be, confidence can utilize the strength information in a different way due to strategic factors that are inherent in all metacognitive judgments. As repeatedly argued by Koriat (e.g., 1993, 1997), all metacognitive judgments are inferential and thus they are not a straightforward function of factors determining accuracy. Participants in our study considered only the strength of the chosen alternative when forming their confidence judgments, discounting all memory information associated with the unchosen alternative, which resulted in a counterintuitive negative confidence-accuracy relationship.

The majority of previous demonstrations of confidence-accuracy dissociations have shown how a factor that does not contribute to memory performance is factored in when making confidence judgments. For example, the work conducted by our research group (Hanczakowski, Pasek, Zawadzka, & Mazzoni, 2013; Hanczakowski, Zawadzka, & Coote, 2014; Hanczakowski, Zawadzka, & Macken, 2015) demonstrated that confidence can be affected by familiarity of contextual elements that accompany recognition trials even when this familiarity is spurious and does not enhance recognition performance. But our recent line of investigation into how recognition memory is shaped by social cues demonstrates clearly that factors affecting performance may be discounted when making confidence judgments. Krogulska, Zawadzka, and Hanczakowski (2016) showed that in a memory conformity paradigm in which by and large accurate social cues are provided (see Jaeger, Lauris, Selmeczy, & Dobbins, 2012), participants incorporate these cues into their recognition decisions, enhancing their recognition performance, but they do not incorporate these cues into their confidence judgments, which results in another confidence-accuracy

dissociation, one in which confidence does not track changes in accuracy (see Zawadzka, Krogulska, Button, Higham, & Hanczakowski, 2016, for related findings). This last dissociation arises because recognition accuracy in a social context depends not only on memory but also on accurate cues provided by external sources. But even when one considers solely memory processes, performance in a recognition task may depend on various components of the recognition process, such as recollection and familiarity, to which confidence may also be differentially sensitive, resulting in dissociations (see Reinitz et al., 2011). Indeed, in a recent study by Beaman, Hanczakowski, and Jones (2014), in which two types of recognition test were employed – one in which accuracy was mainly a function of familiarity of individual items and one in which accuracy was shaped by retrieval of specific associations – another pattern of negative confidence-accuracy relationship was observed. Together, all these studies demonstrate that positive confidence-accuracy relationship across various experimental conditions may be a less prevalent pattern than it is commonly considered.

Our results can also be considered in the light of a discussion about the basis of recognition decisions in 2AFC recognition task. Researchers sometimes use 2AFC tasks as they are thought to provide a measure of memory ability uncontaminated by changes in bias. For example, in studies of the revelation effect, the increased propensity to endorse probes presented after performing a simple operation on them (e.g., Peynircioğlu & Tekcan, 1993; Westerman & Greene, 1996), it has been argued that the use of 2AFC recognition task reveals a memory impairment caused by the revelation manipulation that is masked by a change in bias when simple recognition tasks are used (Hicks & Marsh, 1998). The argument for the superiority of 2AFC recognition is that, in these kinds of tasks, participants weigh two alternatives and simply choose one that is supported by more evidence – a relative strategy for arriving at recognition decisions. However, recently Starns, Staub, and Chen (2015; see also Jou, Flores, Cortes, & Leka, 2016, for related findings) proposed that absolute strategy can also play a prominent role in 2AFC tasks. Using the eye-tracking methodology, Starns et al. have shown that participants faced with a 2AFC task often fixate on only one of the alternatives

before arriving at a recognition decision, which suggests a use of an absolute strategy for arriving at recognition decisions. The use of an absolute strategy implies that performance in 2AFC tasks may not be a bias-free measure of underlying memory processes. Our results have no straightforward bearing on this debate because it is possible that participants rely heavily on the relative strategy for arriving at a recognition decision and then focus exclusively on the chosen alternative in what is in effect an absolute strategy of arriving at a confidence assessment. However, it seems at least plausible that the reason why we found strong evidence for the absolute strategy for confidence patterns is that our participants did rely on the absolute strategy throughout the task, often not taking the unchosen alternative into account before endorsing an alternative they considered first and providing their confidence judgment.

The investigations of performance and confidence in 2AFC recognition judgments gained most prominence in the paradigm investigating the role of choice similarity developed by Tulving (1981). As discussed earlier, the analysis of results obtained in this paradigm concentrate on how manipulating the similarity of lures to the particular target tested on a given trial increases accuracy while reducing confidence. However, data reported both by Tulving and Dobbins et al. (1998) indicate that similarity of lures to the overall set of the tested items also produces an unexpected finding by which correct decisions made in the presence of weak and strong lures are made at the same level of confidence. These results do not seem consistent with either the relative or the absolute strategy of assigning confidence. We suggest thus that they can reflect a mix of these two strategies. The argument is that when multiple alternatives are presented for a recognition task, people may choose the most adequate strategy based on factors like overall level of memory, perceptual features of the tested materials, number of alternatives (see Charman, Wells, & Joy, 2011; Hanczakowski, Zawadzka, & Higham, 2014) or some combination of all those factors. The investigations into choice similarity commonly used perceptually rich materials in the form of halves of pictures. It is possible that with much perceptual detail available for presented alternatives participants are more inclined to engage in relational analysis in support of both recognition

decisions and confidence assessments. Indeed, a recent similar line of investigation into confidence judgments for line-up decisions (Horry & Brewer, 2015) provided support for the relative account of confidence that has been absent from our results documented here. The line-up situation requires judgments concerning faces, another example of perceptually rich materials. It is worth noting that a study by Heathcote et al. (2009), that utilized the choice similarity paradigm of Tulving (1981) with faces as study materials, documented a pattern of higher overall confidence for A-X' than A-B' pairs, which is inconsistent with our results as well as results of Tulving and Dobbins et al. (1998).

The overview presented here points to a type of materials used in a given experiment as a likely moderator of strategies used to arrive at confidence judgments. This suggestion underscores the limitation of our study that used only one specific set of materials. Clearly, additional studies are necessary for elucidating the basis of confidence judgments across variations in to-be-remembered stimuli. A particularly important direction should be to thoroughly investigate confidence judgments for 2AFC decisions in studies using standard words as targets and lures. Studies that used such standard materials in 2AFC tests have employed confidence judgments to assess the validity of various models of recognition memory (see Jang, Wixted, & Huber, 2009; Smith & Duncan, 2004) and thus the investigation of how such confidence judgments are arrived at may have important consequences for this particular method of model validation. Related to this point, a recent study by Jou et al. (2016) looked at recognition decisions in 2AFC tests utilizing standard word materials. Their Experiment 1B included null trials along with old-new pairs and asked participants to make confidence judgments. One interesting finding from this experiment was that participants were more confident in their decisions for old-old pairs than they were for their recognition decisions for old-new pairs. As noted by the authors (p. 36): "If the relative familiarity were the basis for the confidence rating, the difference in familiarity between the two items in the normal pairs is much greater than this difference in the both-old pairs, and therefore the former should be given a higher confidence rating than the latter, which, however, was not the case". This analysis led Jou et al. to conclude that confidence could have been based on the absolute familiarity of a chosen item, which

remains entirely consistent with the results of the present investigation.[5] Given the difference in materials across these studies, the results of Jou et al. suggest certain generalizability of our findings. Still, the issue of how variable the use of relative and absolute strategies in 2AFC tasks is under changing circumstances awaits further empirical scrutiny.

[5] We have become aware of the details of the study by Jou et al. (2016) when revising the present paper.

References

Beaman, C. P., Hanczakowski, M., & Jones, D. M. (2014). The effects of distraction on metacognition and metacognition on distraction: Evidence from recognition memory. *Frontiers in Psychology*, 5:439. http://dx.doi.org/10.3389/fpsyg.2014.00439

Brewer, G. A., Marsh, R. L., Clark- Foos, A., & Meeks, J. T. (2010). Noncriterial recollection influences metacognitive monitoring and control processes. *The Quarterly Journal of Experimental Psychology, 63,* 1936-1942. http://dx.doi.org/10.1080/17470210903551638

Buchler, N. G., Faunce, P., Light, L. L., Gottfredson, N., & Reder, L. M. (2011). Effects of repetition on associative recognition in young and older adults: Item and associative strengthening. *Psychology and Aging, 26,* 111-126. http://dx.doi.org/10.1037/a0020816

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26-48. http://dx.doi.org/10.3758/BF03210724

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identification. *Law and Human Behavior, 35,* 479-500. http://dx.doi.org/10.1007/s10979-010-9261-1

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 232-238. http://dx.doi.org/10.1037/0278-7393.23.1.232

Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence-accuracy inversions in scene recognition: A remember-know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1306-1315. http://dx.doi.org/10.1037/0278-7393.24.5.1306

Hanczakowski, M., Pasek, T., 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., Collie, H., & Macken, B. (2016). Metamemory in a familiar place: The effects of environmental context on feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000292

Hanczakowski, M., Zawadzka, K., & Coote, L. (2014). Context reinstatement in recognition: Memory and beyond. *Journal of Memory and Language, 72,* 85-97. http://dx.doi.org/10.1016/j.jml.2014.01.001

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., & Macken, B. (2015). Continued effects of context reinstatement in recognition. *Memory & Cognition, 43,* 788-797. http://dx.doi.org/10.3758/s13421-014-0502-2

Heathcote, A., Bora, B., & Freeman, E. (2010). Recollection and confidence in two-alternative forced choice episodic recognition. *Journal of Memory and Language, 62,* 183-203. http://dx.doi.org/10.1016/j.jml.2009.11.003

Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review, 16,* 824-831. http://dx.doi.org/10.3758/PBR.16.5.824

Horry, R., & Brewer, N. (2015, June). *Understanding confidence judgments in lineup decisions through manipulations of target-filler similarity*. Paper delivered at the biennial meeting of the Society for Applied Research in Memory and Cognition, Victoria, BC.

Jaeger, A., Lauris, P., Selmeczy, D., & Dobbins, I. G. (2012). The costs and benefits of memory conformity. *Memory & Cognition, 40,* 101-112. http://dx.doi.org/10.3758/s13421-011-0130-z

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138,* 291-306. http://dx.doi.org/10.1037/a0015525

JASP Team (2016). JASP (Version 0.7.5 Beta 2)[Computer software]

Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in Two-Alternative-Forced-Choice recognition: Is the test criterion-free? *Acta Psychologica, 167*, 30-44. http://dx.doi.org/10.1016/j.actpsy.2016.03.014

Kelley, R. & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 701-722. http://dx.doi.org/10.1037/0278-7393.27.3.701

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

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124,* 311-333. http://dx.doi.org/10.1037/0096-3445.124.3.311

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

Krogulska, A., Zawadzka, K., & Hanczakowski, M. (2016). *Whom can I trust? Discrimination of social source reliability in a memory conformity paradigm.* Manuscript submitted for publication.

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

Malmberg, K. J., & Xu, J. (2007). On the flexibility and fallibility of associative memory. *Memory & Cognition, 35,* 545-556. http://dx.doi.org/10.3758/BF03193293

Migo, E., Montaldi, D., Norman, K. A., Quamme, J., & Mayes, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *The Quarterly Journal of Experimental Psychology, 62,* 1198-1215. http://dx.doi.org/10.1080/17470210802391599

Peynircioğlu, Z. F., & Tekcan, A. I. (1993). Revelation effect: Effort or priming does not create the sense of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 382-388. http://dx.doi.org/10.1037/0278-7393.19.2.382

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518-535. http://dx.doi.org/10.1037/0033-295X.99.3.518

Reinitz, M. T., Peria, W. J., Séguin, J. A., & Loftus, G. R. (2011). Different confidence-accuracy relationships for feature-based and familiarity-based memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 507-515. http://dx.doi.org/10.1037/a0021961

Reinitz, M. T., Séguin, J. A., Peria, W. J., & Loftus, G. R. (2012). Confidence-accuracy relations for faces and scenes: Roles of features and familiarity. *Psychonomic Bulletin & Review, 19,* 1085-1093. http://dx.doi.org/10.3758/s13423-012-0308-9

Roediger, H. L., Wixted, J. T., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84-118). Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199920754.003.0004

Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject? *Journal of Memory and Language, 40,* 432-453. http://dx.doi.org/10.1006/jmla.1998.2623

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence for ROC curves. *Journal of Memory and Language, 43,* 67-88. http://dx.doi.org/10.1006/jmla.1999.2701

Smith, D. G., & Duncan, M. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 615-625. http://dx.doi.org/10.1037/0278-7393.30.3.615

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & Cognition, 42,* 1357-1372. http://dx.doi.org/10.3758/s13421-014-0432-z

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language, 70,* 36-52. http://dx.doi.org/10.1016/j.jml.2013.09.005

Starns, J. J., Staub, A., and Chen, T. (2015, November). *Implication of response time and eye movement data for models of forced choice recognition*. Paper delivered at the 54th Annual Meeting of the Psychonomic Society, Chicago, IL.

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior, 20,* 479-496. http://dx.doi.org/10.1016/S0022-5371(81)90129-8

Vokey, J. R. (2016). Single step simple ROC curve fitting via PCA. *Canadian Journal of Experimental Psychology*. Advance online publication. http://dx.doi.org/10.1037/cep0000095

Westerman, D. L., & Greene, R. L. (1996). On the generality of the revelation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1147-1153. http://dx.doi.org/10.1037/0278-7393.22.5.1147

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

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341-1354. http://dx.doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133,* 800-832. http://dx.doi.org/10.1037/0033-2909.133.5.800

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

Table 1.
Mean Recognition Accuracy and Mean Confidence for Correct and Incorrect Recognition Decisions as a Function of the Strength of the Target Alternative (T1 Versus T3), Strength of the Lure Alternative (L1 Versus L3) and Group (Target-Target Versus Lure-Lure) in Experiments 1 and 2. Standard Errors of Means are Given in Parentheses.

	T1-L1	T1-L3	T3-L1	T3-L3
Experiment 1				
Target-target group				
Accuracy	.55 (.04)	.56 (.04)	.70 (.03)	.61 (.05)
Confidence in correct	3.36 (0.23)	3.60 (0.20)	4.02 (0.17)	4.15 (0.24)
Confidence in incorrect	3.08 (0.24)	3.71 (0.20)	3.24 (0.21)	3.87 (0.22)
Lure-lure group				
Accuracy	.56 (.04)	.45 (.04)	.71 (.04)	.67 (.03)
Confidence in correct	3.27 (0.19)	3.54 (0.18)	3.88 (0.20)	4.17 (0.21)
Confidence in incorrect	3.12 (0.21)	3.56 (0.14)	3.33 (0.20)	4.20 (0.22)
Experiment 2				
Target-target group				
Accuracy	.61 (.03)	.52 (.04)	.67 (.03)	.66 (.04)
Confidence in correct	1.88 (0.08)	2.04 (0.11)	2.23 (0.09)	2.51 (0.07)
Confidence in incorrect	1.98 (0.12)	1.85 (0.08)	1.88 (0.10)	2.16 (0.12)
Lure-lure group				
Accuracy	.57 (.03)	.56 (.04)	.70 (.03)	.66 (.04)
Confidence in correct	2.08 (0.09)	2.07 (0.07)	2.28 (0.09)	2.48 (0.07)
Confidence in incorrect	1.83 (0.10)	2.03 (0.10)	1.86 (0.12)	2.11 (0.11)

Note: T1 = target presented once at study; T3 = target presented thrice at study; L1 = lure's parent word presented once at study; L3 = lure's parent word presented thrice at study.

Table 2.
Means of Confidence Judgments for Responses to Null Pairs Consisting Either of Two Targets (Target-Target Group) or Two Lures (Lure-Lure Group) Presented as a Function of Strength of Both Alternatives. Standard Errors of Means are Given in Parentheses.

	T1-T1	T3-T3	L1-L1	L3-L3
Experiment 1				
Target-target group	3.33 (0.18)	4.36 (0.16)	-	-
Lure-lure group	-	-	3.23 (0.14)	3.69 (0.14)
Experiment 2				
Target-target group	2.02 (0.07)	2.30 (0.07)	-	-
Lure-lure group	-	-	1.88 (0.09)	2.08 (0.08)

	Word 1	count	Word 2	
Target-target group	t1	10	t1	
	t3	10	t3	
	t1	10	l1	Lure-lure group
	t1	10	l3	
	t1	10	n	
	t3	10	l1	
	t3	10	l3	
	t3	10	n	
	l1	10	l1	
	l3	10	l3	

Figure 1. Trial types used in Experiments 1-3. “T” denotes targets, “l” denotes lures derived from studied parent words, and “n” denotes new lures. “1” and “3” refer to the number of presentations of the parent words at study.

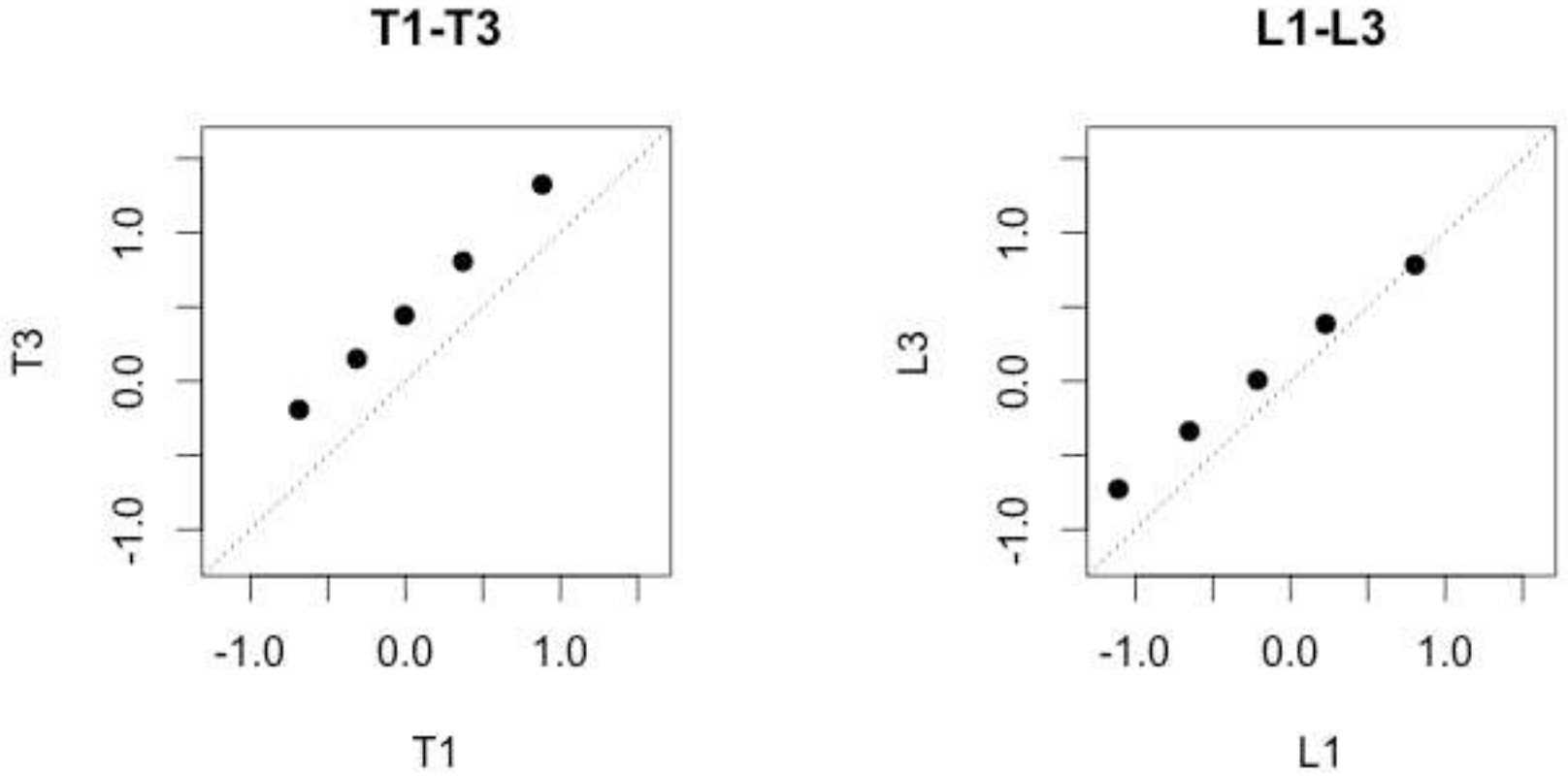

Figure 2. z-ROCs constructed from confidence ratings for non-strengthened (T1) and strengthened targets (T3), and non-strengthened (L1) and strengthened lures (L3).