Investigating Strength and Frequency Effects in Recognition Memory Using Type-2 Signal Detection Theory

Philip A. Higham
University of Southampton

Timothy J. Perfect
University of Plymouth

Davide Bruno
University of Southampton

Criterion- versus distribution-shift accounts of frequency and strength effects in recognition memory were investigated with Type-2 signal detection receiver operating characteristic (ROC) analysis, which provides a measure of metacognitive monitoring. Experiment 1 demonstrated a frequency-based mirror effect, with a higher hit rate and lower false alarm rate, for low frequency words compared with high frequency words. In Experiment 2, the authors manipulated item strength with repetition, which showed an increased hit rate but no effect on the false alarm rate. Whereas Type-1 indices were ambiguous as to whether these effects were based on a criterion- or distribution-shift model, the two models predict opposite effects on Type-2 distractor monitoring under some assumptions. Hence, Type-2 ROC analysis discriminated between potential models of recognition that could not be discriminated using Type-1 indices alone. In Experiment 3, the authors manipulated Type-1 response bias by varying the number of old versus new response categories to confirm the assumptions made in Experiments 1 and 2. The authors conclude that Type-2 analyses are a useful tool for investigating recognition memory when used in conjunction with more traditional Type-1 analyses.

Keywords: recognition memory, signal detection theory, mirror effect, Type 2, ROC analysis

The aim of this article is to introduce the use of Type-2 signal detection theory (SDT) as a means of investigating the cognitive and metacognitive processes involved in old/new recognition memory judgments. To illustrate the insights to be gained from the use of this methodology, we report three experiments. The first is on the word frequency effect, and the second is on the strength effect. In the third experiment, we manipulated response bias by varying the number of old/new response categories available at test, and we conducted this experiment to test some assumptions that we made in Experiments 1 and 2. However, before we introduce the background to these different effects in recognition memory, we begin with a brief review of standard (or Type-1) SDT and its less well-known variant, Type-2 SDT.

SDT

In old/new recognition memory experiments, participants are typically given a list of items to study (e.g., words or pictures) and then administered a recognition test, consisting of (old) target items from the study phase randomly intermixed with (new) distractors. At test, participants are required to judge each individual item as either “old” (previously studied) or “new” (not previously studied). Data from recognition memory experiments are commonly analyzed using SDT (e.g., Macmillan & Creelman, 2005; Rotello, Macmillan, & Reeder, 2004; Yonelinas, 1994). According to SDT, targets and distractors on the recognition memory task are each distributed over a psychological strength-of-evidence dimension, with targets having higher mean strength than distractors (see Figure 1). To make a recognition decision, participants are assumed to adopt a criterion (c in Figure 1) somewhere along the strength-of-evidence dimension. If a test item has strength equal to or above the criterion, it is judged “old”; otherwise, it is judged “new.”

As shown in Figure 1, the criterion splits the target and distractor distributions into two, yielding four areas. The areas under the target distribution above and below the criterion constitute the hit rate (HR) and the miss rate (MR), respectively, whereas the areas under the distractor distribution above and below the criterion constitute the false alarm rate (FAR) and correct rejection rate (CRR), respectively. By convention, recognition memory performance is usually described by the HR and FAR only. The HR and FAR can be used to determine two SDT indices corresponding to bias-free accuracy or discrimination (i.e., the ability to discriminate between targets and distractors) and response bias (i.e., the tendency to respond “old” regardless of the type of item being judged). For example, commonly used indices of discrimination and response bias are d’ and c, respectively. d’ is the distance between the means of the target and distractor distributions in standard deviation units, whereas c is the distance between the...
Type-1 Versus Type-2 SDT

The kind of SDT analysis of recognition data outlined above has been very influential in guiding the development of models of recognition memory (e.g., see Rotello & Macmillan, 2006; Rotello et al., 2004; Verde & Rotello, 2004; Yonelinas, 1994, 1997). Technically, it is described as Type-1 or stimulus-contingent discrimination because the observer is required to discriminate between experimenter-defined stimuli (e.g., targets vs. distractors). However, a lesser known SDT task involves the discrimination between the observer’s own correct and incorrect responses, which is known as Type-2 or response-contingent discrimination (e.g., Clarke, Birdsall, & Tanner, 1959; see Galvin, Podd, Drga, & Whitmore, 2003, for a review). Because of this critical difference in the nature of the discrimination task, the discrimination index for Type-2 SDT is not a bias-free measure of accuracy, as with Type-1 SDT, but rather a metacognitive index reflecting the relationship between confidence and accuracy (e.g., for discussion, see Higham, 2002, 2007; Higham & Arnold, 2007a, 2007b; Higham & Gerrard, 2005; Higham & Tam, 2005). Essentially, it measures the degree to which participants metacognitively monitor the accuracy of their own responses. The meaning of the bias index also varies between Type-1 and Type-2 SDT; the Type-2 bias index measures the overall tendency to judge responses as either “correct” or “incorrect” with confidence ratings.

To illustrate Type-2 discrimination, consider a situation in which participants are required to make an old/new recognition decision and then rate the accuracy of their response by means of a high versus low confidence rating. Someone with good metacognitive monitoring (i.e., good Type-2 discrimination) will tend to assign high confidence to a high proportion of their correct responses and a low proportion of their incorrect responses. These two proportions correspond to the Type-2 HR and FAR, respectively.

Figure 3 shows a flow chart for Type-1 and Type-2 discrimination applied to recognition, although the approach could equally be used in any memory test in which responses can be classified as objectively correct or incorrect. In this figure, and throughout the remainder of this article, we use the subscripts “1” and “2” to denote terms specific to Type-1 and Type-2 discrimination, respectively, and the subscripts “T” and “D” denote terms specific to targets and distractors, respectively. There are two counterintuitive points to note from Figure 3. First, it is possible to calculate a Type-2 HR and MR for distractors (i.e., HR_{D2} and MR_{D2}) and to calculate a Type-2 FAR and CRR for targets (i.e., FAR_{T2} and CRR_{T2}). Readers used to standard, Type-1, SDT in which the HR (HR_{1}) and FAR (FAR_{1}) are specific to targets and distractors, respectively, may...
find this rather peculiar. It occurs because both correct and incorrect responses can be made to both targets and distractors. For instance, it is possible for participants to believe confidently that a target item is new: In Type-2 signal detection analysis this would be a false alarm to a target, because high confidence is assigned to an error. Conversely, a low confidence response that a target is old would be a miss in the Type-2 domain, because a low confidence rating is assigned to a correct response.

The second point to note follows the observations above. Type-2 miss responses ($M_2/M_{2T}/M_{2D}$) correspond to correct Type-1 responses, whereas Type-2 correct rejections ($CR_2/CR_{2T}/CR_{2D}$) correspond to incorrect Type-1 responses. This occurs because $M_2/M_{2T}/M_{2D}$ and $CR_2/CR_{2T}/CR_{2D}$ are referring to the accuracy of the Type-2 discrimination (i.e., whether participants can tell the difference between their own correct and incorrect responses), not to the accuracy of Type-1 responses. Hence, $M_2/M_{2T}/M_{2D}$ is a correct Type-1 response classified as inaccurate Type-2 discrimination because the observer misclassified it as incorrect. Similarly, $CR_2/CR_{2T}/CR_{2D}$ is an incorrect Type-1 response classified as accurate Type-2 discrimination because the observer correctly classified it as incorrect (see Figure 3).

It is not necessary that participants make an explicit, dichotomous “correct” versus “incorrect” decision in addition to an “old” versus “new” response to analyze performance data using Type-2 SDT. Just as with Type-1 SDT, it is possible to apply the same methodology to other forms of memory test, such as cued-recall (Higham, 2002), or in conditions in which participants have to decide whether to report a response or withhold it rather than explicitly rate their confidence (Higham, 2007, Experiment 2). It is also possible to plot a Type-2 ROC curve in the same manner as Type-1 SDT, using confidence data to manipulate the response criterion. For the experiments reported in this article, the “correct” versus “incorrect” decision is obtained from Type-2 confidence ratings regarding the accuracy of their response made on a 6-point scale. Different “correct/incorrect” criteria are assumed for each value on the scale, which generates several HRs and FARs for each participant. These values can then be plotted on a Type-2 ROC.

Table 1 shows how to obtain the ROC values for both Type-1 and Type-2 discrimination. For this table, the term “cc” has been included to denote the criterion associated with a particular level of confidence (e.g., confidence level 4 on a 6-point scale). Two types of confidence are relevant. The term “cc1” refers to confidence-in-oldness, a rating about a stimulus. The term “cc2” refers to confidence-in-accuracy, a rating about a response. To clarify the notation, consider first HR1 for Type-1 detection (see top left part of Table 1). This entry reads as “the number of ‘old’ responses assigned specifically to targets at or above a given level of confidence-in-oldness (Type-1 confidence criterion = cc1) divided by the number of targets.” If all levels of confidence are being considered (i.e., cc1 is maximally liberal), or no confidence rating is taken, then HR1 simply becomes the number of “old” responses assigned to targets, divided by the number of targets. If cc1 = 4 on a 6-point scale, then the HR1 is equal to the number of “old” responses assigned to targets with confidence-in-oldness of 4 or higher divided by the number of targets.

Type-2 confidence-based ROCs are calculated in the same way as traditional Type-1 ROCs (see Higham, 2007, for details). However, it is critical that the confidence rating pertain to the accuracy of the response and not to the oldness of the item (which is a Type-1 confidence rating). To highlight the difference, consider two 6-point scales, the first used for a Type-1 rating, ranging from 1 (certain new) to 6 (certain old), whereas the second is used for a Type-2 rating, ranging from 1 (guess) to 6 (certain correct). Consequently collected targets may be assigned high confidence regardless of whether the confidence
correspondingly, Type-1 ROC curves are plotted using confidence- and bias of judgments about correct/incorrect decisions. Consequently, they are positioned further to the right on the strength-of-evidence dimension. It is assumed that participants are sensitive to the higher strength of low frequency targets, and so adopt a higher hit rate, compared with high frequency words (e.g., Hirshman & Palij, 1992; Malmberg & Murnane, 2002). The word frequency mirror effect (e.g., Glanzer & Adams, 1985, 1990; Glanzer & Adams, 1990; Glanzer & Bowles, 1976) is a commonly observed phenomenon in recognition memory that occurs when some manipulation has one effect on the HR_T but an opposite effect on the FAR_T. For example, many studies have demonstrated that words with low lexical frequency exhibit a higher HR_T as well as a lower FAR_T, compared with high frequency words (e.g., Criss & Shiffrin, 2004; Hirshman & Palij, 1992; Malmberg & Murnane, 2002). The word frequency effect occurs regardless of whether word frequency is manipulated between lists, such that all items on a given study list are either high frequency or low frequency, or within lists, such that the study list consists of both high and low frequency items.

There are at least two SDT models that can account for the mirror effect (e.g., Hirshman, 1995; Stretch & Wixted, 1998). Model A is shown in the top panel of Figure 4. According to this model, which is often attributed to Brown, Lewis, and Monk (1977), low frequency targets (T_L) have higher strength-of-evidence than high frequency targets (T_H). Consequently, they are positioned further to the right on the strength-of-evidence dimension. It is assumed that participants are sensitive to the higher strength of low frequency targets, and so adopt a higher hit rate.
more stringent old/new criterion for low frequency items ($C_L$) than high frequency targets ($C_H$). However, the criterion is not moved as far up the strength-of-evidence dimension as the mean of the $T_L$ distribution. The result is that, relative to high frequency items, the $HR_1$ for low frequency targets is increased, and the $FAR_1$ is decreased: the mirror effect.¹

Model B, shown in the bottom panel of Figure 4, produces the mirror effect without a criterion shift. According to this model, the mirror effect occurs because there are two distractor distributions and two target distributions. As with Model A, it is assumed that $T_L$ is located to the right of $T_H$. In contrast, $D_L$ is located to the left of $D_H$. The positioning of the target distributions would occur if low frequency targets elicited more conscious recollection than high frequency targets, whereas the positioning of the distractor distributions

¹ Brown et al.'s (1977) original hypothesis has been subject to some criticism over the years. For example, some research (Greene & Thapar, 1994; Wixted, 1992) has shown that participants do not assume that low frequency items are more memorable than high frequency items. In fact, they sometimes hold the opposite assumption—that high frequency words are more memorable than low frequency words. However, more recently, Guttentag and Carroll (1998) and Benjamin (2003) have shown that participants come to the realization that low frequency words are more memorable than high frequency words, but only during the course of the recognition test.
in this manner, a mirror effect will occur even if the criterion remains stable throughout the test list (c).

Part of the difficulty of distinguishing between these two models derives from the fact that SDT indices are relative, not absolute, measures. For example, the discrimination index \( d' \) indicates the distance between the target and distractor distributions, and the response bias measure \( c \) indicates the distance between the criterion and the intersection point of the target and distractor distributions, both measured in standard deviation units. Neither indicates where, in absolute terms, either the distributions or criterion is located on the strength-of-evidence dimension. As a case in point, consider Model A shown in the top panel of Figure 4. Although the criterion has shifted between the low and high frequency conditions in this model, the measure \( c \) would equal zero in both the high and low frequency conditions (because it is located at the intersection point of the relevant distributions in both cases), suggesting that no shift had occurred. In short, Type-1 indices on their own are insufficient for discriminating between the models.

Type-2 SDT may have the potential to discriminate between the models, and it is this potential that we hope to demonstrate in the current research. Models A and B differ primarily in how each accounts for the FAR1 portion of the mirror effect, with a distractor distribution difference in Model B versus a criterion shift in Model A. Consequently, although these two accounts yield identical Type-1 indices, they have opposing effects on distractor monitoring (i.e., Type-2 discrimination of distractors). In particular, Model A predicts that low frequency items should be monitored worse than high frequency items, whereas Model B predicts that they should be monitored better.

To understand this, consider the distributions in Figure 5. The three panels on the left of Figure 5 show distractor distributions on a Type-1 dimension ranging from certain new (−4) to certain old (+4), each with an old/new criterion (c). In Figure 5, we have referred to the Type-1 and Type-2 dimensions as confidence-in-oldness and confidence-in-accuracy, respectively, rather than the more common “strength” or “strength-of-evidence.” Our terminology assists in distinguishing between the two dimension types, and we use them for the remainder of the article in places where making this distinction is critical. The top-left distractor distribution corresponded to high frequency items, with a mean slightly less than zero (−0.75) and an old/new criterion (c) at zero. The criterion splits the distractor distribution into two—FA1s and CR1s—with fewer of the former than the latter. In all the experiments reported here, we collected only confidence ratings about the accuracy of the response, whether that response is “old” or “new,” not confidence ratings about the oldness of the item. However, to generate predictions and test Type-1 models, we first need to resolve the problem of how Type-1 confidence-in-oldness is mapped onto Type-2 confidence-in-accuracy. Our solution was to compute absolute values of Type-1 confidence for use on the Type-2 dimension, with higher absolute values of confidence-in-oldness corresponding to higher confidence-in-accuracy. For example, an item located at −3 on the confidence-in-oldness dimension would be a correct rejection with low confidence-in-oldness. This item is translated into a correct decision assigned a high level of confidence-in-accuracy (+3) on the Type-2 dimension, which ranges from guessing to certain correct. The same is true of an item located at +3 on the confidence-in-oldness dimension. On the basis of these assumptions, both a Type-2 incorrect response (noise) distribution (derived from incorrect Type-1 decisions: false alarms) and a Type-2 correct response (signal-plus-noise) distribution (derived from correct, Type-1 decisions: correct rejections) is generated, which is shown on the top-right of Figure 5. By adopting several Type-2 criteria on the Type-2 confidence-in-accuracy dimension, it is possible to generate an ROC curve corresponding to the high frequency distractor distribution. This curve is labeled “Top panel in Figure 5” and appears as filled black circles in Figure 6. It serves as a baseline that can be used to determine the effect of a conservative criterion shift (Model A) or a distribution shift to the lower extreme (Model B).

Now consider the middle panel of Figure 5 that could correspond to low frequency items. For the Type-1 decision, it is assumed that a more stringent old/new criterion has been adopted (c set at +0.5), but the low frequency distribution has the same mean as the high frequency distribution above it (−0.75). The adoption of the stringent criterion for low frequency items lowers the FAR1 and corresponds to Model A (see Figure 4). However, when the correct and incorrect Type-1 decisions to low frequency items are translated into Type-2 space (shown on the right-middle panel of Figure 5), the distributions differ dramatically from the correct and incorrect high frequency distractor distributions shown in the top-right of Figure 5. In particular, the movement of the Type-1 criterion from 0 to 0.5 eliminates incorrect false alarms, and adds correct rejections in the region 0–0.5 on the confidence-in-accuracy dimension. Thus, a clear prediction that emerges from Model A is that monitoring is reduced for low frequency items relative to high frequency items because, for low frequency items, the mean of the incorrect item distribution increases, whereas the mean of the correct item distribution decreases. Furthermore, because the FAR2 equals 1.0 for all criteria less than 0.5 on the confidence-in-accuracy dimension, this would create an intercept of the right hand y-axis of the Type-2 ROC curve. This curve is labeled “Middle panel in Figure 5” and appears as crosses in Figure 6.

In contrast, consider the bottom panel in Figure 5, which represents the predictions that derive from Model B. In this case, recognition judgments might again be made to low frequency items. However, in contrast to the previous example, the mean of the low frequency distribution on the confidence-in-oldness dimension is less (−1.25) than the mean of the high frequency distribution (−0.75; top panel), whereas the criterion (c) is placed at 0 in both cases. The different distribution positions mean that the FAR1 is lower for low frequency distractors than high frequency distractors, and corresponds to Model B’s account of the FAR1 portion of the mirror effect (see Figure 4). However, when correct (CR1s) and incorrect (FA1s) item distributions are generated in Type-2 space (bottom-right of Figure 5), it is clear that the mean of the former increases, whereas the mean of the latter decreases. These changes would have the impact of increasing monitoring, as shown in the ROC with the label “Bottom panel in Figure 5,” which appears as unfilled squares in Figure 6. Also, in contrast to Model A, no intercept is predicted on the right-hand y-axis of the Type-2 ROC curve.

A set of related predictions can be generated for targets. Both Models A and B assume that the low frequency target distribution is shifted to the right relative to the high frequency target distribution (see Figure 4). This shift toward the upper end of the scale...
Figure 5. Type-1 distractor distributions (left) and the Type-2 correct (on the basis of Type-1 correct rejections) and incorrect (on the basis of Type-1 false alarms) response distributions (right) that are generated from them under the direct-translation hypothesis. With respect to Experiment 1, the top panel shows high frequency items, the middle panel shows low frequency items after a Type-1 criterion shift (Model A), and the bottom panel shows low frequency items with lower confidence-in-oldness than high frequency items (Model B). Although the predictions are discussed with respect to high and low frequency distractors for Experiment 1, similar predictions are made for other normal distractor distributions with similar placements and criteria. $c$ = the old/new criterion.
would reduce distractor monitoring. However, for targets, responses below the criterion are incorrect ($M_{1}$s), whereas items above it are correct ($H_{1}$s), which is opposite to distractors (i.e., correct $CR_{1}$s and incorrect $FA_{1}$s are below and above the criterion, respectively). Consequently, the shift to the upper extreme for low frequency targets has the opposite effect; that is, it increases, rather than reduces, monitoring compared with high frequency targets. The predicted increase is bolstered for Model A because there is the additional assumption of a conservative criterion shift. Whereas this shift decreases monitoring for distractors, it increases it for targets. Again, this reversed prediction occurs because correct and incorrect Type-1 responses for targets versus distractors are in opposite locations with respect to the criterion. Hence, the prediction for targets is the same for both Model A and Model B: better monitoring for low compared with high frequency targets.

In summary, Model A and Model B make two differential predictions regarding Type-2 ROC curves for distractors: Model A predicts that low frequency distractors should be monitored worse than high frequency distractors, whereas Model B predicts that they should be monitored better. A secondary prediction is that only Model A predicts that distractor monitoring should drop below chance, with the ROC intercepting the right-hand y-axis for low frequency items at liberal criteria (see Figure 6). Finally, both models predict that low frequency targets should be monitored better than high frequency targets. The only difference between the two models is that Model A, which incorporates the assumption of a conservative criterion shift, predicts that the difference between low and high frequency target monitoring should be more extreme.

Clearly the models and theoretical ROCs shown in Figures 5 and 6 are founded upon some assumptions, such as the Type-1 criterion for high frequency items being set at 0 rather than some other value, and the mean of the distractor distribution having a negative value. We believe both of these assumptions are plausible and provide a good starting point for making predictions about Type-2 ROCs. Critically, however, violations of these assumptions do not make any difference to the central prediction that a criterion shift toward conservativeness (Model A) on the confidence-in-oldness dimension reduces distractor monitoring, whereas a distribution shift toward the lower end of the confidence-in-oldness dimension (Model B) increases it. Another central assumption is that Type-2 confidence-in-accuracy can be determined by computing absolute values of confidence-in-oldness and that the psychological distance between any two scale values are equal within and between dimensions. In other words, confidence-in-accuracy can be directly translated from confidence-in-oldness, something we hereafter refer to as the direct-translation hypothesis. Clearly, this is not the only translation possible, so we test its validity more fully in Experiment 3 and discuss its implications in the General Discussion section.

Overview of the Experiments

To demonstrate the potential of Type-2 SDT for discriminating between possible models of recognition performance, we conducted three recognition experiments. In Experiment 1, we examined the word-frequency mirror effect, comparing recognition performance between words of high and low lexical frequency. In Experiment 2, we investigated the repetition-based strength effect, comparing recognition performance between words presented several times during study (strong) with words presented only once (weak; e.g., Benjamin, 2001; Dewhurst & Anderson, 1999; Hilford, Glanzer, & Kim, 1997; Morrell, Gaitan, & Wixted, 2002; Shiffrin, Huber, & Marinelli, 1995; Stretch & Wixted, 1998). In Experiment 3, we manipulated Type-1 response bias directly and observed the effect on target and distractor monitoring.

The old/new recognition data were analyzed in the traditional way, generating a $H_{1}$ and $FA_{1}$ for each participant, and determining Type-1 discrimination and response bias. This analysis allowed us to test whether a mirror effect was obtained. However, in addition, we also gathered Type-2 confidence ratings in each experiment; that is, regardless of whether an "old" or "new" response was made, participants were asked to rate their confi-
dence that the response was correct. Using these confidence data, we generated Type-2 ROC curves to test the predictions concern-
ing these curves derived from various models of the mirror effect.

Experiment 1

Method

Participants. Twenty-four undergraduate volunteers from the University of Plymouth (20) and the University of Southampton (4) participated for course credit. All participants gave their in-
formed consent to participate in advance of the experimental
session.

Materials. A list of 144 English words, selected from the Medical Research Council (MRC) psycholinguistic database was
used in the experiment. All words were nouns, either five or six
letters in length. Half of the words were high frequency, with a
Thornike–Lorge written frequency of greater than 500, whereas
the other half were low frequency words, with a Thornike–Lorge
written frequency between 5 and 17. In addition, four words were
added to act as primacy and recency buffers at study. Study words
were presented on a computer screen in black capital letters on a
white background as part of a PowerPoint slideshow presentation.
Test words were presented on four separate sheets of white paper
in black ink capital letters.

Design. The experimental design was 2 (prior presentation:
target/distractor) × 2 (frequency: high/low) and within subjects.
At study, 74 words were presented to participants via a slideshow
on a computer monitor. The first and last two words in the list were
treated as primacy and recency buffers and were not presented at
test. Half of the remaining 70 words were high frequency words,
and the other half were low frequency words. At test, 140 words
were presented, including all 70 target words. Additionally, 70
distractor words were presented, half of which were high fre-
quency words, and half were low frequency words. Four different
study lists and two different test lists were used to rotate the items
through the four experimental conditions (low frequency target,
high frequency target, low frequency distractor, high frequency
distractor) across participants, with fixed quasi-random presenta-
tion orders used at study and test.

Procedure. All participants were tested individually. Prior to
beginning the experiment, participants were informed that the
study was about recognition memory, and they were asked to read
and sign a consent form. They were told they would be requested
to list items and later would be tested for their memory
of those items. Subsequently, participants were provided with
study instructions in written form and requested to read the in-
structions. Participants were required to read words presented
individually on a computer monitor and to try to learn them for a
later memory test. After participants had finished reading the
instructions, they pressed the space bar to initiate a slideshow
presentation, with each word presented for 3 s separated by a 1-s
interstimulus interval. In total, the study phase lasted approxi-
mately 5 min.

After the study phase, participants were provided with a test
booklet. Test instructions were printed on the first page of the test
booklet requiring participants to read each test word and to make
a decision about whether it was presented at study. Participants
indicated their recognition decision by writing “O” (meaning
“old”) or “N” (meaning “new”) next to each of 140 test words,
which were printed on four sheets (following the instruction sheet)
of 38, 38, 38, and 26 items each. Participants were also required to
provide a judgment of how confident they were that
their recognition response was correct in a second column next to
the test items. The confidence judgment was made on a 6-point
scale ranging from 1 (not at all confident) to 6 (very confident).
It was made clear to participants that they had to provide a judgment
of how confident they were that the response they had provided
was correct and not of how confident they were that an item was
old. Moreover, it was made clear to participants that “new” re-
sponses could elicit a high confidence rating as well as “old”
responses. In particular, the following instructions were included:

Note that you can be just as confident about a “new” decision as you
might be about an “old” decision. For example, you are probably very
confident that your name did not appear on the study list (which
it didn’t!), because you would have remembered it if it did. So, if one
of the test words was your name, you would respond “new” to it on the
test, and rate your confidence as “6.” On the other hand, you probably
remember quite well that “university” and “Santa” actually did appear
near the beginning of the list. So, if they were shown on the test, you
would respond “old,” and rate your confidence as “6.”

The instruction sheet contained examples of correct and incor-
correct methods of filling out the required information for each item.
After completing the test, which was self-paced, participants were
debriefed and dismissed.

Results

Type-1 analysis. HR\(_1\) and FAR\(_1\) were calculated from the
number of positive responses (“O”) given to test items. Also, d’
(an SDT measure of discrimination) and c (an SDT measure of
response bias) were calculated from HR\(_1\) and FAR\(_1\) (Macmillan &
Creelman, 2005). To avoid undefined values, we applied
Snodgrass and Corwin’s (1988) correction factor to the frequen-
cies making up the HR\(_S\) and FAR\(_S\) prior to calculating d’ and c.
On the other hand, for statistical tests conducted directly on the
HR\(_S\) and FAR\(_S\), uncorrected scores were used. The mean uncor-
rected HR\(_S\) and FAR\(_S\), along with mean d’ and c, are reported in
Table 2 as function of word frequency level.

Table 2

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Index</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR(_1)</td>
<td>.794</td>
<td>.673</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.143)</td>
<td>(.165)</td>
<td></td>
</tr>
<tr>
<td>FAR(_1)</td>
<td>.195</td>
<td>.335</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.127)</td>
<td>(.186)</td>
<td></td>
</tr>
<tr>
<td>d’</td>
<td>1.821</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.861)</td>
<td>(0.595)</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>.032</td>
<td>.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.282)</td>
<td>(.394)</td>
<td></td>
</tr>
</tbody>
</table>

Note. HR\(_S\) = Type-1 hit rate; FAR\(_S\) = Type-1 false alarm rate; d’ =
Type-1 discrimination; c = Type-1 response bias. HR\(_1\) and FAR\(_1\) are
untransformed, whereas d’ and c are based on the transformed values
of HR\(_S\) and FAR\(_S\).

*a Indicates that the difference between high and low frequencies was
statistically significant.
Figure 8. Type-2 distractor receiver operating characteristics for Experiment 1: Open circles represent low frequency targets, and filled circles represent high frequency targets. HR = hit rate; FAR = false alarm rate; 2 = Type-2.

\[ \eta^2 \geq .232, \text{ except "6" and "2+," largest } F(1, 21) = 1.583, MSE = .036, p = .222, \eta^2 = .070. \] The main effect of confidence was not significant, \( F(4, 84) = 2.127, MSE = .051, p = .109, \eta^2 = .092. \)

A single mean index of monitoring was computed for low frequency targets, high frequency targets, low frequency distractors, and high frequency distractors by collapsing the arithmetic difference between HR\(_2\) and FAR\(_2\) across "6," "5+," "4+," "3+," and "2+." This mean value was then compared against chance performance (HR\(_2\) – FAR\(_2\) = 0) with four single-sample t-tests corresponding to each item type. These analyses showed that monitoring was above chance for low frequency targets, \( t(23) = 10.594, p < .001 (M = .342, SD = .158); \) high frequency targets, \( t(23) = 10.192, p < .001 (M = .295, SD = .142); \) and low frequency distractors, \( t(21) = 3.164, p = .005 (M = .108, SD = .160). \) However, monitoring for high frequency distractors did not differ from chance performance, \( t(23) = 0.385, p = .704 (M = .012, SD = .153). \)

**Discussion**

The current experiment produced a frequency-based mirror effect, replicating previous demonstrations of this effect (e.g., Glanzer & Adams, 1985, 1990; Glanzer et al., 1993; Glanzer & Bowles, 1976). That is, low frequency words were associated with both a higher HR\(_1\) and a lower FAR\(_1\), leading to better discrimination, compared with high frequency words. There was no difference between the word types with respect to Type-1 response bias. As we noted in the introduction to this study, this pattern of results is consistent with both Model A and Model B (see Figure 4).

However, it was possible to derive differential predictions from the two models in terms of Type-2 analyses. Model B predicted better monitoring for low frequency distractors than high frequency distractors, whereas Model A makes the opposite prediction. In addition, Model A predicted an intercept on the y-axis for low-frequency items at liberal criteria. The results from this study were therefore unambiguous in supporting Model B and rejecting Model A: In the empirical data, monitoring of low frequency distractors was significantly greater than for high frequency distractors at most levels of confidence, and there was no evidence of an intercept on the right-hand y-axis.

Model A (see Figure 4), in which participants are assumed to adopt more stringent criteria for more memorable items, is typically attributed to Brown et al. (1977). Do the current results, which specifically reject Model A, mean that memorability plays no role in the control of FA\(_s\)? We do not believe so. Although high memorability might cause criterion shifts under some circumstances, it might also contribute to the positioning of the high frequency and low frequency distractor distributions, with low frequency distractors being monitored better than high frequency distractors because they have less confidence-in-oldness associated with them. We consider this possible role of memorability further in the General Discussion section.

The ROC analysis revealed that metacognitive monitoring was higher for low frequency targets compared with high frequency targets, but only at high levels of confidence (5+ and 6). Better target monitoring was predicted by both Models A and B because both assume that the low frequency targets are, on average, located higher on the confidence-in-oldness dimension than high frequency targets, which would result in more correct high confidence judgments. The interaction between confidence level and frequency was not anticipated but might be attributable to conscious recollection, which has been proposed as a source of the HR\(_1\) portion of the mirror effect (e.g., Joordens & Hockley, 2000). Research in metacognition has shown that conscious recollection is monitored very well; generally speaking, people are metacognitively aware when they are accurately recollecting information versus when they are failing to do so. For example, Higham (2002; Higham & Tam, 2005, 2006) has shown that monitoring of cued recall with word pairs is excellent as long as there is no preexisting associative relationship between the pairs. As recall of such items is almost exclusively reliant on conscious recollection, these results are consistent with the idea that conscious recollect-

### Table 3

<table>
<thead>
<tr>
<th>Item and confidence level</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Targets, &quot;6&quot;</td>
<td>.551 (.210)*</td>
</tr>
<tr>
<td>Targets, &quot;5+&quot;</td>
<td>.547 (.281)*</td>
</tr>
<tr>
<td>Targets, &quot;4+&quot;</td>
<td>.358 (.330)</td>
</tr>
<tr>
<td>Targets, &quot;3+&quot;</td>
<td>.195 (.225)</td>
</tr>
<tr>
<td>Targets, &quot;2+&quot;</td>
<td>.058 (.116)</td>
</tr>
<tr>
<td>Distractors, &quot;6&quot;</td>
<td>.067 (.176)</td>
</tr>
<tr>
<td>Distractors, &quot;5+&quot;</td>
<td>.175 (.214)*</td>
</tr>
<tr>
<td>Distractors, &quot;4+&quot;</td>
<td>.181 (.299)*</td>
</tr>
<tr>
<td>Distractors, &quot;3+&quot;</td>
<td>.136 (.303)*</td>
</tr>
<tr>
<td>Distractors, &quot;2+&quot;</td>
<td>-.019 (097)</td>
</tr>
</tbody>
</table>

Note. HR\(_2\) = Type-2 hit rate; FAR\(_2\) = Type-2 false alarm rate.

* Indicates that the difference between high and low frequencies was statistically significant.
Single factor repeated-measures analyses of variance (ANOVs) were applied to the HR$_1$, FAR$_1$, $d'$, and $c$ scores. A significant main effect of frequency was observed for HR$_1$, $F(1, 23) = 10.576$, $MSE = .017$, $p = .004$, $\eta^2 = .314$, with a greater HR$_1$ for low frequency than high frequency words. The FAR$_1$ was also reliably lower for low frequency words than high frequency words, $F(1, 23) = 18.614$, $MSE = .013$, $p < .001$, $\eta^2 = .447$. Consistent with these results, a significant effect of frequency was observed on $d'$, $F(1, 23) = 37.775$, $MSE = .246$, $p < .001$, $\eta^2 = .622$, such that greater discrimination was seen for low frequency words than for high frequency words. In contrast to the preceding analyses, there was no effect of word frequency on the measure of response bias ($c$), $F < 1$.

The current experiment revealed a frequency-based mirror effect. However, the Type-1 results are completely ambiguous with regard to which model (A or B) better accounts for the data. Both models predict a higher HR$_1$ and $d'$, lower FAR$_1$, and no difference in $c$, for low frequency items compared with high frequency items. Hence, examination of the Type-1 SDT indices cannot distinguish between the models, so we now turn to the Type-2 analyses.

Type-2 analysis. Separate Type-2 ROCs, which give an indication of metacognitive monitoring, were created for targets and distractors, which are displayed in Figures 7 and 8, respectively. To create the ROC for targets, we determined the number of H$_1$s and M$_1$s, which respectively form the denominators for all the HR$_S$ (y-axis) and FAR$_S$ (x-axis) plotted on the ROC (see Table 1). Different numerators are generated for each point on the ROC, which correspond to the number of H$_1$s and M$_1$s at or above a given level of confidence. For example, the number of H$_1$s assigned confidence level “6” divided by all H$_1$s constitutes the most conservative HR$_2$ (bottom left). Similarly, the number of M$_1$s assigned confidence level “6” divided by all M$_1$s constitutes the most conservative FAR$_2$. For the next most conservative point, the number of H$_1$s and M$_1$s assigned confidence level “5” were added to the numerators of the HR$_2$ and FAR$_2$, respectively. This process continued until all points on the ROC were generated. The same process is used to create the ROC for distractors, except that the HR$_S$ and FAR$_S$ are generated from CR$_S$ (correct responses) and FA$_S$ (incorrect responses), respectively.

To obtain a single measure of monitoring for each level of the confidence scale, we subtracted FAR$_S$s from HR$_S$s for both targets and distractors. These data are shown in Table 3. A 2 (frequency: high, low) × 5 (confidence: “6,” “5+,” “4+,” “3+,” “2+”) repeated-measures ANOVA on the target data revealed a main effect of confidence, $F(4, 92) = 22.900$, $MSE = .111$, $p < .001$, $\eta^2 = .499$, which was qualified by an interaction between frequency and confidence, $F(4, 92) = 10.202$, $MSE = .021$, $p < .001$, $\eta^2 = .307$. Pairwise comparisons revealed that participants monitored their responses better with low frequency targets than with high frequency targets only at the higher levels of confidence (6 and 5+); “6,” $F(1, 23) = 20.905$, $MSE = .021$, $p < .001”; “5+,”$ $F(1, 23) = 10.905$, $MSE = .021$, $p = .003$. At all other levels of confidence, participants’ monitoring of the recognition response was not affected by word frequency, all Fs $(1, 23) < 3.860$, all ps > .06. The main effect of frequency was not significant, $F(1, 23) = 2.540$, $MSE = .052$, $p = .125$, $\eta^2 = .099$.

For the analogous ANOVA applied to distractors, 2 participants had to be excluded in the low frequency condition because they did not produce any FA$_S$s. The main effect of frequency was significant, $F(1, 21) = 6.052$, $MSE = .078$, $p = .023$, $\eta^2 = .224$, which was qualified by an interaction between frequency and confidence, $F(4, 84) = 3.654$, $MSE = .036$, $p = .020$, $\eta^2 = .148$. Pairwise comparisons revealed that participants monitored their responses to low frequency distractors better than high frequency distractors at all levels of confidence, all Fs $(1, 21) > 6.333$, all ps ≤ .004, all ps ≤ .004.

2 We would have preferred to use $A_c$ to index Type-2 discrimination. However, $A_c$ was not used because of the large number of cases for which not all of the confidence ratings were used by participants. When this happens, extreme values (1 or 0) occur frequently when calculating cumulative proportions, leading to undefined values of $A_c$. Had the Type-2 ROCs resembled the Type-1 equal-variance Gaussian ROCs shown in Figure 2, the simplest solution would have been to compute $d'$ on the basis of the HR and FA$_1$ collapsed across confidence levels to obtain a single discrimination measure per participant. However, just as with Type-1 ROCs, the measure of metacognitive monitoring (Type-2 discrimination) that should be used depends on the nature of the ROCs. The empirical Type-2 ROCs obtained in all our experiments differed substantially from the theoretical ones shown in Figure 2, suggesting that a simple analysis of $d'$ would not have been appropriate. More generally, very little is known about the nature of the underlying correct and incorrect item distributions on which Type-2 ROCs are based. Consequently, we chose to investigate monitoring using a measure that would reflect potential changes in discrimination across different levels of confidence. We do not advocate this type of analysis to investigate monitoring for all Type-2 ROC curves. Just as with Type-1 SDT, if the nature of the underlying distributions is understood and the ROCs conform to those distributions, alternate measures of discrimination (and bias) can be chosen accordingly.

3 For this and any subsequent analyses for which there was a violation of sphericity, more conservative Huynh-Feldt values are reported. Also, for several ANOVAs and pairwise comparisons, it was necessary to eliminate data from some participants because of missing data, the number of which are declared and can also be determined from the degrees of freedom reported with the relevant analyses. For such cases, we report the means from the analyses, not the means for all available participants.
tion leads to good monitoring. Similarly, Payne, Jacoby, and Lambert (2004) have found a moderate relationship between confidence and accuracy (a measure of metacognitive monitoring) in a two-alternative forced choice memory task. However, the confidence-accuracy relationship increased dramatically when it was calculated solely on an estimate of conscious recollection, which had been isolated with the process dissociation procedure (Jacoby, 1991). If conscious recollection is accounting for the difference in monitoring efficacy between low and high frequency targets, then it is likely that this difference would be reflected mostly in high confidence judgments because at a subjective level, conscious recollection is generally understood to be compelling (e.g., Dougal & Schooler, 2007), which would lead to accurate high confidence judgments. Because conscious recollection is anticipated to have an effect specifically on high confidence, it would result in an interaction between frequency and confidence level in the Type-2 ROC analysis, just as we observed in the current experiment.

It is also worth noting that monitoring of high frequency distractors was no better than chance. In other words, confidence ratings did not distinguish between accurate (“new”) versus inaccurate (“old”) judgments to these items. Although the difference between low frequency and high frequency distractors is interpretable in terms of differential memorability, participants’ complete inability to discriminate between correct and incorrect responses to high frequency distractors was unexpected. More seriously, the empirical ROC for high frequency distractors does not resemble the theoretical “High frequency (Expt 1)” ROC shown in Figure 6, which shows above-chance monitoring. Chance-level monitoring could easily be modeled by raising the mean of the high frequency distractor distribution from \(-0.750\), shown in the top panel of Figure 5, to 0, while keeping the criterion positioned at 0. This increase would mean that \(FA_1\) (incorrect decisions above the mean) would be assigned confidence equal to that assigned to \(CR_1\) (correct decisions below the mean), which would yield chance level monitoring. The problem with this amendment to the model is that it predicts that the \(FA_1\) would equal 0.500, whereas it was only 0.335 in the current experiment. We reserve further discussion of this point until after Experiment 2, where we once again report data on participants’ ability to monitor distractors.

**Experiment 2**

In Experiment 1, we manipulated lexical frequency to produce a mirror effect. However, another method used to produce a mirror effect is to manipulate the number of presentations (i.e., strength) of the targets during study (e.g., Benjamin, 2001; Dewhurst & Anderson, 1999; Hilford et al., 1997; Morrell et al., 2002; Shiffrin et al., 1995; Stretch & Wixted, 1998). If strength is manipulated between test lists, a strength-based mirror effect is nearly always obtained (e.g., Benjamin, 2001), such that the \(HR_1\) is higher, but the \(FA_1\) is lower, for strong items than for weak items. Conversely, if strength is manipulated within the same test list, the \(HR_1\) portion of the mirror effect is usually obtained, but logically there is only one \(FA_1\) (because distractors, by definition, are not presented during study and hence not subject to the strength manipulation). Hence, in many within-list experiments on the strength effect, it is not possible to compare the \(FA_1\) between weak and strong distractors to establish whether a mirror effect has occurred.

However, to get around this problem, Stretch and Wixted (1998) investigated both the strength effect and the word frequency effect within the same experiment and used colors to cue the strength categories. For example, in their Experiment 3, high and low frequency words were presented within the same study phase, but high frequency words were selectively strengthened by presenting them five times each throughout the study list, whereas low frequency words were presented only once. At test, (strong) high frequency items were presented in one color (e.g., red), whereas (weak) low frequency items were presented in another color (e.g., green). Therefore, participants could have used either the lexical frequency of the items and/or their color to determine whether they belonged to the strong or weak category, and this was true of both targets and distractors. Stretch and Wixted reasoned that if participants varied their old/new decision criterion on an item-by-item basis throughout the test list, adopting a more stringent criterion for strong (red, high frequency) items and a more lenient criterion for weak (green, low frequency) items, then both the \(HR_1\) and \(FA_1\) portions of the word frequency effect should be affected. Indeed, the strength manipulation caused the \(HR_1\) for strong high frequency items to be greater than the \(HR_1\) for weak low frequency items. However, the \(FA_1\) for strong high frequency items remained greater than that for weak low frequency items. In other words, selectively strengthening high frequency items reversed the \(HR_1\) portion of the word frequency effect, but the \(FA_1\) portion of the word frequency effect remained intact. On the basis of these results, and the results of several other experiments, they concluded that item-by-item criterion shifts do not occur. This result is consistent with other research on the strength effect suggesting that \(FA_1\)s are not affected when strength is manipulated within lists (e.g., Morrell et al., 2002). Indeed, the consistency with which strength-based mirror effects occur between lists, but not within lists, has led some to conclude that the strength effect has completely different bases in between-lists versus within-list paradigms (e.g., Morrell et al., 2002; Stretch & Wixted, 1998).

That is, the between-lists strength effect is based on a Type-1 criterion shift (producing a full mirror effect; see Model A in Figure 3), whereas no criterion shift occurs when strength is manipulated within lists (producing the \(HR_1\) portion of the mirror effect, but not the \(FA_1\) portion).

For Experiment 2, we investigated the within-list strength effect, again using Type-2 SDT, and used living and nonliving words as strength category cues. Thus, for a given participant, all words designating living things (e.g., “chicken”) were presented at one level of strength in the study list, whereas words designating nonliving things (e.g., “iron”) were presented at the other strength level (counterbalanced across participants). Because this is not a particularly salient cue, we also used a unique labeling methodol-

---

4 It should be noted that Stretch and Wixted (1998) also used the terms “type-1” and “type-2.” However, these terms were not used to distinguish between different detection tasks as they are used here. Instead, “type-1” was used to identify a mirror effect caused by a shift in the response criterion (our Model A), whereas “type-2” was used to identify a mirror effect observed in spite of the fact that the response criterion remains fixed (our Model B).
ogy during the recognition memory test to highlight the difference between weak and strong items. Two slightly different labeling methodologies were used. For one group of participants \((n = 24)\), instead of simply requiring old/new judgments of each test item, two columns were presented next to each test word, one with the heading “saw 1 time (O) or new (N)?” and the other with the heading “saw 3 times (O) or new (N)?” (Tam, 2006). Space for an old/new recognition rating was provided next to each test item under only one of the column headings so that it was clear which strength category each item belonged to regardless of whether it was a target or a distractor. For another group of participants \((n = 26)\), two response alternatives were presented next to each test item. For targets and distractors assigned to the strong condition, the two response alternatives were “3” and “0” (i.e., presented either three times or not at all during study). For targets and distractors in the weak condition, the two response alternatives were “1” and “0” (i.e., presented either once or not at all during study). Participants were instructed to circle just one response alternative next to each test item. Such labeling, coupled with the living/nonliving cue, might be enough for participants to adopt different Type-1 response criteria for strong versus weak test items, producing a difference in the FARs and a mirror effect. On the other hand, if these manipulations did not make the strong versus weak distinction salient enough, then comparable FARs might be obtained, similar to Stretch and Wixted’s (1998) results.

More important, whatever Type-1 results we obtained in Experiment 2 should be reflected in the Type-2 analysis. In particular, if a mirror effect is obtained, then both Model A and Model B shown in Figure 4 are feasible candidates (only with high frequency and low frequency designations replaced with weak and strong, respectively). Following the same reasoning as in Experiment 1, compared with the weak condition, Model A predicts better monitoring of targets, but worse monitoring of distractors, in the strong condition (see Figures 5 and 6). On the other hand, Model B predicts better monitoring of both targets and distractors in the strong condition relative to the weak condition.

Conversely, if strength affects the HRs, but no difference in FARs is observed (i.e., a mirror effect is not obtained), Models A and B are no longer viable candidates because they predict mirror effects. Instead, alternative models must be considered, two of which are shown in Figure 9. In Model C, there is neither a criterion shift nor separate distractor distributions, so the FARs do not differ between the strength conditions. However, because there are separate target distributions reflecting the manipulation of strength, different HRs will be obtained. As far as Type-2 analyses are concerned, Model C predicts better monitoring of strong targets compared with weak targets because of the different positioning of the target distributions on the strength-of-evidence scale. However, no difference in the monitoring of strong versus weak distractors is predicted because there is only one distribution and one criterion.

Model D in Figure 9 again shows that strength has affected the positioning of the strong and weak target distributions, but it has also affected the positioning of the distractor distributions. In particular, the strong distractor distribution has more strength-of-evidence than the weak distractor distribution. This difference might occur if there is greater similarity amongst items within rather than between the living/nonliving categories used to designate the strong and weak conditions. If so, then strengthening the targets within one of the categories might simultaneously strengthen distractors from the same category because targets and distractors share features. In response to the higher strength-of-evidence for both strong targets and strong distractors, participants adopt a more stringent criterion for those items, moving it the same distance up the strength-of-evidence dimension as the gain in strength-of-evidence for strong distractors. The result will be similar FARs, but different HRs, between the strong and weak conditions. Because the criterion has shifted to be more conservative for strong items, the effect on distractor monitoring will be similar to that observed from a criterion shift in Model A. That is, strong distractor monitoring is predicted to be much worse than weak distractor monitoring under Model D. This effect is bolstered by the gain in strength-of-evidence for the strong compared with the weak distractor distribution. On the other hand, better target monitoring is predicted for strong targets compared with weak targets under Model D. In summary, regardless of whether a mirror effect is obtained in Experiment 2, Type-2 SDT analyses can be used to disambiguate the results. If a mirror effect is obtained, then Type-1 indices are inadequate for distinguishing between Models A and B; both predict similar Type-1 data. However, Model A predicts that distractor monitoring will be worse in the strong compared with the weak condition, whereas Model B makes the opposite prediction. If a mirror effect is not obtained, and a single FAR is observed but there is a difference in HRs between the strong and weak conditions, then either Model C or Model D is viable. Again, however, Type-1 data cannot differentiate between these models. On the other hand, Type-2 SDT predicts that there will be no difference in strong versus weak distractor monitoring under Model C, but that the former will be worse than the latter under Model D.

Method

Participants. Participants were 50 undergraduates from the University of Plymouth who participated for partial course credit. Twenty-four participants were assigned to the O–N group, and 26 were assigned to the 0–1–3 group. All participants gave their informed consent prior to commencing the study.

Materials. A list of 144 English words was used for the experiment, again taken from the MRC psycholinguistic database. All were common English nouns, between three and eight letters in length, with a Thorndike–Lorge frequency between 50 and 1,000. Half of the words represented living items (e.g., parsley, goat), whereas the other half represented nonliving items (e.g., spoon, clock). Study words were presented on a computer screen in black capital letters on a white background as part of a PowerPoint slideshow presentation. Test words were presented on four separate sheets of white paper in black ink capital letters.

Design. The experiment was a 2 (test type: 3–1–0/O–N) × 2 (prior presentation: target/distractor) × 2 (strength: strong/weak) mixed design, with test type manipulated between subjects. At study, 72 words were presented. Half of these words (36) were presented only one time (weak condition), whereas the other half (36) were presented three times (strong condition) for a total of 144 study trials. For half the participants, all living words were presented thrice, whereas all nonliving words were presented once. This strength/item type association was reversed for the other half.
of participants. At test, 144 words were presented, printed on four test sheets of 36 items each. Of these words, 72 were the previously seen targets, and 72 were distractors. Half of the 72 distractors were living, and half were nonliving. Because of the strength/item type association of the targets, distractors of a given item type were thus assigned to either the weak or strong condition, depending on the counterbalance condition. Four different study lists and two different test lists were used to rotate the items through the four experimental conditions (target-weak, target-strong, distractor-weak, and distractor-strong). Items were presented in a fixed quasi-random order at both study and test.

Procedure. The procedure of Experiment 2 was identical to that of Experiment 1, aside from the following differences. At study, participants were informed that words would represent something either living or nonliving, and they were instructed to use this difference to help memorize the words. There was no interstimulus interval between words presented at study. At test, the instructions either correctly informed participants that living items were presented three times and that nonliving items were presented once, or vice versa, depending on the counterbalance version. Two different types of test were used across participants: Both types were designed to remind participants of the number of times they had potentially seen a word at study and to make the strength distinction for distractors more explicit on an item-by-item basis. In both cases, each of the four test sheets containing test words (which followed the instruction sheet) consisted of four columns. In the first, far-left column, the test words were listed in a fixed random order that differed from the study order.

For 24 participants in the O–N group, next to the list of test words were two additional columns, the first with the heading “saw 1 time (O) or New (N)?” and the second with the heading “saw 3 times (O) or New (N)?” If a test word had been presented

---

**Figure 9.** Two hypothetical signal detection theory models to account for the within-list strength effect in which a difference in the hit rate, but no difference in the false alarm rate, is observed. Model C assumes that there is a single distractor distribution and a single criterion, whereas Model D assumes that there are different positioning of the distractor distributions and a concomitant criterion shift. D = distractor; T = target; S = strong; W = weak; c = the old/new criterion.
once at study or was a distractor assigned to the weak condition, a line appeared only under the first column with no space available under the second column to make an O/N rating. Space was available for an O/N rating under the second column, but not the first, if the test item had been presented thrice at study or was a distractor assigned to the strong category. The fourth column (with space next to every test word) was used for a rating of confidence about the accuracy of the response ranging from 1 (not at all confident) to 6 (very confident).

For 26 participants in the 3–1–0 group, next to the list of test words were two additional columns. The first column was used to indicate that the item had been seen during study and the second to indicate that it had not. For strong targets and distractors, the choices available were “3” (saw three times) and “0” (new item), whereas for weak targets and distractors, the available choices were “1” (saw once in study) and “0” (new item). Participants were instructed to circle one of the available alternatives for each test item. Also for the 3–1–0 group, the fourth column was used for a rating of confidence (1–6) about the accuracy of the response.

**Results**

Preliminary analyses indicated that the manipulation of test type did not exert a significant main effect, nor did it interact with any of the other independent variables, largest $F(1, 49) = 2.698$, $MSE = .045$, $p = .107$, $\eta^2 = .053$. Consequently, all subsequent analyses were conducted after collapsing over this factor.

**Type-1 analysis.** As in Experiment 1, HR$_{1}$, FAR$_{1}$, $d’$, and $c$ were calculated for each participant, and each measure was subjected to repeated-measures ANOVA. To avoid undefined values, we applied Snodgrass and Corwin’s (1988) correction factor to the frequencies used to create the HR$_{s}$ and FAR$_{s}$ prior to calculating $d’$ and $c$. On the other hand, for statistical tests conducted directly on the HR$_{s}$ and FAR$_{s}$, uncorrected scores were used. Mean values of all SDT indices are shown in Table 4 as a function of strength.

There was a main effect of strength on uncorrected HR$_{1}$, $F(1, 49) = 80.594$, $MSE = .008$, $p < .001$, $\eta^2 = .622$, with a greater HR$_{1}$ for strong targets compared with weak targets. However, there was no significant strength effect on false alarms, $F < 1$. Overall, discrimination ($d’$) was greater for strong items than for weak items, $F(1, 49) = 82.925$, $MSE = .180$, $p < .001$, $\eta^2 = .629$, and the measure of response bias ($c$) was lower in the strong condition than in the weak condition, $F(1, 49) = 28.863$, $MSE = .093$, $p = .001$, $\eta^2 = .371$.

**Type-2 analysis.** Similar to Experiment 1, to assess how strength affected participants’ monitoring, ROCs for targets and distractors were generated, and a 2 (strength: strong and weak) $\times$ 5 (confidence: “6,” “5+,” “4+,” “3+,” and “2+”) repeated-measures ANOVA was conducted on the dependent variable $HR_{2} – FAR_{2}$, ROCs for targets and distractors are reported in Figures 10 and 11, respectively, and the mean values of $HR_{2} – FAR_{2}$ at different levels of confidence by type of item (targets vs. distractors) and strength level (strong vs. weak) are reported in Table 5. Fifteen participants had to be excluded from the analysis in the strong-target condition, as they produced no $M_{s}$. Similarly, 3 participants had to be excluded from both the strong and weak distractor conditions because they produced no $NA_{s}$.

When the ANOVA was carried out on targets, there was a main effect of strength, $F(1, 34) = 6.103$, $MSE = .064$, $p = .019$, $\eta^2 = .152$, a main effect of confidence, $F(4, 136) = 40.614$, $MSE = .106$, $p < .001$, $\eta^2 = .544$, both of which were qualified by a significant interaction: $F(4, 136) = 2.846$, $MSE = .019$, $p = .026$, $\eta^2 = .077$. Pairwise comparisons showed better monitoring for strong targets than for weak targets at high levels of confidence (6 and 5+); “6.” $F(1, 34) = 16.158$, $MSE = .019$, $p < .001$, $\eta^2 = .322; “5+,” F(1, 34) = 13.579$, $MSE = .019$, $p < .001$, $\eta^2 = .285$. At lower levels of confidence, there was no effect of strength on target monitoring, all $Fs(1, 34) \leq 1.421$, all $ps \geq .241$, all $\eta^2$s $\leq .040$. When the analogous ANOVA was applied to distractors, there was no main effect of strength, no main effect of confidence, nor a significant interaction between strength and confidence, largest $F(4, 148) = 1.902$, $MSE = .069$, $p = .141$, $\eta^2 = .049$. These results suggest that strength did not have any effect on participants’ ability to monitor distractors.

As in Experiment 1, four one-sample $t$-tests compared mean values of $HR_{2} – FAR_{2}$ (collapsed across confidence levels) against chance (i.e., $HR_{2} – FAR_{2} = 0$) for strong targets, weak targets, strong distractors, and weak distractors. These analyses showed that participants performed well above chance for both strong and weak targets, $t(34) = 8.282$, $p < .001$ ($M = .329$, $SD = .235$), and $t(49) = 11.165$, $p < .001$ ($M = .284$, $SD = .180$), respectively. In contrast, monitoring was at chance for both strong and weak distractors, $t(42) = 0.156$, $p = .877$ ($M = .005$, $SD = .198$), and $t(43) = 1.529$, $p = .134$ ($M = .043$, $SD = .186$), respectively.

**Discussion**

In the current experiment, we investigated the within-list strength effect and, consistent with much other research on the topic (e.g., Bruno, Higham, & Perfect, 2008, Experiment 1; Hirshman, 1995; Morrell et al., 2002; Stretch & Wixted, 1998), we found that strength had an effect on the $HR_{1}$ but not effect on the $FAR_{1}$. Hence, Models A and B (see Figure 4) were eliminated as possible candidates because they predict mirror effects. On the other hand, both Models C and D (see Figure 9) predict the pattern of results that was obtained.
was associated with a higher HR in which distractor monitoring was better for the class of items that result stands in contrast to the effect of frequency in Experiment 1, conditions, and strength did not interact with confidence. This null was no difference in monitoring between the strong and weak sentences. In contrast, Type-2 ROC analysis indicated that there was should exert an effect predominantly on high confidence judg-

tments. In Experiment 1, that is, participants monitored strong targets better than weak targets, but only at higher levels of confidence. Again, we attribute this effect to an increase in the conscious recollection of targets, which, following the same reasoning as in Experiment 1, should exert an effect predominantly on high confidence judgments. In contrast, Type-2 ROC analysis indicated that there was no effect of the strength manipulation on distractors; that is, there was no difference in monitoring between the strong and weak conditions, and strength did not interact with confidence. This null result stands in contrast to the effect of frequency in Experiment 1, in which distractor monitoring was better for the class of items that was associated with a higher HR, (low frequency items). More important, similar levels of distractor monitoring between the strength conditions was predicted by Model C, whereas Model D predicted that strong distractors will be monitored worse than weak distractors. Thus, Type-2 SDT was able to discriminate between models of recognition performance that Type-1 SDT could not.

As with high frequency items in Experiment 1, the monitoring of both strong and weak distractors was very poor and not significantly above chance. As mentioned above, if it is assumed that the distractor distribution has a mean of zero and the Type-1 criterion is also positioned at zero, monitoring would be equal to chance (see Figure 5). However, this scenario also predicts that the FAR would be equal to 0.500, which was not the case with either distractor type (see Table 4). This issue deserves more attention and is addressed more fully in the General Discussion section.

**Experiment 3**

An alternative interpretation of the collective results of Experiments 1 and 2 is that differences in distractor monitoring only occur if there are differences in the Type-1 FAR, with good distractor monitoring for low FAR, and poor distractor monitoring for high FAR. The data are perfectly consistent with this interpretation. In Experiment 1, in which there were differences in FAR, as a function of frequency, there was also a corresponding variation in distractor monitoring. In Experiment 2, in which there was no difference in the FAR as a function of strength, there was no variation in distractor monitoring.

Sensitivity of distractor monitoring to the mere magnitude of the FAR might occur if the direct-translation hypothesis is invalid. This hypothesis assumes that it is possible to translate directly the confidence-in-oldness scale to the confidence-in-accuracy scale by computing absolute values, and that the zero point of both dimensions is fixed. However, another possible translation would be to set the zero (guess) point of Type-2 confidence at the Type-1 criterion, wherever that may be, rather than at the zero point of confidence-in-oldness (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000). Under this assumption, the Type-2 scale would shift according to the placement of the criterion on the confidence-in-oldness dimension, an assumption we hereafter refer to as the shifting-scale hypothesis. This is not an unreasonable hypothesis.

![Figure 10. Type-2 target receiver operating characteristics for Experiment 2: Filled circles represent strong targets, and open circles represent weak targets. HR = hit rate; FAR = false alarm rate; 2 = Type-2.](image1)

![Figure 11. Type-2 distractor receiver operating characteristics for Experiment 2: Filled circles represent strong distractors, and open circles represent weak distractors. HR = hit rate; FAR = false alarm rate; 2 = Type-2.](image2)
the Type-1 criterion marks the crossover point of the old/new decision, so it is not unrealistic to assume that this is also the point of maximum uncertainty (guessing) about accuracy.

It is critical to determine the validity of the shifting-scale hypothesis because the usefulness of Type-2 SDT rests upon it. Consider again the effect on distractors as the criterion is made more conservative. This effect is depicted in the top and middle panels of Figure 5, which correspond to Model A. Under the direct-translation hypothesis, as the criterion shifts toward conservativeness from the top-left to the middle-left panels of Figure 5, the Type-2 model shown in the middle-right panel is created, as described in detail above with respect to a manipulation of word frequency. However, if the zero-point on the confidence-in-accuracy scale shifts with the Type-1 criterion, then the same criterion shift will produce the Type-2 model shown in the bottom-right panel of Figure 5, which corresponds to a distractor distribution shift and model B. In other words, if the shifting-scale hypothesis is valid, then both Model A and Model B would produce the same Type-2 model (bottom-right of Figure 5), meaning that they would be indistinguishable on the basis of Type-2 data.

To ascertain the validity of the shifting-scale hypothesis, we conducted a third experiment in which the FAR, was manipulated with a variable known to affect the response criterion but have no effect on discrimination. In particular, Donaldson’s (1996) procedure that varied the number of “old” versus “new” response categories was adopted. In the NNNO group, participants were presented with three response categories to respond “old,” whereas the ratio of “new” to “old” was 1:1. Finally, in the NNOO condition, there were two categories each for responding “old” and “new.” Different levels of the same response category within each group were used to designate varying levels of Type-1 confidence-in-oldness. Type-2 confidence-in-accuracy ratings were also required after the recognition judgment. On the basis of Donaldson’s results, we expected conservative Type-1 responding in the NNNO group, liberal Type-1 responding in the NOOO group, and intermediate Type-1 responding in the NNOO group. However, we expected no between-groups differences in Type-1 discrimination. In other words, this manipulation should vary the FAR, by altering the placement of the Type-1 criterion on the confidence-in-oldness dimension, but it should have no effect on d’ or the placement of the target and distractor distributions on the same dimension.

By varying the criterion in this manner, it was possible to pit the direct-translation and shifting-scale hypotheses against each other. If distractor monitoring is inversely related to the FAR, as the shifting-scale hypothesis predicts, then decreasing the FAR, from the liberal NOOO group through to the conservative NNNO group should increase distractor monitoring. Conversely, if monitoring and the FAR, are positively correlated, as the direct-translation hypothesis predicts, then the opposite result should obtain.

**Method**

**Participants.** Participants were 55 A-level students from Totton College, Hampshire, United Kingdom. They were split into seven groups of participants between 3 and 14 people in size. Their age was between 16 and 18 years: For students under the age of 18 years, participation in the experiment was authorized by their parents or their adult legal representatives, who signed a consent form to allow the treatment of their data. Students who were 18 years of age or above gave consent for the treatment of their own data personally. Two participants were eliminated because of incomplete tests, leaving 53 participants in total.

**Materials.** A list of 144 English words taken from the MRC psycholinguistic database was used for the experiment. All words were common English nouns, between three and nine letters in length, with a Thorndike–Lorge frequency between 50 and 1,000. In addition to this, four buffer items were used, two at the beginning and two at the end of the study list. Study items were presented on a white board with a projector connected to a Macintosh computer. Test items were presented on four separate sheets of paper.

**Design.** The experiment was a 2 (prior presentation: target/distractor) × 3 (test type: NOOO, NNNO, NNNO) mixed design with test type manipulated between subjects. At study, 72 critical words and four buffer words were presented. Three study lists were generated, each with the same buffer words, but a different random selection (and order) of 72 critical words from the pool of 144 words. For the NNNO group (n = 19), 6, 8, and 5 participants received study lists 1, 2, and 3, respectively. The analogues values for the NNOO (n = 17) and NOOO (n = 17) groups were 6, 6, and 5 for Study Lists 1, 2, and 3, respectively. At test, 144 words were presented on four separate sheets of paper, 3 with 41 items and a 4th with 21 items. A different random order of test items was used for each test type group.

**Procedure.** Prior to beginning the experiment, participants were informed that the study was aimed at investigating recognition memory, and that they would be required to learn a list of words and then later be tested for their memory of those words. Subsequently, participants were informed the list of words would be

---

*Note.* HR<sub>2</sub> = Type-2 hit rate; FAR<sub>2</sub> = Type-2 false alarm rate.

---

*Table 5*

<table>
<thead>
<tr>
<th>Item and confidence level</th>
<th>Strength</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targets, “6”</td>
<td>.507 (.336)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.374 (.216)&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Targets, “5+”</td>
<td>.520 (.312)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.398 (.227)&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Targets, “4+”</td>
<td>.381 (.287)</td>
<td>.366 (.253)</td>
</tr>
<tr>
<td>Targets, “3+”</td>
<td>.186 (.334)</td>
<td>.159 (.286)</td>
</tr>
<tr>
<td>Targets, “2+”</td>
<td>.052 (.260)</td>
<td>.013 (.202)</td>
</tr>
<tr>
<td>Distractors, “6”</td>
<td>.049 (.241)</td>
<td>.040 (.144)</td>
</tr>
<tr>
<td>Distractors, “5+”</td>
<td>.034 (.282)</td>
<td>.098 (.268)</td>
</tr>
<tr>
<td>Distractors, “4+”</td>
<td>-.004 (.334)</td>
<td>-.032 (.320)</td>
</tr>
<tr>
<td>Distractors, “3+”</td>
<td>-.022 (.232)</td>
<td>-.005 (.294)</td>
</tr>
<tr>
<td>Distractors, “2+”</td>
<td>.002 (.192)</td>
<td>.010 (.221)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Indicates that the difference between strong and weak conditions was statistically significant.

---

<sup>5</sup> Although the earlier description relates specifically to criterion shifts resulting from a manipulation of word frequency, the predictions of Model A with respect to distractor monitoring can be generalized to any of a variety of manipulations that cause a conservative criterion shift.
projected onto the white board and that they should attend to it. Study words were presented for 0.5 s each, with an interstimulus interval of 1.5 s. The total amount of time participants were engaged in the study phase was 5 min.

After completion of the study phase, participants were required to make old/new recognition judgments to test words (printed on the left side of four test sheets) by choosing one of four response options (appearing in four columns to the right). In the NOOO test-type group, there was one response category for responding “new” and three categories for responding “old” (with varying levels of confidence). These categories were: “New,” “Old1” (“not so sure” old), “Old2” (“quite sure” old), and “Old3” (“very sure” old). In the NNOO test-type group, there were two categories each for both “new” and “old” judgments: “New1” (“not so sure” new), “New2” (“very sure” new), “Old1” (“not so sure” old), and “Old2” (“very sure” old). Finally, in the NNNO test-type group, there were three categories for “new” judgments but only one for “old” ones: “New1” (“not so sure” new), “New2” (“quite sure” new), “New3” (“very sure” new), and “Old.”

After the recognition response was made, participants also were asked to provide a judgment of how confident they were that their recognition response was correct. The confidence judgment was made for each test response (in a column to the far right of the page) with a 6-point scale, in which 1 = not at all confident, 3 or 4 = somewhat confident, and 6 = very confident. It was made clear to participants that they had to provide a judgment of how confident they were that the response they had provided was correct and not of how confident they were that an item was old or new. Moreover, it was also clarified that “new” responses could elicit a high confidence rating and not just “old” responses. At the end of the test phase, participants were debriefed.

**Results**

**Type-1 analysis.** HR₁s and FAR₁s were calculated from the number of positive (“old”) responses given to test items. Also, d’ and c were calculated from the corrected HR₁s and FAR₁s as in Experiments 1 and 2. Means of all four indices can be found in Table 6. A one-way between-subjects ANOVA was conducted on HR₁, and a main effect of test type was observed, F(2, 50) = 25.419, MSE = .017, p < .001, η² = .504, showing that the greatest HR₁ value was in the NOOO test type condition, and the lowest was in the NNNO test type condition. Post hoc analyses (Tukey’s honestly significant difference [HSD]) showed that the difference in HR₁ between the NNNO and the NOOO conditions was significant, mean absolute difference = .154, p = .002, and so was the difference between the NOOO and the NNNO conditions, mean absolute difference = .153, p = .003.

An analogous ANOVA also was conducted on FAR₁, and, again, a main effect of test type was observed, F(2, 50) = 33.725, MSE = .017, p < .001, η² = .574, showing that the greatest FAR₁ value was in the NOOO test type condition and the lowest was in the NNNO test type condition. Post hoc analyses (Tukey’s HSD) showed that, as with the analysis of HR₁, the difference in FAR₁ between the NNNO and the NOOO conditions was significant, mean absolute difference = .020, p < .001, and so was the difference between the NOOO and the NNNO conditions, mean absolute difference = .155, p = .003.

<table>
<thead>
<tr>
<th>Test type group</th>
<th>Index</th>
<th>NNNO</th>
<th>NNOO</th>
<th>NOOO</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR₁</td>
<td>.492</td>
<td>.645</td>
<td>.798</td>
<td></td>
</tr>
<tr>
<td>FAR₁</td>
<td>.125</td>
<td>.327</td>
<td>.482</td>
<td></td>
</tr>
<tr>
<td>d'</td>
<td>1.250</td>
<td>0.890</td>
<td>0.946</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>.647</td>
<td>.050</td>
<td>-.414</td>
<td></td>
</tr>
</tbody>
</table>

Note. HR₁ = Type-1 hit rate; FAR₁ = Type-1 false alarm rate; d' = Type-1 discrimination; c = Type-1 response bias. HR₁ and FAR₁ are untransformed, whereas d' and c are based on from the transformed values of HR₁ and FAR₁. NNNO group = participants were presented with three response categories to respond “new” but only one for responding “old”; NNOO group = participants were presented with two categories each for responding “old” and “new”; NOOO group = participants were presented with three response categories to respond “old” but only one for responding “new.”

Finally, analogous ANOVAs were conducted on both the SDT measures d’ and c to examine whether test type had an effect on discrimination and/or response criterion placement, respectively. No main effect of test type was observed on d’, F(2, 50) = 2.074, MSE = .329, p = .136, η² = .077. In contrast, a main effect of test type was detected with c, F(2, 50) = 53.325, MSE = .096, p < .001, η² = .681. Post hoc analyses (Tukey’s HSD) revealed that the difference in c between the NNNO and the NOOO conditions was significant, mean absolute difference = .597, p < .001, and so was the difference between the NOOO and the NNNO conditions, mean absolute difference = .464, p < .001.

**Type-2 analysis.** As in the previous experiments, ROCs for targets and distractors were generated, which are shown in Figures 12 and 13, respectively. Two 3 (test type: NOOO, NNOO, NNNO) × 5 (confidence: “6,” “5+,” “4+,” “3+,” and “2+”) mixed ANOVAs, one for targets and the second for distractors, were conducted on the dependent variable HR₂ – FAR₂, with test type as the between-subjects factor. The mean values of HR₂ – FAR₂ at different levels of confidence by type of item (targets vs. distractors) and test type are reported in Table 7. One participant had to be excluded from the analysis in the NNNO condition, as they produced no FA₁.s.

The ANOVA conducted on targets revealed a main effect of test type, F(1, 50) = 8.579, MSE = .116, p = .001, η² = .225, a main effect of confidence, F(4, 200) = 73.504, MSE = .037, p < .001, η² = .595, both of which were qualified by a significant interaction: F(8, 200) = 3.039, MSE = .037, p = .016, η² = .108. Pairwise comparisons were then conducted comparing the NOOO and NNNO conditions, where the greatest difference in bias was detected. The comparisons showed better monitoring in the NNNO group, where responding was more conservative, than in the NOOO group, at all levels of confidence, all Fs(1, 34) ≥ 6.514, all ps ≤ .015, except the lowest (“2+”), F < 1.

When the ANOVA was applied to distractors, there was a main effect of test type, F(1, 49) = 5.034, MSE = .212, p = .010, η² = .170, but no main effect of confidence, F < 1. The main effect of test type was qualified by a significant interaction: F(8, 196) = 3.298, MSE = .030, p = .006, η² = .119. Pairwise comparisons
were again conducted comparing the NOOO and NNNO groups. In contrast to the target monitoring results, these comparisons showed better monitoring in the NOOO group, where responding was more liberal, than in the NNNO group, at all levels of confidence, all Fs(1, 33) ≥ 11.733, all ps ≤ .002, except the lowest (“2+”), $F < 1$.

As can be seen in Figures 12 and 13, monitoring was good for targets but poor for distractors. As in Experiments 1 and 2, six one-sample t-tests, one for each experimental condition, compared mean $HR_s - FAR_s$ against chance (collapsed across confidence levels). For targets, monitoring was above chance in all three experimental conditions: $t(18) = 10.976$, $p < .001$ ($M = .411$, $SD = .163$); $t(16) = 10.392$, $p < .001$ ($M = .361$, $SD = .143$); and $t(16) = 5.705$, $p < .001$ ($M = .207$, $SD = .149$), for the NNNO, NNOO, and NOOO groups, respectively. In contrast, distractor monitoring was significantly below chance in the most conservative NNNO group, $t(17) = -2.451$, $p = .025$ ($M = -.159$, $SD = .275$), whereas it did not differ from chance in either the NNOO group, $t(16) = -0.583$, $p = .608$ ($M = -.025$, $SD = .177$), or the NOOO group, $t(16) = 1.821$, $p = .087$ ($M = .060$, $SD = .136$).

**Discussion**

In summary, the Type-1 analysis indicated that test type successfully manipulated the response criterion (cf. Donaldson, 1996). That is, increasing the number of “old” categories from one (NNNO group) to two (NNOO group) to three (NOOO group) monotonically increased the $HR_s$, $FAR_s$, and $c$. However, the test type manipulation had no effect $d'$, suggesting that the underlying distributions were unaffected.
Table 7
Mean (Standard Deviation) HR, = FAR, by Level of Confidence, Type of Item, and Test Type Group (NNNO, NNOO, and NOOO) in Experiment 3

<table>
<thead>
<tr>
<th>Item and confidence level</th>
<th>NNNO</th>
<th>NNOO</th>
<th>NOOO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targets, “6”</td>
<td>.641 (.215)*</td>
<td>.480 (.203)</td>
<td>.360 (.251)*</td>
</tr>
<tr>
<td>Targets, “5+”</td>
<td>.602 (.207)*</td>
<td>.474 (.183)</td>
<td>.314 (.213)*</td>
</tr>
<tr>
<td>Targets, “4+”</td>
<td>.585 (.221)*</td>
<td>.418 (.204)</td>
<td>.216 (.225)*</td>
</tr>
<tr>
<td>Targets, “3+”</td>
<td>.259 (.206)</td>
<td>.292 (.220)</td>
<td>.096 (.194)*</td>
</tr>
<tr>
<td>Targets, “2+”</td>
<td>.095 (.148)</td>
<td>.139 (.136)</td>
<td>.047 (.115)</td>
</tr>
<tr>
<td>Distractors, “6”</td>
<td>-.205 (.309)*</td>
<td>.014 (.163)</td>
<td>.057 (.224)*</td>
</tr>
<tr>
<td>Distractors, “5+”</td>
<td>-.207 (.373)*</td>
<td>.007 (.235)</td>
<td>.117 (.232)*</td>
</tr>
<tr>
<td>Distractors, “4+”</td>
<td>-.194 (.350)*</td>
<td>-.018 (.235)</td>
<td>.079 (.195)*</td>
</tr>
<tr>
<td>Distractors, “3+”</td>
<td>-.142 (.292)*</td>
<td>-.070 (.214)</td>
<td>.059 (.168)*</td>
</tr>
<tr>
<td>Distractors, “2+”</td>
<td>-.046 (.212)</td>
<td>-.058 (.181)</td>
<td>-.011 (.083)</td>
</tr>
</tbody>
</table>

Note. HR, = Type-2 hit rate; FAR, = Type-2 false alarm rate; NNNO group = participants were presented with three response categories to respond “new” but only one for responding “old”; NNOO group = participants were presented with two categories each for responding “old” and “new”; NOOO group = participants were presented with three response categories to respond “old” but only one for responding “new.” *Indicates a statistically significant difference between the NNNO and NOOO groups.

As in Experiments 1 and 2, target monitoring was good in Experiment 3, whereas distractor monitoring was poor. However, one important difference between this experiment and the previous ones was that distractor monitoring in the NNNO group was significantly below chance rather than at chance. This result is predicted by the direct-translation hypothesis for liberal portions of the Type-2 ROC for cases in which the Type-1 criterion is conservative, such as in the NNNO group. For example, the theoretical ROC with crosses in Figure 6 (which corresponds to a conservative criterion shift) shows below-chance monitoring for Type-2 confidence levels of 4 and below. However, the same theoretical ROC in Figure 6 also predicts above-chance distractor monitoring at conservative Type-2 criteria, which is not apparent in the NNNO group’s empirical ROC in Figure 13. Although the NNOO group shows a pattern resembling the theoretical curve (i.e., the ROC for that group is slightly above chance for conservative Type-2 criteria but slightly below chance for liberal Type-2 criteria), it is really the NNNO group that is the best test of the theoretical prediction because responding is maximally conservative in that group (see Table 6). Another difference between the conservative-criterion theoretical ROC and NNNO group’s empirical ROC is that the former predicts that the liberal points will intercept the right-hand y-axis, whereas the latter shows no evidence of any intercept. We discuss possible reasons for these differences in the General Discussion section.

General Discussion

This article has introduced a new methodology for investigating cognitive and metacognitive processes in recognition memory: Type-2 SDT with ROC analysis. Although Type-2 SDT has recently been successfully applied to a number of different research domains in cognitive psychology—including cued recall (e.g., Higham, 2002; Higham & Tam, 2005, 2006), the strategic regulation of accuracy on multiple-choice tests (e.g., Higham, 2007; Higham & Arnold, 2007a, 2007b), answer changing on multiple-choice tests (Higham & Gerrard, 2005), unconscious perception (Kunimoto, Miller, & Pashler, 2001), and implicit learning (e.g., Tunc, & Shanks, 2003)—to our knowledge, it has not yet been used with recognition memory research. Given the popularity of more traditional, Type-1 (stimulus-contingent) ROC analysis with recognition memory, the absence of Type-2 analysis is rather surprising. We suspect that one reason for this absence is that the methodology for conducting such analysis has not heretofore been delineated. Our hope is that the current article describes the methodology in enough detail so that this gap in the literature is filled.

The application of Type-2 SDT to the word frequency effect in Experiment 1 yielded results that discriminated between two models of performance (see Models A and B in Figure 4). In short, the mirror effect obtained in Experiment 1 was attributable to Model B, with different positioning of both the target and distractor distributions with respect to frequency. Model A, which incorporated a within-list criterion shift, was rejected. A similar contribution was made in Experiment 2: Although the Type-1 data effectively eliminated Models A and B, they could not distinguish between Models C and D. However, Type-2 SDT demonstrated decisively that Model C, which incorporated a target distribution difference, a single distractor distribution, and no within-list criterion shift, was a better model of performance than Model D. Experiment 3 lent support for the direct-translation hypothesis that was used in Experiments 1 and 2 to map Type-1 confidence-inoldness onto Type-2 confidence-in-accuracy. Together, the results of both experiments demonstrate that the use of Type-2 in conjunction with Type-1 SDT can discriminate between competing models of recognition performance that cannot be discriminated with Type-1 SDT alone.

Type-1 analyses can sometimes distinguish between competing models of recognition performance when more complex designs are used. Thus, our argument is not that Type-1 SDT on its own has no value, but that using Type-1 and Type-2 SDT together can add diagnostic power to model testing. Importantly, the conclusions we reached regarding how best to model frequency-based and strength-based effects in recognition are generally consistent with other Type-1 research. For example, we concluded that both the frequency-based mirror effect in Experiment 1 and the strength-based effect on HR1 in Experiment 2 occurred because of distribution differences rather than within-list criterion shift. Morrell et al. (2002) and Stretch and Wixted (1998) reached similar conclusions using very different (Type-1) methodologies.

More recently, evidence of within-list criterion shifts has been obtained (e.g., Dobbins & Kroll, 2005; Hockley & Niewiadomski, 2007; Rhodes & Jacoby, 2007; Singer & Wixted, 2006; see also Verde & Rotello, 2007). However, mirror effects that occur as a function of strength within a test list of randomly intermixed strong targets, strong distractors, weak targets, and weak distractors, which would typically be seen as evidence of a within-list criterion shift, are rare. Indeed, to our knowledge, our experiments (e.g., Bruno et al., 2008; see also Tam, 2006) are the only ones to demonstrate such mirror effects. These experiments have provided some insight as to why mirror effects occur in some cases but not others: If participants deem the overall conditions of the experiment to be poor for memory (i.e., low global subjective memorability), then strength-based mirror effects occur within lists. Under
these conditions, participants appear to adopt different strategies for strong versus weak items if these items (predominantly distractors) do not provide any memorial information at test. In particular, lack of memory information for strong distractors is considered diagnostic of their newness, so they are confidently rejected via the metacognitive strategy (Strack & Bless, 1994), leading to a low FAR1. Conversely, forgetting is presumed for a portion of the weak items that provide no memory information. That is, Strack and Bless’s (1994) presuppositional strategy is adopted, which causes the proportion of “old” responses and the weak FAR1 to increase and a mirror effect to occur. On the other hand, if global subjective memorability is high, a similar metacognitive strategy is adopted for all distractors, yielding no difference between strong and weak FAR1s and no mirror effect.

In terms of other studies demonstrating within-list criterion shifts, one other seems particularly relevant to the current research. Singer and Wixted (2006) manipulated retention intervals and found that participants adopted a more conservative recognition criterion for items belonging to categories that were presented a short time before compared with items from different categories presented 48 hr before. Furthermore, this criterion shift occurred within lists suggesting that it was adjusted on an item-by-item basis. However, participants did not adjust their criterion in this manner if the long retention interval was only 20 min or 40 min instead of 48 hr. Although not a manipulation of strength via repetition per se, this result is generally consistent with the global subjective memorability hypothesis. That is, it is likely that participants would have assessed the global memory conditions to be poor if the test list contained items presented 48 hr before, but they were less likely to do so if the longest retention interval for any item was less than 1 hr. Consequently, when the long retention interval was 48 hr, participants may have attended to the category cues designating the retention interval for a given item at test and adopted different (metacognitive and presuppositional) strategies accordingly. Such strategic switching would have produced the mirror effect that was observed.

The Nature of the Strength of Evidence Dimension

A number of researchers have suggested that the FA1 portion of the word frequency effect is due to differential levels of memorability between low frequency and high frequency items (e.g., Brown et al., 1977). Usually, these memorability accounts rely on a concomitant criterion shift such that a more stringent criterion is used to judge the more memorable class of items. However, the current results suggest that criterion shifts did not occur between low and high frequency items, which might be interpreted to mean that subjective memorability played no role in creating the FA1 portion of the mirror effect in Experiment 1. However, in our view, subjective memorability need not take the form of variable criterion setting as Brown et al. (1977) proposed. If low frequency distractors are more subjectively memorable than high frequency distractors, this means that they possess evidence for newness, perhaps causing them to be located lower on the confidence-in-oldness scale than high frequency distractors (i.e., toward the “certain new” end of the Type-1 dimension).

If it is assumed that the FA1 portion of frequency-based mirror effect obtained in Experiment 1 was attributable to differential subjective memorability, then why did subjective memorability not exert a similar effect in Experiment 2? At test, there was a living/nonliving strength cue as well as test labels that clearly identified which distractors were weak and which were strong, so participants could easily have used these cues to reject strong and weak distractors at different rates. This result is particularly remarkable given that the inclusion of test labels for each item made strength highly salient compared with previous studies (e.g., Stretch & Wixted, 1998), in which participants were required to remember the association between strength and the strength-designating cue (e.g., color). One possible explanation might have to do with the fact that word frequency is an intrinsic cue, whereas strength is an extrinsic one. According to Koriat’s (1997) cue-utilization hypothesis, cues for confidence can be separated into three categories: intrinsic, extrinsic, and mnemonic. He found that participants are typically influenced by intrinsic cues (at least initially), but they discount extrinsic cues. Although the cue accessibility hypothesis was developed primarily for understanding prospective judgments-of-learning, it is conceivable that there are differential effects of intrinsic versus extrinsic cues on Type-1 and Type-2 retrospective confidence as well. As discussed above, it may be the case that extrinsic cues such as strength only influence retrospective Type-1 confidence (i.e., confidence-in-oldness) to distractors, leading to different FAR1s, when the overall conditions of the experiment are perceived to be poor for memory (Bruno et al., 2008).

The Direct-Translation Hypothesis

In all of our Experiments, we gathered only confidence-in-accuracy ratings. However, to derive predictions that could be tested with Type-2 SDT analysis, it was necessary to make some assumptions regarding the relationship between the underlying Type-1 and Type-2 dimensions. In making our specific predictions, we used a simple transformation, mapping absolute values on the Type-1 dimension onto the Type-2 dimension, meaning that confidence-in-newness and confidence-in-oldness transformed symmetrically into confidence-in-accuracy (e.g., Type-1 values of -3 [high confidence new] and 3 [high confidence old] both translated into 3 [high confidence correct]). This transformation rests on the assumption that confidence in the accuracy of the Type-1 response (i.e., Type-2 confidence) is based on exactly the same evidence used to make the Type-1 response itself. Analogous assumptions have been made by other researchers (e.g., Galvin et al., 2003; Kunimoto et al., 2001, Appendix B), and it forms an important basis of the direct-translation hypothesis.

The data from Experiment 3 were generally consistent with the direct-translation hypothesis, but the hypothesis might not be valid in other situations if there are variables that affect Type-2 confidence that have no effect on Type-1 accuracy or vice versa. In this vein, Koriat’s (1997) cue-utilization hypothesis, although developed primarily for prospective judgments-of-learning, suggests that several intrinsic, extrinsic, and mnemonic cues affect confidence not always related to memory access. For example, Kelley and Lindsay (1993) found that the ease with which memory candidates come to mind has an effect on retrospective confidence but no effect on accuracy. Even more relevant, however, are the results of Busey et al. (2000, Experiment 3) who found that high luminance of faces presented on a recognition memory test increased participants’ Type-2 confidence in the accuracy of their
recognition responses, but it had no effect on Type-1 recognition accuracy. Indeed, if the faces had originally been studied with low luminance, accuracy was quite poor despite the high confidence. In experimental situations such as these, in which there are variables that affect Type-2 confidence that do not affect Type-1 accuracy, generating a Type-2 model based on Type-1 data may be a more complex process than that which we used here.

Turning to the Type-2 ROC curves from Experiment 2 with these considerations in mind, we note that the shapes of Type-2 ROC curves were qualitatively quite different (see Figure 11) even though overall monitoring efficacy did not differ significantly between strong and weak distractors. Perhaps what is most striking about the distractor ROCs is that the FAR\textsubscript{2} at the highest level of confidence (6; bottom left in the ROC space) was considerably higher (0.228) in the strong condition than in the weak condition (0.088), causing monitoring to drop below the chance line in the former case but not in the latter. This difference in FAR\textsubscript{2} was significant, \(F(1, 37) = 7.978, \text{MSE} = .035, \text{p} = .008, \text{\textit{\textit{\textit{\textit{\textit{n}}}}} = 0.177.\) Critically, however, there was no corresponding strength difference in the FAR\textsubscript{1}s. In other words, participants were more willing to assign high confidence to strong distractors (judged “old”) than weak distractors (judged “old”; i.e., strong FAR\textsubscript{2} > weak FAR\textsubscript{1}) despite the fact that there was no difference in the actual proportion of “old” responses (i.e., strong FAR\textsubscript{2} = weak FAR\textsubscript{1}). This pattern of results constitutes a confidence–accuracy dissociation, and suggests that the strength manipulation had an effect on confidence-in-accuracy that it did not have on confidence-in-oldness. Such dissociations are problematic for the direct-translation hypothesis.

If there are factors affecting confidence-in-accuracy that have no effect on confidence-in-oldness, this may partly explain why the distractor ROCs in the NNNO group of Experiment 3 (see Figure 13) did not conform perfectly to the theoretical prediction for a conservative Type-1 criterion (see middle panel of Figure 5 and ROC curve with crosses in Figure 6). In particular, the liberal points of the empirical Type-2 ROC did not intercept the right-hand y-axis as predicted by the direct-translation hypothesis. The curve is predicted to intercept this axis because as the Type-2 criterion is moved toward liberality to generate the Type-2 ROC, there are no incorrect responses (derived from FA\textsubscript{1}s) remaining once the point on the scale associated with the Type-1 criterion is reached, causing the FAR\textsubscript{2} to equal 1.0. In Figure 5, this point on the scale is 0.5. However, if confidence-in-accuracy judgments are not strictly determined by confidence-in-oldness, and extraneous factors add systematic influence or noise to them, then the FAR\textsubscript{2} might fall short of 1.0 at lower Type-2 confidence levels, as was observed in Experiment 3.

Another potential problem with the direct-translation hypothesis is that it does not predict the chance-level monitoring of high frequency distractors in Experiment 1 or of strong distractors in Experiment 2. As noted above, chance-level monitoring could be achieved by setting both the Type-1 criterion and the mean of the distractor distribution at zero on the Type-1 dimension. This would then cause the correct (CR\textsubscript{s}) and incorrect (FA\textsubscript{1}s) Type-2 item distributions to overlap completely, rendering chance-level Type-2 discrimination. However, such an alteration would produce a FAR\textsubscript{2} of 0.500, which is inconsistent with the empirical FAR\textsubscript{2}s (0.335 for high frequency distractors in Experiment 1 and 0.167 for strong distractors in Experiment 2). It is important to note, however, that a model might be created that maintains the fundamental assumptions of the direct-translation hypothesis but that does not have equal spacing between scale values on the confidence-in-oldness dimension, or spacing between scale values that is unequal between the confidence-in-oldness and confidence-in-accuracy dimensions. For example, consider again the Type-1 model in the top-left of Figure 5, which is translated into Type-2 space in the top-right of the figure. In this model, the FAR\textsubscript{1} is less that .50, which is similar to the empirical data from our experiments. Suppose that instead of equal distance between all scale values on the confidence-in-oldness dimension, the distance between the confidence-in-oldness scale values above the criterion (applicable to the FAR\textsubscript{1}) was half the distance between the confidence-in-oldness scale values below the criterion (applicable to the CRR\textsubscript{1}). In such a model, the height of the FAR\textsubscript{2} portion of the curve on the confidence-in-oldness scale would approach zero at +4 instead of +2 as is currently shown. The effect of this alteration after taking absolute values to translate the model into Type-2 space would be to stretch the incorrect response distribution (corresponding to the FAR\textsubscript{1}) toward the high end of the confidence-in-accuracy scale. In other words, the mean of the incorrect item distribution shown in the top-right of Figure 5 would increase as the distribution is stretched to the right, whereas the correct response distribution would remain the same. This increase in the mean would decrease monitoring, potentially to chance levels, despite the fact that the FAR\textsubscript{1} was below .50.7

Importantly, however, the data from Experiment 3 suggest that potential problems with the direct-translation hypothesis were unlikely to have undermined the central conclusions we reached in the current research that rested upon it. Nonetheless, the validity of the direct-translation assumption is worthy of further scrutiny if Type-2 SDT is to be used in conjunction with Type-1 SDT in future research on recognition memory. This future research might focus on a more detailed, quantitative description of the mapping of the Type-1 dimension onto the Type-2 dimension or the identification of factors that might affect Type-1 versus Type-2 confidence differently. For now, our aim is simply to introduce the application of Type-2 SDT analyses to old/new recognition in the hopes that it might provide new ways of understanding recognition phenomena and testing models of recognition memory that Type-1 SDT cannot offer on its own.

6 Twelve participants were eliminated from this analysis because of missing data (i.e., they produced FA\textsubscript{s}).

7 We thank John Wixted for raising this possibility.

References
TYPE-2 SIGNAL DETECTION THEORY AND RECOGNITION MEMORY


Received December 18, 2007
Revision received June 27, 2008
Accepted July 2, 2008

Call for Nominations: Psychology of Violence

The Publications and Communications (P&C) Board of the American Psychological Association has opened nominations for the editorship of Psychology of Violence, for the years 2011–2016. The editor search committee is chaired by William Howell, PhD.

Psychology of Violence, to begin publishing in 2011, is a multidisciplinary research journal devoted to violence and extreme aggression, including identifying the causes and consequences of violence from a psychological framework, finding ways to prevent or reduce violence, and developing practical interventions and treatments.

As a multidisciplinary forum, Psychology of Violence recognizes that all forms of violence and aggression are interconnected and require cross-cutting work that incorporates research from psychology, public health, neuroscience, sociology, medicine, and other related behavioral and social sciences. Research areas of interest include murder, sexual violence, youth violence, inpatient aggression against staff, suicide, child maltreatment, bullying, intimate partner violence, international violence, and prevention efforts.

Editorial candidates should be members of APA and should be available to start receiving manuscripts in early 2010 to prepare for issues published in 2011. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

Candidates should be nominated by accessing APA’s EditorQuest site on the Web. Using your Web browser, go to http://editorquest.apa.org. On the Home menu on the left, find “Guests.” Next, click on the link “Submit a Nomination,” enter your nominee’s information, and click “Submit.”

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Emnet Tesfaye, P&C Board Search Liaison, at Emnet@apa.org.

Deadline for accepting nominations is January 31, 2009, when reviews will begin.