2014

Feedback as a Source of Criterion Noise in Recognition Memory

Bryan Franks
Louisiana State University and Agricultural and Mechanical College

FEEDBACK AS A SOURCE OF CRITERION NOISE IN RECOGNITION MEMORY

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Arts

in

The Department of Psychology

by
Bryan A. Franks
B.A., University of New Mexico, 2012
August 2014

Table of Contents

Abstract

In two experiments, I investigated whether providing accuracy feedback on recognition memory tests affects discriminability of encoded targets from lures. The primary hypothesis was that feedback is a source of criterion noise which leads to lower discriminability. Additionally, it was predicted that separate sources of criterion noise might have additive effects. In both experiments, the presence of feedback was manipulated within-subjects. In Experiment 1, participants completed two recognition tests in which they made either "old/new" decisions or responded using an 8-point confidence scale. Feedback lowered discriminability for both response type conditions, although a slightly larger deleterious effect was observed in the "old/new" response condition. Whether people responded either with "old/new" decisions versus on an 8-point confidence scale had no effect on discriminability. In Experiment 2, I manipulated the strength of study items whereby half of the items were studied once (weak) and the other half were studied four times (strong). At test, these targets were intermixed with an equal number of lures. Additionally, the presence of color cues indicating the expected strength of test items was varied between-subjects. Feedback decreased discriminability, although this was primarily for the strong items. The presence of color cues marking expected strength had no effect on discriminability. Taken together, these results suggest that feedback has a deleterious effect on recognition discriminability and that this may result via feedback introducing criterion noise into the recognition decision.

Introduction

Null effects of corrective test feedback on discriminability of studied from nonstudied items dominate the recognition memory literature. Feedback is generally unhelpful to discriminability when using continuous recognition (Estes & Maddox, 1995), under manipulations of base rates (Rhodes & Jacoby, 2007; Kantner & Lindsay, 2010; Selmeczy & Dobbins, 2013), memory strength (Verde & Rotello, 2007; Hicks & Starns, 2014), with older adults (Jennings & Jacoby, 2003), or with exotic stimuli such as complex melodies or paintings (Lindsay & Kantner, 2011). These results are somewhat surprising given that we might expect feedback to improve discriminability in several possible ways, such as allowing people to adapt a more optimal response criterion (Kantner & Lindsay, 2010) or by enhancing metacognitive monitoring of test stimuli (Selmeczy & Dobbins, 2013). Some formal models of signal detection theory (SDT) rely explicitly on the integration of feedback and stimulus representations for making recognition decisions. For example, Turner, Van Zandt, and Brown (2011) proposed that when feedback is present, it can help improve recognition performance by allowing people to update their information regarding both signal and noise distributions.

Although most researchers that have investigated the effects of feedback on recognition memory have typically noted no difference in discriminability for participants in control or feedback conditions, few, if any, have considered feedback to be a source of harm on recognition judgments. For example, Kantner and Lindsay (2010) conducted four experiments expecting a positive feedback effect but instead found three null effects and a significant negative effect of feedback in Experiment 2. They also noted that in all their experiments, discriminability was numerically lower in the feedback conditions. Kantner and Lindsay (2010) dismissed the possibility that their lack of positive effects was due to a Type II error in their study because there was no trend for feedback to improve discriminability. Selmeczy and Dobbins (2013) also dismissed the possibility of a Type II error, noting that their pattern of results did not display a trend toward feedback improving recognition sensitivity. However, neither study entertained the possibility that with sufficient statistical power, feedback may actually have a negative effect on performance.

One plausible explanation is that introducing feedback may harm recognition via criterion noise. According to Benjamin, Diaz, and Wee's (2009) noisy decision theory of signal detection (ND-TSD), criterion noise is introduced into recognition decisions via two mechanisms: the maintenance and updating of decision criteria. Both of these processes place an encumbrance on memory, which can subsequently lead to poorer recognition performance (Benjamin et al., 2009). Incorporating feedback into recognition decisions can easily be thought of as a way to promote appropriate updating and placement of response criteria. In fact, using feedback to better control criterion placement, rather than to influence discriminability, is often the primary reason that feedback is applied (e.g., Verde & Rotello, 2007). Thus, the primary goal of this study is to assess whether or not corrective feedback at test can introduce criterion noise leading to a decrement in memory discriminability. In the next few sections I first discuss a signal detection framework for recognition memory decisions. Next, I present evidence that feedback appears to be detrimental to recognition discriminability. Finally, I discuss how feedback may be viewed as causing criterion noise (or criterion variance).

Signal Detection Theory

Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005), has provided a successful framework in which to understand recognition memory. Consider a typical laboratory recognition study in which people study a list of words and take a test that contains a mixture of studied and non-studied words. According to SDT, these two types of test items can be represented as separate distributions that vary on the singular dimension of familiarity or memory strength. Studied items comprise the target (signal + noise) distribution and sit farther to the right of the lure distribution (noise only). This is depicted below in Figure 1.

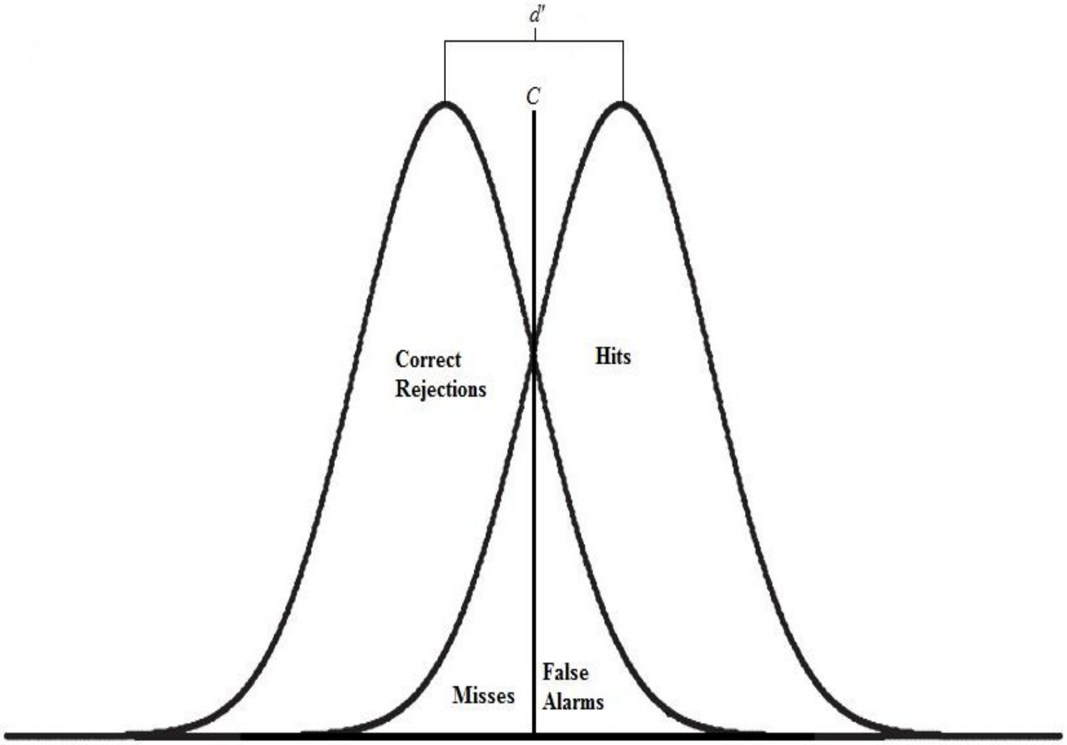


Figure 1. Theoretical target and lure distributions plotted on an axis of memory strength with stronger items in memory farther to the right. The vertical line labeled “C” represents an optimal criterion.

Because memory is not perfectly veridical and people come into the laboratory with some pre-existing level of familiarity with all test items (at least when they are known words), these distributions overlap to some degree. This overlap requires that people set a criterion on this axis of memory strength, which is essentially a threshold by which test items are judged. This criterion is often assumed to be fixed throughout the entire test (e.g., Stretch & Wixted, 1998). Importantly, most applications of SDT to recognition memory assume that unlike noise imparted by the test stimuli, there is no decision noise introduced by the criterion itself. That is, criterion variance equals zero (Green & Swets, 1966). If a test item exceeds this criterion it will be called "old" otherwise it will be called "new." Items correctly called "old" are denoted as hits and those incorrectly called "old" are false alarms, the proportions of which can be used to calculate an overall hit rate (HR) and false alarm rate (FAR). These measures can be used to further define recognition performance by calculating a measure of discriminability (*d'*) which indexes the distance between the peaks of the signal and noise distributions in standardized units. Additionally, various measures of response bias and/or criterion placement can be calculated which represents either a person's overall tendency to call test items "old" or to estimate the point along the familiarity axis at which a criterion sits. The ideal observer will set a criterion that maximizes the HR and minimizes the FAR.

How Does Feedback Affect Recognition Discriminability?

The vast majority of studies have found either a null or negative effect of feedback on recognition memory discriminability. Some of the earliest research suggesting a feedback-induced improvement came from Titus (1973) who had participants study CVC trigrams and take a recognition test in which he manipulated the presence of feedback at test as well as participants' awareness of the base-rates of test items. For all subjects, only 20% of test items were old. This proportion of test items requires that people set a very conservative criterion in order to be most optimal. Titus analyzed HRs and FARs across the 75-item test in 3 blocks and found that when people were unaware of the base-rates of test items but received feedback, a conservative shift in criterion was observed as participants' HRs and FARs decreased across blocks, although the decrease in the FAR was most apparent. Statistically, neither prior knowledge of the probability that a test item was studied nor feedback affected recognition discriminability. However, *d'* values I estimated from the reported HRs and FARs (based on an equal variance assumption) suggest that feedback helped improve performance overall, most notably when subjects had no knowledge about the probability of target items being old (i.e., an improvement from 1.61 to 1.97). This is what would be expected when criterion placement is nonoptimal in the control condition. The results also suggested an improvement in *d'* when subjects were informed about the target probability, as performance increased from 1.61 in the control condition to 1.84 with the additional instruction.

An early study by Clark and Greenberg (1971) suggests that the presence of feedback (or knowledge of results) harms recognition memory. Following the learning of 18 CVC trigrams, participants took 3 successive blocks old/new recognition tests. In each block, the same 18 trigrams were the targets and a unique set of 12 trigrams was used as lures. Averaged over the blocks, d' was 1.30 for the no-feedback group and 1.02 for the feedback group. It should be noted that this main effect was obtained in the context of interactions with other variables, including performance across blocks (1 through 3) and another factor of induced anxiety during the test procedure.

This line of research perplexingly remained at a standstill over 20 years until Estes and Maddox (1995) studied the effects of feedback using a modified continuous recognition paradigm. In their paradigm, a set of stimuli was studied initially and people were instructed to call each item on this list "new." After this phase, stimuli were continuously introduced in a hybrid learning/testing procedure. Each item was presented for an "old/new" decision in different testing blocks. Within each block, three types of items occurred: brand new items, items from the initial study list, and items repeated from the prior block. Items from the study list and those repeated from prior blocks are considered targets. Brand new items in each block are considered lures. This stands in contrast to the more common procedure in which people study a set of items in a learning phase that is entirely distinct from the testing phase. They manipulated stimulus type between-subjects such that participants studied random digits, letter trigrams, or words. Feedback was also manipulated across participants along with the base-rates of test items whereby either 67% or 33% of the test items were old. Participants were unaware of the base-rate manipulation. Across two experiments, marginally significant positive feedback effects were found for digit and letter stimuli. For the word stimuli, the presence of feedback again exerted no effect of discriminability, though there was a numerical benefit for those receiving feedback, particularly in Experiment 2. Regarding response bias, participants appropriately adopted either a liberal or conservative “old”-saying bias for the 67% and 33% old conditions when feedback was present, although feedback did not significantly impact subjects’ criterion for words (Estes & Maddox, 1995). As with the Titus (1973) study, feedback influenced criterion-setting in this study by getting people to shift to a criterion placement more consistent with the base rates of the target and lure items and seemed to nominally improve discriminability. However, the results of this study should interpreted cautiously for two reasons. First, few subjects were tested in each condition, which limits both statistical power and generalizability. Second, the use of the hybrid continuous recognition paradigm makes it unclear whether feedback is impacting processes at encoding or retrieval.

Rhodes and Jacoby (2007) included feedback with recognition tests in order to assess whether participants could dynamically shift their criterion when different base-rates of old items covaried with a particular study location. In Experiment 3, participants completed 4 study-test blocks in which they received feedback in either the first two test blocks or the last two. They found that the presence of feedback and awareness of the base-rate manipulations were necessary for participants to appropriately shift their decision criteria. They also found that when feedback was given for the first two test blocks, discriminability was best in the first block and subsequently dropped off for the other three blocks. For participants who received feedback on the last two test blocks, discriminability was consistent across on blocks 1 and 2 and declined when they received feedback for blocks 3 and 4. It is difficult to ascertain the overall influence of feedback across these conditions, because explicit d' values were not presented separately for each block within each condition, but feedback was associated with criterion shifting.

Likewise, Verde and Rotello (2007) also found that the presence of feedback was needed in order for participants to optimally shift their criterion for test items that varied in memory strength. Although they did not manipulate feedback within a particular experiment, their findings across two experiments are noteworthy. In both Experiments 2 & 5, participants studied a list of words in which some were studied once (weak condition) and some words were studied four times (strong condition) and took a recognition test comprised of targets and lures from each item class of items. No participants received feedback in Experiment 2 and all participants in Experiment 5 received trial-by-trial accuracy feedback. Otherwise the procedures in these experiments were identical. Although a criterion shift was observed only in the presence of feedback (Exp. 5), discriminability was numerically better for both strong and weak items when feedback was absent (Verde & Rotello, 2007). This study is somewhat unusual in that the strong and weak items were tested in a particular sequence, with 40 strong items and 40 lures in the first test block and the 40 weak items and lures in the last test block. Feedback in Experiment 5 prompted people to decrease the HR first (strong) testing block relative to Experiment 2, leaving the FAR relatively unaffected. In addition, feedback increased both the HR and FAR in the second (weak) testing block, leaving discriminability only slightly lowered in the feedback experiment. Of course, one drawback is that feedback is being compared across experiments, rather than being manipulated within a single experiment.

Feedback was also examined in a study by Han and Dobbins (2008) who were interested in whether people could shift their criterion without manipulations of memory strength or base-rates of test items and without participants' awareness of test manipulations. In Experiment 1, participants completed 4 study-test cycles as in Rhodes & Jacoby (2007) and corrective feedback was given in 2 of the 4 blocks. However, unlike Rhodes and Jacoby (2007) discriminability was the same for feedback and no-feedback blocks. Experiments 2 & 3 introduced two types of false positive feedback in order to see if participants would adjust their criterion in response to the feedback. Indeed, participants were able to use the feedback to shift their criterion without changes in discriminability. However, both of these latter experiments lacked a no-feedback group so the impact of the different types of feedback on discriminability could not be assessed.

This limitation was addressed in a follow-up study in which two study-test blocks that included different false feedback manipulations preceded two additional blocks in which no feedback was given (Han & Dobbins, 2009). Again, the authors were primarily interested in participants' ability to incorporate feedback in order to shift their decision criterion. Criterion shifts were readily observed in both experiments while discriminability was either unaffected by presence of feedback (Exp. 1) or declined across blocks (Exp. 2; Han & Dobbins, 2009). This finding echoes that of Rhodes and Jacoby (2007) who found that when feedback was present in the first two study-test blocks, discriminability was best in the first block and declined afterwards. One potential limitation of this study is that Han and Dobbins (2009) did not counterbalance the order in which participants received feedback, as blocks with feedback always came before no-feedback blocks.

Some of the most comprehensive work recently that has examined how feedback affects recognition memory is that of Kantner and Lindsay (2010). They were interested in whether feedback could enhance discriminability when participants completed a single study phase followed by a single test. They manipulated feedback between-subjects and across four appreciably distinct experiments found null feedback effects in three of them. However, in Experiment 2 in which they also manipulated the base rates of test items, feedback significantly lowered discriminability.

These null effects led the authors in a later study to consider whether the stimuli used in a recognition paradigm would affect whether or not feedback was helpful. Specifically, in a multitude of experiments, Lindsay and Kantner (2011) examined if feedback could enhance recognition for complex stimuli such as Korean melodies, famous paintings, and verses of poetry. For Korean melodies, two experiments yielded a small but significant positive effect of feedback. In contrast, two conceptual replications of these experiments again using Korean melodies produced null effects of feedback. In one replication, the authors manipulated both feedback and recognition responses whereby participants make either “yes/no” decisions or respond on a 6-point confidence scale. Discriminability was numerically lower in the feedback condition, indicating no benefit from feedback. For the experiments using either famous paintings or poetry, all of them with the exception of one of the poetry studies again resulted in a null effect of feedback. These results were found in concert with a variety of other manipulations such as participants’ motivation, test format (yes/no or rating scale response), orientation tasks, and study list presentation (human or robot voices). Appropriately, the authors warn caution when considering the small positive effects found in this study as 13 of the 16 experiments resulted in null effects.

Another recent examination of feedback in recognition memory was by Hicks & Starns (2014) who were interested in delineating the circumstances under which within-list strength-based criterion shifts could be facilitated by manipulating test composition. Similarly to Verde and Rotello (2007), memory strength was manipulated via study repetitions whereby weak items were studied only once and strong items studied four times. Tests were comprised of 80 items that had an equal number of strong targets and lures as well as an equal number of weak targets and lures. In some conditions, the lures were designated as strong and weak only by a color cue, setting up the expectation that participants would treat them differently based on their expected strength (i.e., in being compared with either strong or weak target items in the same color). In each of their first two experiments, they manipulated the presence vs. absence of this color cue marking, with the prediction that color marking should enable criterion shifting between strong and weak test blocks, whereas the lack of such marking would not. Test items were presented in strong and weak blocks, the length of which was varied between-subjects. Additionally, feedback was not given in Experiment 1, but all participants in Experiment 2 received corrective feedback. Results indicated that both the presence of color cues and the presence of feedback independently harmed discriminability (Hicks & Starns, 2014).

Selmeczy and Dobbins (2013) explored the interplay between metacognitive monitoring and feedback using cues about the probability of a given test item being old or new. The probability cues consisted of indications prior to each test item about its likelihood of being old, with either a “likely old” or “likely new” statement. Some trials were preceded by these cues and other trials were not. In their first experiment, these cues were correct 75% of the time. Feedback was manipulated between subjects. Selmeczy and Dobbins predicted that optimal criterion shifts and better discriminability would result for cued versus uncued trials and that feedback would help people adopt appropriate criteria. In both experiments, participants made “old/new” decisions followed by a confidence judgment. These confidence judgments were later correlated with recognition accuracy to produce a measure of metacognitive monitoring. Overall, discriminability was better in cued trials but feedback did not improve discriminability across the board nor did it selectively help only the cued trials. Although they only presented analyses that are collapsed across feedback groups, as feedback did not yield a significant effect, the authors mentioned two findings of importance. First, numerically the feedback had a negative effect. Second, the worst discriminability was observed on cued trials when feedback was present (Selmeczy & Dobbins, 2013). In Experiment 2, they manipulated cue validity in which test items were preceded by a screen that correctly indicated the cues given were correct either 65% or 85% of the time. Again, they replicated their results from Experiment 1 such that feedback did not improve performance selectively or overall and that discriminability was best for cued trials. Interestingly, in both experiments the metacognitive monitoring scores were lower for those in feedback conditions (Selmeczy & Dobbins, 2013). This suggests that feedback might introduce some uncertainty into the recognition decision that is not present for those who receive no feedback.

Null Effects or Type II Errors?

Much of the foregoing analysis is summarized below in Table 1 as a listing of feedback-related effect sizes from prior work.

Table 1. Negative effects of feedback on recognition discriminability.

Study	Experiment	Sample Size	Cohen’s d
Kantner & Lindsay (2010)	Exp. 1	46	.40
	Exp. 2	71	.58
	Exp. 3	43	.27
	Exp. 4	77	.28
Lindsay & Kantner (2011)	All 16	538	.03
Clark & Greenberg (1971)	Exp. 1	30	.17
Verde & Rotello (2007)	Exps. 2 & 5	53	.35
Han & Dobbins (2008)	Exp. 1	16	.00
Hicks & Starns (2014)	Exps. 1 & 2	596	.26

In this table, I included only those estimates in which effect size was reported or could be reasonably estimated. Titus (1973) did not include *d’* values or estimates of variability for them Estes and Maddox (1995) did not report any measures of variability for their reported *d’* values, and only reported that *F* values were less than one in their ANOVA models examining the influence of stimulus type and feedback on *d’*. Hence the effect size for these experiments could not be reasonably estimated. It is important to note these absences from Table 1, because they also represent the only cases in which at least nominal positive feedback effects have been reported. Additionally, the Rhodes and Jacoby (2007) study did not report feedback vs. no-feedback conditions in enough detail to reasonably estimate an effect size. The descriptions of their analyses imply that feedback either had a true null effect or perhaps a slight negative one. The remaining entries in Table 1 suggest an overall trend for negative influences of feedback when there are any above-zero effect sizes. Only the no-stress condition from the Clark and Greenberg (1971) study was included in this table, because it is most comparable to the other listed studies and to the experiments reported later.

Although most of the findings regarding feedback were not statistically significant as reported in their respective publications, the range of effect sizes in Table 1 is considerable, with some studies finding a true null effect size (Han & Dobbins, 2008) up to a medium-to-large effect size showing a decline in discriminability (Cohen's d = .53) in Kantner & Lindsay's (2010) study. Using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007), Figure 2 displays the sample sizes needed to produce a significant negative influence of feedback for various levels of statistical power for a within-subjects manipulation assuming the following parameters: population effect size of Cohen's d = .30, an estimated population correlation between repeated measures of ρ = .503 derived from pilot data, and a Type I error rate of .05.

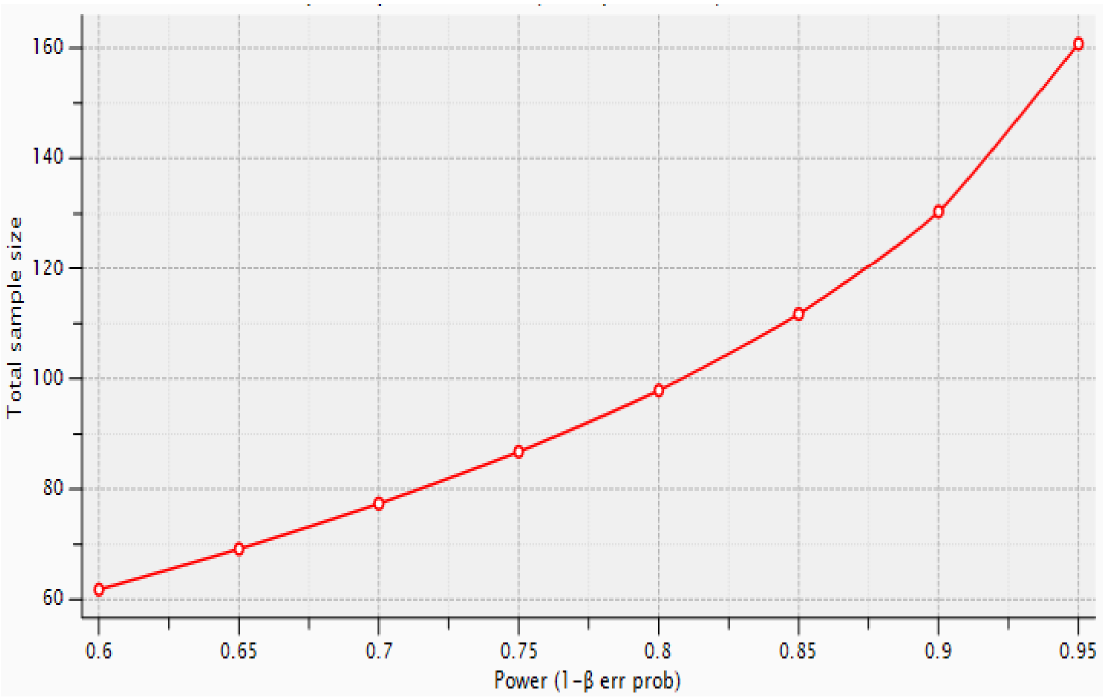


Figure 2. Sample size requirements for a given level of a prior statistical power to detect an influence of feedback using a within-subjects design assuming a population effect size of Cohen's d = .30 and a population correlation between repeated measures of ρ = .503. Type I error rate equals .05.

When power is set at .80 (i.e., 80%), a within-subjects manipulation of feedback would require a total sample size of 90 participants, whereas a between-subjects manipulation of feedback would require 352 participants. Note that the Hicks and Starns (2014) between-subjects comparison of feedback versus no feedback involved over 500 subjects. Although significant results with large effect sizes have been found with smaller samples (e.g., Kantner & Lindsay, 2010), this illustrates the point that although noteworthy effects of feedback might exist, they can be difficult to detect as significant depending on the experimental design and manipulations used.

Noisy Decision Theory of Signal Detection

The overall results from Table 1 suggest very small effect sizes associated with positive influences of feedback, but small-to-moderate effects sizes in the negative direction. On average, the data suggest that feedback is likely more harmful than helpful. Given the assumption that feedback might be doing some harm, one must consider a theoretical basis for it. One candidate process is increased criterion variability caused by feedback. The noisy decision theory of signal detection (ND-TSD; Benjamin et al., 2009) is a recent example highlighting the possibility that criterion noise can disrupt recognition memory processes. This instantiation of SDT primarily contrasts with the classic SDT as outlined above in its assumptions regarding response criterion. As mentioned earlier, classic STD (Green & Swets, 1966; Macmillan & Creelman, 2005) assumes that criterion variance is non-existent or negligible. In contrast, ND-TSD postulates that the response criterion is a random variable allowed to vary from trial to trial (Benjamin et al., 2009). As a consequence, criterion noise can be introduced into recognition decisions. Criterion noise is essentially a memory burden that can result from simply trying to maintain a response criterion or by attempting to update a criterion. Regarding the maintenance of response criterion, the authors posit that the use of a criterion to make recognition decisions requires that a person remember what that criterion value is from trial to trial. In a basic recognition paradigm where people study a single list of words and make simple “old/new” decisions, only a single criterion is needed for the entire test and participants must remember to apply that standard of evidence to every test item. In contrast, consider the case in which at test participants make recognition decisions using a 6 point confidence scale. In order to make recognition decisions in this context, participants must establish and use five different criterion values, one for each confidence boundary. Because of the additional memory resources needed to maintain and switch between multiple criterion values, discriminability performance could be worse in this situation. Thus, ND-TSD predicts that having to use and remember multiple criterion values creates criterion noise subsequently leading to worse recognition performance (Benjamin et al., 2009).

Benjamin et al. (2009) aimed to demonstrate that criterion noise contributed significantly to recognition decisions by having participants complete an ensemble recognition task and then modeling individuals’ response frequencies to evaluate whether their data better fit statistical models of discriminability that assume either zero (SDT) or non-zero (ND-TSD) criterion variability. In their experiment, participants studied a list of words and took an ensemble recognition test in which set size was manipulated. On each test trial, participants were presented with one, two, or four items together and asked to make a recognition decision on the entire set of items. The items in each ensemble were either all old or all new items. This manipulation of set size was intended to allow the authors to separately estimate the unique contribution of stimulus and decision noise, where stimulus noise is the uncertainty introduced by the test items themselves and decision noise reflects criterion variability. Benjamin et al. reasoned that set size should affect stimulus noise but not criterion noise because each ensemble is supposed to be evaluated with a single criterion. The results of their model fitting favored ND-TSD over traditional SDT theory, which led the authors to suggest that criterion noise plays a large, meaningful role in recognition decisions (Benjamin et al., 2009).

Benjamin, Tullis, and Lee (2013) have recently provided further evidence in support of ND-TSD. In this study, they evaluate the claim of ND-TSD that maintaining a criterion introduces noise by manipulating test format. After studying a list of words, participants took a recognition test in which they made simple “old/new” decisions, or responded on a 4-point or 8-point confidence scale. Because making recognition decisions using confidence scale ratings requires a participant to maintain multiple confidence criteria, updating these multiple criteria should produce more criterion noise. Thus, ND-TSD predicts that discriminability should be worse when confidence ratings are used. In line with their predictions, Benjamin et al. (2013) found that recognition discriminability was best when participants made “old/new” judgments and dropped significantly as more decision points were added.

With regard to the adjustment or updating of a criterion relevant to the present focus on feedback, manipulations attempting to get participants to change their criterion also introduce noise into the recognition decision (Benjamin et al., 2009). Again, this is because doing so places a non-trivial memory load on the recognizer. Consider the case where participants make “old/new” decisions and accuracy feedback is either present or absent. When feedback is not given, participants have no basis for updating their criterion and subsequently only have to remember a single criterion value that may not change much throughout the test. Conversely, when feedback is present, it serves as an external recommendation by which participants attempt to adjust their criterion. When a participant responds “old” and is given feedback that her decision was wrong, she may adjust the criterion for the next few test items to be slightly more conservative. Similarly, when she responds “new” incorrectly, she may adopt a slightly more liberal response criterion. This constant adjustment of a criterion across a test is not only cognitively demanding (Rhodes & Jacoby, 2007), but also requires that a participant remember the updated criterion value and then use it accordingly. Thus, over the course of a memory test, a no-feedback condition only requires participants to remember a singular criterion value, whereas a feedback condition prompts participants to adjust and remember new criterion values multiple times throughout the course of a test. It may be crucial whether the adjustment of a criterion is systematic and helpful, versus random and unhelpful. For example, when feedback is applied to encourage people to adjust from a nonoptimal criterion placement to a more optimal one, feedback should likely help performance. Indications of this type of help can be seen in the work by Titus (1973) and Estes and Maddox (1995), although Kantner and Lindsay's (2010) second experiment also used extreme base rates and found a negative influence of feedback. But when people may already be optimal in their spontaneous criterion placement, feedback may cause them to adjust in nonoptimal ways, creating noise. The results in other experiments by Kantner and Lindsay (2010) and by Hicks and Starns (2014) are consistent with this possibility.

Although recent evidence has supported ND-TSD (Benjamin et al., 2009, 2013), the theory is not without its critics. Kellen, Klauer, and Singmann (2012) re-analyzed the data set from Benjamin et al. (2009) and also provided new data from a recognition test in which they manipulated their subjects' responses by having them give either confidence ratings or ranking judgments. On confidence rating test trials, participants saw a single test probe and were asked to give it a confidence judgment using a 6 point rating scale. For ranking trials, participants were shown four test items and, knowing that only one of them was old, asked to rank order each item on its probability of being previously studied. This ranking task was supposed to be analogous to a forced-choice alternative task whereby participants can make decisions based solely on memory strength without reference to any criterion and thus free of any criterion noise. Kellen et al. (2012) argue that the results of their modeling of the recognition data indicate that ND-TSD does not provide a substantive account of signal detection over and above classic SDT. For most of their sample, criterion variability was estimated to be zero and the amount of variability for the few who displayed any at all was negligible. Hence, the authors concluded that ND-TSD does not provide a substantive account of recognition memory beyond that of traditional SDT. These findings stand in stark contrast to Benjamin et al. (2009) who argued that the presence of criterion noise has a substantial impact on recognition performance.

The Hicks and Starns (2014) work also represents another way in which ND-TSD's predictions have not borne out. In their work, the finding that color cues indicative of memory strength actually lowered discriminability contradicts a particular claim of ND-TSD. According to Benjamin et al. (2009), when test items vary on a particular dimension (e.g. memory strength), criterion variability is greater when the observer samples test items that have a larger range on that dimension. When the testing environment does not readily allow the observer to treat distinct classes of test items differently (e.g. strong or weak items), they will sample across the entire range of old items, thereby increasing the range for criterion variance as well. However, when these classes of items are readily distinguished, as is the case with the color cue manipulation, participants should be able to treat these strong and weak items differently and thus separately estimate the range of memory strength for weak and strong items. Consequently, rather than having large criterion variability across all test items, this separate estimation reduces criterion variability (and hence criterion noise) because each class of items has its own amount of criterion variability which is smaller than the variability that comes from treating all old items similarly. Thus, ND-TSD predicts that the color marking manipulation of Hicks and Starns should actually reduce criterion variability leading to better discriminability.

Hicks and Starns (2014) reported the opposite result: color marking produced a decrement to recognition discriminability. On the surface, this result suggests that the use of color marking to adjust one's criterion may still have created noise. In many other studies, information regarding test cues indicating base rates (ex. Van Zandt, 2000; Aminoff et al., 2012) or memory strength (ex. Verde & Rotello, 2007; Hicks & Starns, 2014) are given as instructions before a testing phase and participants are required to use the cues on their own. That is, they must keep particular information or rules in mind about what different cues represent, select the appropriate rule for each test item, and then try to use the cue to make a memory decision. Because of the extra cognitive effort required to use cues in these situations, criterion noise may be more apparent. However, one must also acknowledge that the effect size associated with the color marking decrement was small. In addition, Hicks and Starns' manipulation of color marking depended on people noting how the colors differentiated strong from weak items and keeping that in mind throughout the test on their own. However, other studies have administered test cues in a manner that reduces the cognitive load on participants. For example, Selmeczy & Dobbins (2013) provided alerting cues for each test item *individually* right before its presentation. Thus, when the testing environment is highly supportive, such as by giving cues individually for each test item (ex. Selmeczy & Dobbins, 2013; Bruno, Higham, & Perfect, 2009), criterion noise may be dramatically reduced.

Another difficulty in offering criterion noise as a mechanism by which feedback harms recognition is that the presence of criterion noise should have noticeable effects on recognition performance whenever it is present. However, if criterion noise affects recognition to such a large degree, why do non-significant feedback effects run rampant in the literature? The most likely reason is that the effects of criterion noise are slightly more moderate than Benjamin et al. (2009) suggest and that the majority of experiments examining feedback in recognition have not possessed sufficient statistical power to find significant effects. However, as reviewed earlier, notable exceptions to this trend exist in which a significant negative effect of feedback was found. Kantner and Lindsay (2010) found a medium-sized effect (Cohen, 1992), while Hicks and Starns (2014) found a small effect. Regardless, criterion noise does represent a potential mechanism for deleterious feedback effects. Moreover, there is a long-standing argumentation in the signal detection literature suggesting that feedback may disrupt recognition processes via added criterion variability (e.g., Clark & Greenberg, 1971; Schoeffler, 1965). Wickelgren (1968) reiterated the importance of considering criteria as having variances that must be considered when comparing different recognition test contexts (e.g., “old/new” recognition versus rating scales).

Overview of Current Study

The goal of this study is to evaluate the idea that accuracy feedback on a recognition test is a source of criterion noise. According to ND-TSD (Benjamin et al., 2009), criteria can vary from trial to trial and criterion noise is created when criteria are maintained and updated (see also Schoeffler, 1965). Traditional SDT (Green & Swets, 1966; Macmillan & Creelman, 2005) assumes that criterion variance is zero. Thus, these theories make competing predictions about the effect of feedback on recognition discriminability. Specifically, ND-TSD predicts that because the aim of feedback is to help participants update their response criterion to optimize performance, an ironic consequence is that this constant adjustment of criterion in response to the feedback actually hurts performance by introducing criterion noise. In contrast, traditional SDT (Green & Swets, 1966; Macmillan & Creelman, 2005) would predict either a null effect of feedback on discriminability because participants are unable to appropriately adjust their criterion or that feedback would provide a benefit to recognition performance by allowing participants to more optimally set their criterion (assuming they may begin nonoptimally). Additionally, the hypothesis that multiple sources of criterion noise can have additive, detrimental effects on recognition discriminability is tested here. That is, I assessed whether criterion noise can be created with other test manipulations (e.g., rating scales, color cues) and whether feedback would further decrease performance beyond these manipulations. Additionally, I collected RT data in both experiments to examine whether a negative impact of feedback on discriminability might be attributed to a speed-accuracy tradeoff.

In Experiment 1, participants completed a recognition test in which both feedback and the length of the rating scale used to make a memory judgment (2 or 8 point) was manipulated. A predicted source of criterion noise in recognition is the length of rating scales used to make memory decisions (see also Wickelgren, 1968). Benjamin et al. (2013) found worse discriminability for tests in which 8 point or 4 point confidence scales were used as compared to simple “yes/no” responses. If maintaining multiple criteria creates criterion noise, then trying to update and optimize multiple response criteria in response to accuracy feedback could create additional criterion noise. RTs were predicted to be slower on the tests with feedback.

In Experiment 2, I manipulated the presence of feedback and color cues indicative of memory strength similar to both Hicks and Starns (2014) and Verde and Rotello (2007). However, both of these earlier studies manipulated feedback across experiments whereas here feedback is varied within-subjects in the same experiment. Additionally, both of these studies presented test items in varying sizes of strength blocks (strong or weak), whereas in this experiment test items were randomly presented. Hicks and Starns (2014) do have one condition where test items were presented randomly, but feedback was not compared to a no-feedback condition in this random test sequence. Here, both the feedback and color cues are crossed in a factorial design with the hypothesis that presenting test items randomly rather than in blocks would increase the likelihood of observing the effects of criterion noise because it would be more difficult to maintain and update criteria on an essentially trial by trial basis rather than when a block of a particular type of item is encountered. As mentioned earlier, ND-TSD predicts that color cues should actually reduce criterion noise (Benjamin et al., 2009). However, based on the work of Hicks and Starns (2014) and some pilot data, color cues were hypothesized to decrease recognition performance. Reaction times were predicted to be slower in the presence of both feedback and color cues.

Experiment 1

Participants

One hundred and twenty-five undergraduate Psychology students from Louisiana State University participated in this experiment to fulfill a partial course requirement or for extra credit.

Materials

Five hundred and sixty unique words with the following characteristics were randomly selected from the MRC psycholinguistic database (http:www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm): concreteness, familiarity, and imageability ratings all between 200 and 600 on scales ranging from 100 to 700, Kučera-Francis written frequency between 10 and 800, and word length between five and nine letters. Four sets of 140 items were used to create study-test lists in four separate computer programs, all of which were created using E-Prime software (Psychology Software Tools, Pittsburgh, PA). These four sets of items were equated for frequency, concreteness, and imageability. For each program, 70 words were randomly chosen to be studied and the remaining 70 words were used as new items at test. This assignment of words to act either as studied targets or new lures at test was counterbalanced. Both the presentation of studied and test items was randomized anew for each participant.

Design

The design was a 2 (feedback: present or absent) × 2 (rating scale length: 2 or 8 point) mixed factorial, with feedback manipulated within-subjects and rating scale length manipulated between-subjects.

Procedure

Participants were tested in groups of 1-3 people and completed two study-test cycles. For the encoding phase, participants were instructed that they would study a list of words for a later memory test. The studied words were presented individually for a duration of 2 seconds each followed by a blank 250 ms ISI. Five primacy and recency buffers were included at the beginning and end of each study list. Immediately following the study list, participants were given instructions for the recognition test. Participants were told that they will be presented with a mixture of studied and non-studied words.

Participants in the 2 point rating scale group were asked to make “old/new” decisions for each test item by pressing the “/” key for “old” and the “z” key for “new” responses. For the participants in the 8 point rating scale group, each test item appeared on the screen above an 8 point confidence rating scale ranging from 1 (“sure new”) to 8 (“sure old”). Participants were asked to press the appropriate number key on the keyboard that corresponded to their level of confidence and were encouraged to use the entire range of responses across the test.

On tests that included feedback, immediately following each response participants saw a screen for 1 second informing them of the status of the word they just judged. If the word they just made a response to was an old word, they saw the message “Studied!” appear in green. If the word they made a response to was a new word, participants saw the message “Not Studied!” in red. For the other recognition test that participants completed, no feedback was given. The order in which participants received feedback (first or second test) was counterbalanced across subjects.

Results

The use of both a 2 and 8 point rating scale presents somewhat of an issue when trying to compare recognition discriminability for these groups. The measure d_a can be calculated when there is more than one point in z-ROC space, which is the case for the 8 point but not the 2 point (yes/no) scale because there is no slope for a singular point in z-ROC space (Macmillan & Creelman, 2005). Thus, d_a was calculated for both the feedback and no feedback conditions in the 8 point scale group and the average slopes derived from these conditions were used to calculate d_a for the 2 point conditions. That is, the slope calculated for the 8 point scale when feedback was absent was used as a slope estimate for the 2 point condition in which there was also no feedback. The same procedure was used to estimate a slope for the 2 point condition when feedback was given from the 8 point rating scale in which feedback was also present.

Additionally, to ensure that the results were not specific to this particular metric of discriminability, A_z was also calculated for each participant. In the 8 point condition, d_a was used to calculate A_z and in the 2 point condition, *d'* was substituted for d_a (Verde, Macmillan, & Rotello, 2006).

Seven participants were excluded from the analyses because they were at or below chance performance in one or both of the recognition tests. Thus, the final sample size for the analyses was 118 participants.

Discriminability

Discriminability was analyzed by submitting the d_a measures to a 2 (feedback) × 2 (response type) × 2 (test order) mixed-factorial ANOVA. Table 2 displays the results below.

Table 2. Group recognition data from Experiment 1 with standard error in parentheses.

	2 Point Ratings		8 Point Ratings	
	No Feedback	Feedback	No Feedback	Feedback
HR	.71 (.02)	.71 (.01)	.72 (.02)	.72 (.01)
FAR	.26 (.02)	.37 (.02)	.28 (.02)	.31 (.02)
d_a	1.30 (.08)	.96 (.07)	1.26 (.08)	1.09 (.06)
A_z	.81 (.01)	.73 (.01)	.79 (.01)	.77 (.01)

There was a main effect of feedback such that overall discriminability was lower in the feedback condition, $F(1, 116) = 29.60$, $MSE = .128$, $p < .001$, $\eta_p^2 = .206$. There was a trend for feedback to interact with response type, although this was not significant, $F(1, 116) = 3.31$, $p = .072$, $\eta_p^2 = .028$. Pairwise comparisons showed that feedback had a somewhat larger negative impact in the 2-point condition, $t(57) = 4.94$, $p < .001$, Cohen's d = .64, than in the 8-point condition, $t(59) = 2.81$, $p = .007$, Cohen's d = .36. This result is depicted visually below in Figure 3. There was no main effect of test order, $F(1, 116) = 1.73$, $p = .19$, or response type, $F(1, 116) < 1$, $p = .61$. Additionally, the interactions between feedback and test order, response type and test order, as well as the three-way interaction were not significant, *Fs* < 1.

A_z measures were also examined with a 2 (feedback) × 2 (response type) × 2 (test order) mixed-factorial ANOVA. A significant main effect of feedback was found such that feedback lowered discriminability, $F(1, 116) = 32.34$, $MSE = .004$, $p < .001$, $\eta_p^2 = .221$. However, this main effect was qualified by a significant feedback by response type interaction, $F(1, 116) = 9.26$, $MSE = .004$, $p = .003$, $\eta_p^2 = .075$.

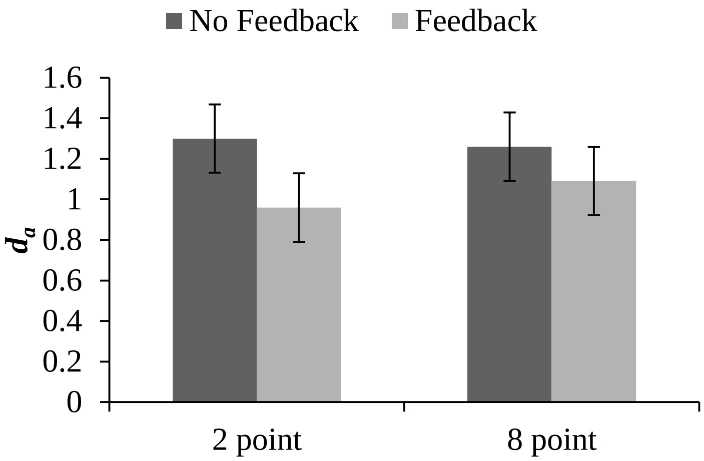


Figure 3. Discriminability measure d_a in Experiment 1. Error bars represent 95% CIs as recommended by Masson & Loftus (2003) for mixed-factorial designs.

Post-hoc tests showed that the negative effect of feedback was larger in the 2-point condition, $t(57) = 6.08$, $p < .001$, Cohen’s d = .80, than in the 8-point condition, $t(59) = 1.83$, $p = .073$, Cohen’s d = .24, and this result is shown below in Figure 4.

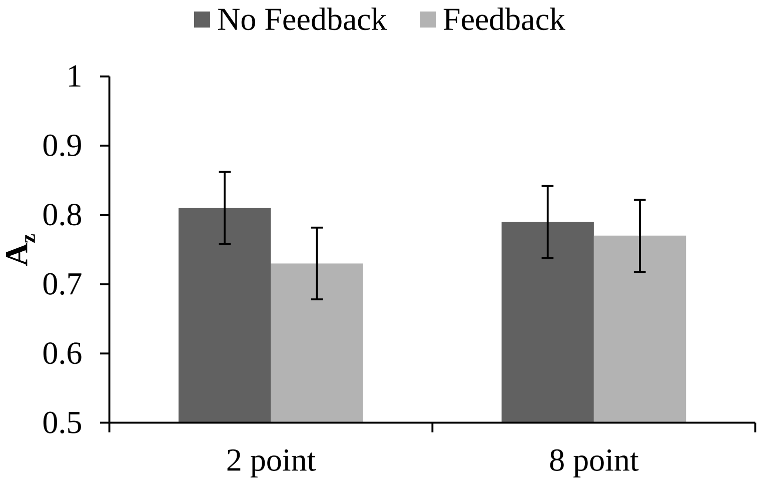


Figure 4. Discriminability measure A_z in Experiment 1. Error bars represent 95% CIs as recommended by Masson & Loftus (2003) for mixed-factorial designs.

The main effects of test order, *F*(1, 116) = 1.34, *p* = .25, and response type, *F*(1, 116) = .80, *p* = .37, were both not significant. Again there were no interactions between feedback and test order, response type and test order, and the three-way interaction was not significant, *Fs* < 1.

Reaction Times

The group RT data is displayed on the following page in Table 3. Median RTs for hits and correct rejections were also examined separately for the 2-point and 8-point conditions with 2 (feedback) × 2 (test order) mixed-factorial ANOVAs. Reaction times in the 2-point group were initially trimmed if they were faster than 300ms or slower than 2000ms. For the 8-point group, RTs faster than 500ms or slower than 4600ms were trimmed prior to analysis[1].

Table 3. Average median RTs from Experiment 1 with standard error in parentheses.

	2 Point				8 Point			
	FB First		FB Second		FB First		FB Second	
	No FB	FB	No FB	FB	No FB	FB	No FB	FB
Hits	906.28 (48.60)	903.25 (51.50)	921.42 (53.92)	861.44 (57.13)	1393.10 (49.38)	1606.55 (52.32)	1591.29 (51.05)	1407.53 (54.09)
Correct Rejections	1036.02 (48.83)	992.08 (55.24)	979.35 (54.17)	945.19 (61.28)	1510.73 (49.61)	1654.39 (56.13)	1770.67 (51.29)	1564.26 (58.03)

For the 2-point condition, there was a tendency for hits to have faster RTs in the feedback condition though this was not significant, *F*(1, 56) = 3.22, *p* = .08. There was no effect of test order, *F*(1, 56) = .12, *p* = .73, and the interaction was also not significant, *F*(1, 56) = 2.63, *p* = .11. For the 8-point condition, there was no effect of feedback or test order on RTs for hits, *Fs* < 1. The interaction between feedback and test was significant, *F*(1, 58) = 19.45, *MSE* = 60763.57, *p* < .001, η_p^2 = .258, indicating that RTs for hits were faster in the no-feedback condition when the test with feedback was completed first but that RTs were faster in the feedback condition when the test including feedback was completed second. That is, RTs simply got faster across the testing session.

Regarding RTs for correct rejections in the 2-point group, there was an effect of feedback such that RTs were faster when feedback was present, *F*(1, 58) = 5.03, *MSE* = 8692.32, *p* = .03, η_p^2 = .082. Test order did not affect RTs, *F*(1, 56) = 1.72, *p* = .20, and the interaction was also not significant, *F*(1, 56) = .08, *p* = .78. For the 8-point group, there was no main effect of feedback on RTs for correct rejections, *F*(1, 58) = .77, *p* = .39. Additionally, the main effect of test order was not significant, *F*(1, 58) = .86, *p* = .36. There was a significant interaction between feedback and test order, *F*(1, 58) = 23.87, *MSE* = 38463.95, *p* < .001, η_p^2 = .292, indicating that RTs were faster on the test without feedback when that test was completed last and RTs were faster on the test with feedback when that test was completed last.

Endnote

[1] Reaction times were also analyzed by examining log-transformed median values of hits and correct rejections. RTs were trimmed at the lower end for the 2 and 8-point groups at 300 and 500ms, respectively. Slower RTs were trimmed if they were 2.5 SDs above an individual’s average RT. This trimming procedure removed only 3% of cases and the subsequent analyses yielded the same results as those reported above.

Discussion

To examine whether feedback introduces criterion noise into recognition memory decisions, participants completed two study-test cycles in which accuracy feedback was present on one test but not the other. Additionally, the type of recognition response was manipulated such that participants made either “yes/no” decisions or gave confidence ratings. As predicted, there was an overall negative effect of feedback such that discriminability was lower when feedback was present. Discriminability was equal between the “yes/no” and confidence ratings groups, which is somewhat problematic for ND-TSD as it predicts that memory should be better when participants are making simple “yes/no” decisions. Additionally, the ordinal interaction between feedback and response is difficult to explain using ND-TSD (Benjamin et al., 2009). Feedback lowered discriminability more in the “yes/no” condition than in the confidence ratings condition. ND-TSD predicts that feedback might interact with rating scale length, though it predicts that feedback should be more harmful to recognition discriminability in the confidence ratings condition. As noted earlier, Benjamin et al. (2013) found that longer rating scale length was associated with lower recognition discriminability, which they interpreted as indicative of criterion noise because making confidence decisions requires the use of multiple criteria. However, unlike Benjamin et al. (2013), we manipulated rating scale length between subjects rather than within-subjects. Hence, our failure to find that feedback lowers discriminability more for rating scale decisions might reflect a couple of possible outcomes: this study lacked sufficient statistical power to detect an effect of rating scale length or that the additive effects of different sources of criterion noise were small or negligible. A detailed discussion of these possible outcomes is deferred to the General Discussion section.

Regarding the RT data, there was a very mixed bag of results. Participants in the 2-point condition were slightly faster in the feedback condition which replicates Kantner and Lindsay (Exp. 2; 2010) who also found slightly faster RTs in a feedback condition, albeit in the context of making 6-point confidence ratings. However, RTs in the 8-point group were not impacted much by the presence of feedback. Rather, the results indicated that they just got faster as the testing session progressed. This also replicates Kantner and Lindsay (Exps. 3 & 4; 2010) who found in two of their experiments that feedback had a null effect on RTs but that RTs generally got faster across testing blocks. The RT data also speak against a simple speed-accuracy tradeoff in regards to why feedback lowers discriminability. For the 2-point group, RTs were only significantly faster for correct rejections when feedback was present, whereas we might expect a speed-accuracy tradeoff to affect both types of correct decisions. Additionally, discriminability was lowered by the feedback in the 8-point condition, though not to the same degree as in the 2-point condition. Regardless, feedback lowered discriminability in 8-point condition but there were no significant RT differences between feedback and no feedback conditions.

Experiment 2

As mentioned earlier, Verde and Rotello (2007) found numerically worse discriminability using a strength manipulation at encoding and by presenting test items in blocks according to their strength. Hicks and Starns (2014) found a significant negative effect of feedback on discriminability using this procedure and also found that the presence of color cues denoting memory strength lowered discriminability. Thus, the aim of this experiment was to assess how feedback and color cues indicative of memory strength (i.e. strong or weak) would affect recognition discriminability when test items are presented randomly rather than in blocks. Because a random test condition requires frequent criterion shifts, I hypothesized that this test format would create more criterion noise than a blocked test and hence would allow for the potentially negative effects of feedback and color cues to be more readily observable. Both feedback and color cues were predicted to lower discriminability and these two variables were predicted to have additive negative effects.

Participants

One hundred and twenty-three undergraduate Psychology students from Louisiana State University participated in this experiment to fulfill a partial course requirement or for extra credit.

Materials

Four hundred unique words with the same properties as the stimuli used in Experiment 1 were randomly selected from the MCR psycholinguistic database. Four programs were created using E-Prime software and 100 words were randomly assigned as stimuli for each of the four programs.

Design

The design was a 2 (feedback: present or absent) × 2 (strength: strong or weak targets) × 2 (color marking: present or absent) × 2 (test order: feedback on first or second test) mixed-factorial, with feedback and strength manipulated within-subjects and color marking at test and test order manipulated between-subjects.

Procedure

Participants were tested individually or in pairs for each experimental session and completed two study-test cycles. Participants were told they would study a list of words for a later memory test and that some words would be presented multiple times. For the encoding procedure, the program randomly selected 40 words from the pool of 100 items, half of which were presented four times (strong targets) and half of which were presented only once (weak targets). Ten filler items were presented at the beginning and end of each encoding to act as primacy and recency buffers. Thus, the encoding phase consisted of 100 presentations which were randomized anew for each participant. Words were presented individually for a 700 ms immediately followed by a blank 100ms ISI. The remaining 40 words served as lures on the test, half of which were assigned to the strong color cue while the remaining 20 were assigned to the weak color cue. These lure stimulus assignments were made by the software even when the color cue was not provided. This aspect of the procedure is identical that used by Hicks & Starns (2014).

After the encoding phase, participants were immediately given test instructions informing them they would take a test composed of studied and non-studied words. For participants in the marked condition, they were informed that test items studied four times would be presented in red font color, words studied once would be presented in green, and that new test items will appear half in red and half green. For participants in the unmarked condition, all test items were presented in black. Participants made an “old” or “new” decision for each test item by pressing the “/” and “z” keys, respectively. In both conditions, test items were randomly presented. For the programs that included feedback, participants were additionally informed that feedback will appear on the test. Specifically, they were told that they would see the message “***ERROR***” when they made an incorrect decision. This feedback screen lasted for 1200 ms. The order in which participants received feedback at test was counterbalanced across participants.

Results

Nineteen participants were excluded from the analyses because they were at or below chance performance in one or both of the recognition tests. Thus, a final sample size of 104 participants was used in the analyses.

Discriminability

Group recognition data are presented below in Table 4. For each recognition test, HR, FAR, and *d’* was calculated separately for strong and weak items.

Table 4. Group recognition data from Experiment 2 with standard error in parentheses.

	Unmarked				Marked			
	No Feedback		Feedback		No Feedback		Feedback	
	Weak	Strong	Weak	Strong	Weak	Strong	Weak	Strong
HR	.67 (.02)	.88 (.02)	.67 (.02)	.86 (.01)	.65 (.02)	.87 (.01)	.67 (.02)	.85 (.01)
FAR	.28 (.02)	.27 (.02)	.26 (.02)	.31 (.02)	.28 (.02)	.26 (.02)	.32 (.02)	.27 (.02)
d’	1.17 (.09)	2.04 (.09)	1.17 (.02)	1.72 (.10)	1.10 (.06)	1.94 (.08)	1.01 (.08)	1.79 (.09)

Discriminability was examined by analyzing *d’* with a 2 (feedback) × 2 (color marking) × 2 (strength) × 2 (test order) mixed-factorial ANOVA. A significant main effect of strength indicated that memory was better for strong than weak items, *F*(1, 102) = 186.60, *MSE* = .313, *p* < .001, η_p^2 = .651. Critically, there was a main effect of feedback, *F*(1, 102) = 8.00, *MSE* = .258, *p* = .006, η_p^2 = .074, whereby discriminability was lower when feedback was present. The main effects of test order, *F*(1, 102) = 2.28, *p* = .13, and color marking, *F*(1, 102) = .94, *p* = .33, were not significant. There was no significant interaction between strength and test order, *F*(1, 102) = .60, *p* = .44, or between strength and color marking, *F*(1, 102) = .75, *p* = .39. However, the there was a significant three-way interaction between these variables, *F*(1, 102) = 7.39, *MSE* = .313, *p* = .008, η_p^2 = .069. When feedback was first, memory for strong items was higher in the unmarked (*M* = 1.99, *SE* = .10) relative to the marked condition (*M* = 1.66, *SE* = .11). However, when feedback was second, memory for strong items was higher in the marked (*M* = 2.04, *SE* = .10) versus the unmarked condition (*M* = 1.76, *SE* = .10).

Feedback did not significantly interact with test order, *F*(1, 102) = 3.74, *p* = .056, though there was a trend for feedback to lower discriminability more when the test including feedback was completed after the test with no feedback. Feedback also did not interact with color marking, *F*(1, 102) = .18, *p* = .68. However, there was an unexpected interaction between feedback and strength, *F*(1, 102) = 4.00, *MSE* = .212, *p* = .049, η_p^2 = .038. Post-hoc tests revealed that feedback significantly lowered discriminability for strong items, *t*(103) = 3.50, *p* = .001, Cohen's d = .34, but not for weak items, *t*(103) = .70, *p* = .49, Cohen's d = .07 (Figure 4). The interaction between test order and color marking was also not significant, *F*(1, 102) = 3.92, *p* = .051, however discriminability was nominally higher in the marked (*M* = 1.58, *SE* = .07) versus the unmarked (*M* = 1.51, *SE* = .08) condition when the test with feedback was completed second but was lower in marked (*M* = 1.32, *SE* = .08) versus the unmarked (*M* = 1.54, *SE* = .08) condition when the test including feedback was completed first. There was no three-way interaction between feedback, strength, and test order, *F*(1, 102) = 2.67, *p* = .11. Additionally, the three-way interaction between feedback, strength, and color marking was also not significant, *F*(1, 102) =

2.37, p = .13. Lastly, the four-way interaction between all variables was not significant, F(1, 102) < .001, p = .99.

Reaction Times

The group RT data is displayed below in Table 5.

Table 5. Average median RTs from Experiment 2 with standard error in parentheses.

	Unmarked				Marked			
	FB First		FB Second		FB First		FB Second	
	No FB	FB	No FB	FB	No FB	FB	No FB	FB
Hits								
Strong	745.96 (26.17)	812.85 (31.91)	730.18 (26.69)	801.18 (32.54)	840.02 (27.24)	954.75 (33.21)	839.86 (24.78)	886.88 (30.21)
Weak	836.79 (37.80)	896.71 (35.51)	798.20 (38.55)	879.92 (36.22)	897.58 (39.34)	1031.25 (36.97)	1022.47 (35.79)	969.31 (33.63)
Correct Rejections								
Strong	903.00 (33.86)	982.83 (32.24)	907.70 (34.53)	943.34 (32.88)	1045.75 (35.24)	1136.10 (33.56)	1055.81 (32.06)	1010.52 (30.53)
Weak	927.94 (29.50)	946.96 (30.63)	899.94 (30.09)	943.42 (31.24)	949.00 (30.71)	1081.85 (31.88)	1020.48 (27.93)	1036.62 (29.01)

Median RTs for hits and correct rejections were separately analyzed for each strength condition with a 2 (color marking) × 2 (feedback) × 2 (test order) mixed-factorial ANOVA. RTs faster than 300ms or slower than 2000ms were trimmed prior to analysis[1].

Regarding hits for strong items, there was a main effect of feedback such that RTs were faster when feedback was absent, F(1, 102) = 32.48, MSE = 8940.00, p < .001, η_p^2 = .245.

Additionally, a main effect of color making revealed that RTs were faster in the unmarked condition, *F*(1, 102) = 17.03, *MSE* = 35335.58, $p < .001$, $\eta_p^2 = .146$. The main effect of test order was not significant, *F*(1, 102) = .83, *p* = .36. Lastly, none of the interactions was significant: feedback × color marking, *F*(1, 102) = .21, *p* = .65; feedback × test order, *F*(1, 102) = 1.46, *p* = .23; marking × test order, *F*(1, 102) = .15, *p* = .70; feedback × marking × test order, *F*(1, 102) = 1.87, *p* = .18.

For weak items, there was a main effect of feedback on hit RTs, indicating that RTs were faster when feedback was absent, *F*(1, 102) = 9.28, *MSE* = 17199.37, *p* = .003, $\eta_p^2 = .085$. A main effect of marking was obtained, whereby RTs were faster in the unmarked condition, *F*(1, 102) = 15.89, *MSE* = 52735.97, $p < .001$, $\eta_p^2 = .137$. There was no main effect of test order, *F*(1, 102) = .004, *p* = .95. However, test order did interact with feedback, *F*(1, 102) = 5.12, *MSE* = 17199.37, *p* = .026, $\eta_p^2 = .049$. When the test with feedback was completed first, RTs were faster on the test without feedback but when the test with feedback was completed last there was no difference between feedback and no feedback RTs. Additionally, the three way interaction between feedback, marking, and test order was significant, *F*(1, 102) = 8.18, *MSE* = 17199.37, *p* = .005, $\eta_p^2 = .076$, whereby in the unmarked condition RTs were always faster when feedback was absent but in the marked condition RTs were faster on whichever test (with feedback or without feedback) was completed last. There was no significant interaction between feedback and marking, *F*(1, 102) = .70, *p* = .40, or between marking and test order, *F*(1, 102) = .86, *p* = .36.

Moving on to RTs for correct rejections of strong items, there was a main effect of feedback, *F*(1, 102) = 6.23, *MSE* = 13376.37, *p* = .014, $\eta_p^2 = .059$, such that RTs were faster on the tests without feedback. There was a significant main effect of marking, *F*(1, 102) = 19.45, *MSE* = 43460.17, *p* < .001, η_p^2 = .163, indicating that RTs were faster in the unmarked condition. A significant feedback by test order interaction, *F*(1, 102) = 6.23, *MSE* = 13376.37, *p* = .014, η_p^2 = .059, revealed that RTs were faster on the test without feedback when it was completed second and that there was no difference in feedback and no feedback RTs when the test with feedback was completed second. There was no significant interaction between feedback and marking, *F*(1, 102) = 1.20, *p* = .28, or between marking and test order, *F*(1, 102) = .49, *p* = .49. The three way interaction was not significant, *F*(1, 102) = 2.02, *p* = .16.

For the RTs of correct rejections of weak items, there was a main effect of feedback, *F*(1, 102) = 15.09, *MSE* = 9588.08, *p* < .001, η_p^2 = .131, indicating that RTs were slower when feedback was present. A main effect of color marking revealed that RTs were slower in the marked condition, *F*(1, 102) = 19.45, *MSE* = 37473.99, *p* = .001, η_p^2 = .106. There was no effect of test order, *F*(1, 102) = .002, *p* = .96. Feedback did not interact with test order, *F*(1, 102) = 2.87, *p* = .09, nor did it interact with marking, *F*(1, 102) = 2.52, *p* = .12. However, the three-way interaction between these variables was significant, *F*(1, 102) = 6.72, *MSE* = 9588.08, *p* = .011, η_p^2 = .063, indicating that RTs in the unmarked group did not vary by test order or feedback but in the marked condition RTs were faster when feedback was absent but only when the test including feedback was completed first. Lastly, the interaction between color marking and test order was not significant, *F*(1, 102) = .29, *p* = .59.

Endnote

[1] Reaction times were also analyzed by examining log-transformed median values of hits and correct rejections. RTs were trimmed if they were faster than 300ms or 2.5 SDs above a person’s mean RT. This trimming procedure removed only 3% trials. Subsequent analyses resulted in similar results to those above with the exception that for correct rejections, there is no main effect of feedback for strong or weak items.

Discussion

In this experiment, participants completed two recognition tests in which feedback at test was either present or absent. Additionally, we manipulated the memory strength of items and the presence of color cues at test indicating memory strength (strong or weak). Overall, feedback was found to lower discriminability as predicted by ND-TSD (Benjamin et al., 2009). In particular, the presence of feedback significantly lowered discriminability for strong items whereas discriminability for weak items was relatively unaffected. These results provide mixed support for ND-TSD (Benjamin et al., 2009) as the theory would predict that feedback should lower discriminability via introducing criterion noise into the recognition decision. However, it is unclear why feedback for strong items would produce more criterion noise than feedback for weak items. Additionally, there was no effect of color cues at test, which is consistent with the ND-TSD prediction that giving people a cue to better recognize differences in target distributions should reduce criterion noise. The overall result suggests that different, multiple sources of criterion noise may not have additive effects, though there was a trend for discriminability to be lower in the marked condition relative to the unmarked condition when the test with feedback was completed first. However, the nature of this marginally significant interaction was disordinal in that participants actually had slightly higher discriminability in the marked condition when the test including feedback was completed after the test with no feedback. If color cues are a source of criterion noise, then they should be an impediment to discriminability regardless of the test order. One potential explanation here is simply that there were practice effects resulting from using the color cues on the first test. A more elaborate discussion of these findings is reserved for the next section.

Regarding RTs, both feedback and marking generally led to longer RTs, suggesting that participants were in fact trying to use the color marking and incorporating the feedback. Additionally, participant RTs generally got faster across the testing session, which likely reflects a practice effect. Taken together, the RT data suggest that the deleterious impact of feedback is not simply a speed-accuracy trade off. If this were the case we would have expected feedback to speed responses only for strong items since feedback only lowered discriminability for strong items, but the results show that feedback speeded responses regardless of strength.

General Discussion

The primary goal of this study was to test the hypothesis that feedback lowers recognition discriminability by means of introducing criterion noise into the decision process. According to ND-TSD (Benjamin et al., 2009) criterion variance is not fixed (i.e. zero) as traditionally assumed by SDT (Green & Swets, 1966; Macmillan & Creelman, 2005) but instead can randomly vary on each trial. The consequence of this criterion variability is that a strain on memory (i.e., discriminability) is experienced when having to maintain a criterion (e.g. Benjamin et al., 2013) or when attempting to update that criterion. Thus, two general predictions can be derived from ND-TSD: any test manipulation that encourages a person to update or adjust their criterion or any circumstance in which a person is required to hold multiple criteria will lower recognition discriminability. Corrective feedback given on recognition tests is meant to help improve performance by allowing participants to make favorable criterion adjustments, but this constant updating of criteria might actually introduce criterion noise and thus lower discriminability.

To investigate the possibility that feedback produces a real and deleterious effect on recognition discriminability, feedback was manipulated within-subjects in two experiments in which we also manipulated rating scale length (Experiment 1) and memory strength and color cues at test indicating strength (Experiment 2). There are four primary results from these experiments which are summarized here. First, in both experiments participants' overall discriminability was lower when feedback was present rather than absent at test. Second, feedback in Experiment 1 produced a larger negative effect when participants made simple "yes/no" decisions to test items as opposed when they made confidence ratings. Third, feedback in Experiment 2 produced lower discriminability for strong items but exerted only a negligible effect on weak items. Fourth, the presence of color cues in Experiment 2 indicative of memory strength had no effect on discriminability.

Regarding the first point, this overall negative effect of feedback replicates the work of Hicks and Starns (2014) who found feedback lowered discriminability in various blocked test conditions when target strength was manipulated. The present results also extended this negative effect of feedback to a random test condition in this same strength-based paradigm (cf. Experiment 2). In addition, feedback in Experiment 1 generally lowered discriminability regardless of response type, although this effect was smaller in the context of confidence ratings. The negative influence of feedback is also generally consistent with the effect size estimates discussed earlier in Table 1. Thus, no hint of a positive influence of feedback was found anywhere in the various conditions of these experiments. When feedback did exert a significant influence, it was a negative one.

Regarding the second point, although the ordinal interaction between feedback and rating scale length was not predicted a priori, it is worth speculating about this finding as it has implications for ND-TSD. Specifically, this theory predicts that more criterion noise should be created as the length of the response rating scale increases (Benjamin et al., 2009; 2013). There are several reasons why we might have failed to find this effect of rating scale length. One possibility is that because of our between-subjects manipulation of rating scale length, we simply lacked statistical power to detect this effect. However, when comparing the 2-point versus the 8-point conditions collapsed across feedback conditions, discriminability was slightly better for the rating scale condition, rather than worse. Experiment 1's outcome therefore replicates Lindsay and Kantner (Exp. 3, 2011) as well as Koen and Yonelinas (2011) who found no differences in discriminability when comparing "yes/no" versus confidence rating decisions. Benjamin et al.

(2013) found a small difference between these types of recognition decisions, though they are also the only study to manipulate response type within-subjects. So it is possible that the influence of response type is so small that it is best studied in a within-subjects context. Alternatively it could be that there is only a negligible amount criterion noise introduced into recognition decisions when participants make confidence ratings (Kellen et al., 2012). Regardless of the interaction produced in Experiment 1, those results suggest that the presence of feedback had much more influence on discriminability than did the type of response context. Further work should focus on replicating the influence of rating scales versus “yes/no” decisions.

Another possible interpretation for feedback exhibiting a larger negative effect in the 2-point versus the 8-point rating decision is that the range in which criterion can vary is smaller in the 8-point condition. In the 8-point rating condition, participants are assumed to set a criterion for each confidence boundary. One consequence of this might be that the criterion variance for any particular level of confidence is artificially restricted in that even when a criterion is variable it does not cross the confidence boundary (i.e. criterion) immediately above or below it. Put another way, even when criterion variance is present for a given confidence level, say dividing a “5” from a “6”, the criterion for that level can only vary within the boundaries of the other confidence levels, as opposed to crossing into the boundary between a “4” and “5” or between a “6” and “7”. A related idea comes from Mueller and Weidemann (2008) who suggest that for each confidence level, criteria can be represented as overlapping normal distributions on an axis of perceptual evidence (e.g. memory strength). For a given amount of evidence, there is a high probability that the criterion corresponding to that particular level of confidence will be selected, though occasionally a criterion is sampled from one of the overlapping criterion distributions. Conversely, when a participant makes “yes/no” decisions, they use only a single criterion. In this case, because there are no other levels of confidence, the criterion is free to vary across a larger range of memory strength because it never encroaches upon any other decision boundary. Admittedly, this explanation is post-hoc and, in its present state, ND-TSD allows for criterion variability even across confidence boundaries, in fact predicting more noise in this response context (Benjamin et al., 2009). Concerning the effect of color cues indicative of the memory strength of test items in Experiment 2, the presence of these cues did not affect discriminability. Hicks and Starns (2014) found that these types of test cues lowered discriminability in their strength-blocked tests of varying sizes. In contrast, color cues did not have a significant effect on discriminability here. This finding replicates Stretch and Wixted (Exp. 3, 1998), though it should be noted that in their study and in the present work discriminability was numerically larger in the unmarked conditions, in the direction of color marking producing a negative effect. Thus, the effect size for this factor is small, as indicated by Hicks and Starns. They found a significant effect in statistical comparisons comprised of over 200 participants in each of their first two experiments.

Prior to conducting this study, I collected pilot data in an experiment identical to Experiment 2, except that feedback was manipulated between-subjects and color cues were manipulated within-subjects (Appendix A). In this pilot study, the presence of color cues significantly lowered discriminability whereas there was only a trend for feedback to harm discriminability. The contrast between the present work and this pilot data illustrates the point that the effects of both feedback and marking are small and often hard to detect depending on the type of experimental manipulation used. Again, another reason for the finding that feedback but not color cues harm discriminability could be that different sources of criterion noise do not have an additive effect. If a random test condition that requires essentially trial-by-trial updating of a response criterion creates more criterion noise than color cues, then the present study may be limited in its ability to detect a significant source of criterion noise above and beyond that created by a random test condition. However, in the present study there is no manipulation of test composition. Thus a direct comparison of discriminability in an unmarked condition between a random and blocked test cannot be made here.

It is important to note that according to ND-TSD (Benjamin et al., 2009), when color cues denoting memory strength are provided at test, the amount of criterion noise should actually be diminished. Specifically, Benjamin et al. (2009) predicted that when participants can treat multiple classes of test items (e.g. weak and strong) differently, criterion variance should be reduced. Again, the purpose of presenting the color cues at test is to get participants to treat strong and weak test items differently. Thus, the results here do not support one of the explicit predictions of ND-TSD. It remains to be seen whether a cleaner manipulation of making people aware of strength differences might produce results in line with the ND-TSD prediction. For example, one could rely on a more extreme manipulation of strength by adding repetitions to the strong items, or perhaps with a more traditional shallow/deep manipulation based on the levels of processing framework (Craik & Lockhart, 1972).

Unexpectedly, the results from Experiment 2 showed that feedback exerted a negative influence on the discriminability of strong but not weak items. In the color marked condition, discriminability for weak items was slightly lower in the feedback condition. However, in the unmarked condition, discriminability for weak items was identical in both feedback conditions. It is rather unclear why feedback would differentially affect strong but not weak items. One possibility is that feedback may only have a beneficial, or at least benign, effect on discriminability when the recognition decision is particularly difficult to make. That is, when a person is required to make a more fine-grained decision (e.g. discriminating a weak target from a lure), feedback may not exert a negative effect because the decision is already moderately noisy. Conversely, there may be a smaller amount of pre-existing noise for easier recognition decisions (e.g. discriminating a strong target from a lure). Thus, feedback might exert an adverse effect only on strong items because it introduces criterion noise into a decision that typically has a low amount of noise. This explanation is consistent with the previously suggested idea that there may be an upper limit to the amount of noise (criterion noise or stimulus noise) that can impact recognition decisions.

Related Accounts of Criterion Variability

Although the idea and significance of criterion variability is a relatively recent matter of debate for memory researchers, the idea of criterion variance has longstanding roots in the psychophysical literature (e.g. Tanner, 1961). For example, Schoeffler (1965) developed an intricate model of learning in the context of SDT in which feedback is explicitly considered as a factor of interest. According to Schoeffler (1965), when feedback is present at test, subject's knowledge of their performance, particularly when they make an error, will lead them to adjust their criterion in order to more appropriately respond on the next trial. However, in the absence of feedback, subjects do not adjust their criterion (i.e. criterion variance is zero) and the results of his modeling predicts that discriminability should be better when no feedback is present. One subtle contrast here with ND-TSD (Benjamin et al., 2009) is that ND-TSD posits that criterion noise exists even without feedback, as simply using and maintaining a criterion introduces noise. Schoeffler's notion that participants use feedback to adjust their criterion was echoed by Kac (1969) who proposed that people will only adjust their criterion in response to error feedback. That is, when a person is made aware that they made a correct decision, they have no motivation to adopt a more liberal or conservative criterion However, when informed of an incorrect decision, people will adjust slightly in the appropriate direction as to minimize the risk of making that error again (i.e. liberal shift for a miss and a conservative shift for a false alarm). However, this assumes that observers equally value errors of omission and commission which may not always be the case (e.g. payoff manipulations).

Another theory of criterion variability comes from Mueller and Weidemann (2008) who propose a decision noise model (DNM) extension of SDT. The authors make the distinction between “stimulus noise” and “decision noise”, with the latter reflecting an analogous concept to criterion noise. According to the DNM, decision noise is essentially a mismapping between an internal state (e.g. familiarity) and an external response (e.g. confidence rating). Similar to ND-TSD, the DNM asserts that on a particular test trial, a criterion value is selected from a distribution of criteria. Additionally, both of these theories predict that recognition discriminability should be better when making “yes/no” decisions versus confidence ratings, although the mechanisms by which that happen are different. ND-TSD predicts that discriminability is lower for rating scale confidence decisions because each level of confidence must be maintained which places a burden on the observer to remember where each criterion maps onto varying levels of memory strength (Benjamin et al., 2009). Alternatively, DNM posits that discriminability is worse for confidence ratings because some of the criterion distributions overlap at any given point of memory strength. That is, for a given familiarity value, an observer may sample from multiple overlapping criterion distributions (i.e. confidence levels) which leads to suboptimal performance (Mueller & Weidemann, 2008).

Summary and Future Directions

Taken together, the results from both experiments provide an empirical point of departure from the extant work on feedback in recognition (Kantner & Lindsay, 2010) in that the data from this study indicate that feedback has a real, albeit small, deleterious impact on discriminability. The primary basis for re-examining this abundance of null results stemmed from predictions based on ND-TSD (Benjamin et al., 2009) which asserts that manipulations designed to get people to adjust their criterion may create criterion noise which subsequently leads to lower discriminability. Although considering feedback as a source of criterion noise is consistent with the spirit of ND-TSD theory, feedback was not a factor considered by Benjamin and colleagues as a potential source of noise. Moreover, other results from this study suggest that embracing ND-TSD theory in its current state would be premature. For instance, in the first experiment the length of the rating scale had no effect on discriminability. One potential way to explore this idea would be to manipulate both feedback and rating scale length within-subjects. This should enable an effect of rating scale length to emerge, if it exists, and also to observe if an ordinal interaction exists between feedback and rating scale length.

Also, even though color marking in Experiment 2 did not significantly harm recognition discriminability, it also did not help performance as predicted by ND-TSD theory. Assuming that color marking should produce a negative influence, contrary to the ND-TSD predictions but consistent with prior work by Hicks and Starns (2014), in the second experiment there was no additive negative effect of feedback and color cues, again indicating that different sources of criterion noise may not accumulate. However, the ability to observe this may again be limited by the experimental design. In all cases, participants completed a random test which may produce so much criterion variance on its own that layering color cues on top of that is essentially a drop in the bucket. This issue could be addressed by also including a condition where participants completed a blocked test as in Hicks and Starns (2014). Additionally, ND-TSD does not give an account of why there might be more criterion noise for strong versus weak items.

Future examinations of feedback in recognition might also do well to follow the lead of Kellen et al. (2012) who examined criterion noise in a ranking task which the authors assumed to be criterion free. Although the results from the present study provide evidence that feedback has a negative impact on recognition discriminability with criterion variance as a proposed mechanism, whether or not that decrement reflects criterion noise per se is, admittedly, not an entirely resolved issue. Although the current work was motivated by ND-TSD, the primary goal here was to examine whether feedback has a genuine, adverse effect on recognition discriminability and not necessarily to champion ND-TSD. If it is truly criterion noise that is causing an adverse effect of discriminability, that decrement should not exist in a task where participants do not typically exhibit bias such as a two-alternative forced choice (2AFC) task. If feedback lowered discriminability on a 2AFC task, those results would be very problematic for ND-TSD or for any theory that might propose criterion noise as a source of the negative influence of feedback. Alternatively, if feedback does not affect performance in a 2AFC task that would provide some indirect evidence that the locus of the decrement is in fact the criterion.

Conclusions

In closing, the results from this study provide evidence that recognition discriminability can be negatively affected in the presence of feedback. Previously, feedback has been regarded as a potential source of advantageous information that should lead to better discriminability. Some support was found for ND-TSD which predicts that feedback is a source of criterion noise. Additionally, no evidence was found which would suggest that independent sources of criterion noise can have additive effects. Models of memory will need to be able to account for the negative effect of feedback at test and current theories which explicitly state that feedback should help recognition discriminability (e.g. Turner et al., 2011) will need to consider the results from this study in future work.

References

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., . . . Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition, 40*(7), 1016-1030. doi: 10.3758/s13421-012-0204-6

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*(1), 84-115. doi: 10.1037/a0014351

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1601-1608. doi: 10.1037/a0031849

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition, 37*(6), 807-818. doi: 10.3758/MC.37.6.807

Clark, W., & Greenberg, D. B. (1971). Effect of stress, knowledge of results, and proactive inhibition on verbal recognition memory (*d'*) and response criterion (L_X). *Journal of Personality And Social Psychology, 17*(1), 42-47. doi:10.1037/h0030422

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1075-1095. doi: 10.1037/0278-7393.21.5.1075

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. doi: 10.3758/BF03193146

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford England: John Wiley.

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition, 36*(4), 703-715. doi: 10.3758/MC.36.4.703

Han, S., & Dobbins, I. G. (2009). Regulating recognition decisions through incremental reinforcement learning. *Psychonomic Bulletin & Review, 16*(3), 469-474. doi: 10.3758/PBR.16.3.469

Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*, *42*(5), 742-754. doi:10.3758/s13421-014-0397-y

Kac, M. (1969). Some mathematical models in science. *Science, 166*, 695-699.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38*(4), 389-406. doi: 10.3758/MC.38.4.389

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review*, *119*(3), 457-479. doi: 10.1037/a0027727

Lindsay, D. S., & Kantner, J. (2011). A search for the influences of feedback on recognition of music, poetry, and art. In P. Higham & J. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea* (pp. 137-154). Houndmills, UK: Palgrave Macmillan.

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

Masson, M. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203-220. doi:10.1037/h0087426

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*(3), 465-494. doi:10.3758/PBR.15.3.465

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 305-320. doi: 10.1037/0278-7393.33.2.305

Selmeczy, D., & Dobbins, I. G. (2013). Metacognitive awareness and adaptive recognition biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 678-690. doi: 10.1037/a0029469

Schoeffler, M. S. (1965). Theory for psychophysical learning. *Journal of The Acoustical Society Of America, 37*(6), 1124-1133. doi:10.1121/1.1909534

Tanner, W. P. (1961), *Physiological implications of psychophysical data*. Annals of the New York Academy of Sciences, 89: 752–765. doi: 10.1111/j.1749-6632.1961.tb20176.x

Titus, T. G. (1973). Continuous feedback in recognition memory. *Perceptual and Motor Skills, 37*(3), 771-776. doi: 10.2466/pms.1973.37.3.771

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118*(4), 583-613. doi: 10.1037/a0025191

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 582-600. doi: 10.1037/0278-7393.26.3.582

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics*, *68*(4), 643-654. doi:10.3758/BF03208765

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*(2), 254-262. doi: 10.3758/BF03193446

Wickelgren, W.A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5*(1), 102-122. doi:10.1016/0022-2496(68)90059-X

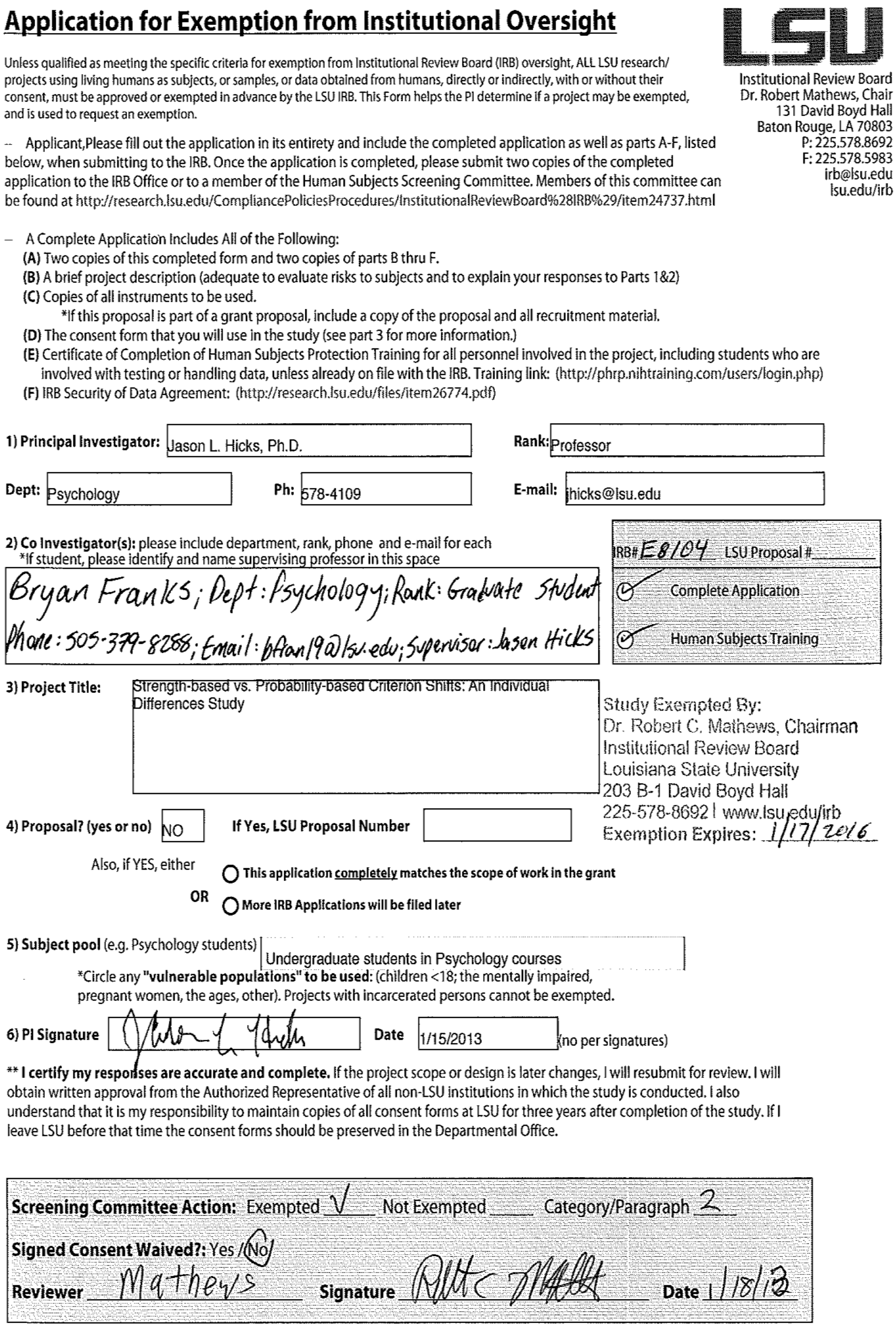
Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*(4), 616-641.

Appendix A: Pilot Experiment Data

Pilot experiment data wherein color cues were manipulated within-subjects and feedback between subjects on a random test. Standard error of the mean in parentheses.

	Unmarked				Marked			
	No Feedback		Feedback		No Feedback		Feedback	
	Weak	Strong	Weak	Strong	Weak	Strong	Weak	Strong
HR	.69 (.01)	.91 (.01)	.69 (.02)	.89 (.01)	.67 (.02)	.86 (.02)	.66 (.02)	.85 (.02)
FAR	.30 (.02)	.28 (.02)	.31 (.02)	.31 (.02)	.31 (.02)	.33 (.02)	.33 (.02)	.34 (.02)
d'	1.15 (.09)	2.13 (.09)	1.14 (.09)	1.91 (.11)	1.08 (.09)	1.72 (.10)	.91 (.07)	1.61 (.11)

Appendix B: IRB Approval Form



Application for Exemption from Institutional Oversight

LSU
Institutional Review Board
Dr. Robert Mathews, Chair
131 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu
lsu.edu/irb

Unless qualified as meeting the specific criteria for exemption from Institutional Review Board (IRB) oversight, ALL LSU research/projects using living humans as subjects, or samples, or data obtained from humans, directly or indirectly, with or without their consent, must be approved or exempted in advance by the LSU IRB. This Form helps the PI determine if a project may be exempted, and is used to request an exemption.

– Applicant,Please fill out the application in its entirety and include the completed application as well as parts A-F, listed below, when submitting to the IRB. Once the application is completed, please submit two copies of the completed application to the IRB Office or to a member of the Human Subjects Screening Committee. Members of this committee can be found at http://research.lsu.edu/CompliancePoliciesProcedures/InstitutionalReviewBoard%28IRB%29/item24737.html

– A Complete Application Includes All of the Following:
(A) Two copies of this completed form and two copies of parts B thru F.
(B) A brief project description (adequate to evaluate risks to subjects and to explain your responses to Parts 1&2)
(C) Copies of all instruments to be used.
*If this proposal is part of a grant proposal, include a copy of the proposal and all recruitment material.
(D) The consent form that you will use in the study (see part 3 for more information.)
(E) Certificate of Completion of Human Subjects Protection Training for all personnel involved in the project, including students who are involved with testing or handling data, unless already on file with the IRB. Training link: (http://phrp.nihtraining.com/users/login.php)
(F) IRB Security of Data Agreement: (http://research.lsu.edu/files/item26774.pdf)

1) Principal Investigator: Jason L. Hicks, Ph.D. **Rank:** Professor

Dept: Psychology **Ph:** 578-4109 **E-mail:** jhicks@lsu.edu

2) Co Investigator(s): please include department, rank, phone and e-mail for each
*If student, please identify and name supervising professor in this space

Bryan Franks; Dept: Psychology; Rank: Graduate Student
Phone: 505-379-8288; Email: bfran19@lsu.edu; Supervisor: Jason Hicks

IRB# E8104 LSU Proposal #______
☑ Complete Application
☑ Human Subjects Training

3) Project Title: Strength-based vs. Probability-based Criterion Shifts: An Individual Differences Study

Study Exempted By:
Dr. Robert C. Mathews, Chairman
Institutional Review Board
Louisiana State University
203 B-1 David Boyd Hall
225-578-8692 | www.lsu.edu/irb
Exemption Expires: 1/17/2016

4) Proposal? (yes or no) NO **If Yes, LSU Proposal Number**

Also, if YES, either ○ **This application completely matches the scope of work in the grant**
OR ○ **More IRB Applications will be filed later**

5) Subject pool (e.g. Psychology students) Undergraduate students in Psychology courses
*Circle any **"vulnerable populations" to be used:** (children <18; the mentally impaired, pregnant women, the ages, other). Projects with incarcerated persons cannot be exempted.

6) PI Signature **Date** 1/15/2013 (no per signatures)

**** I certify my responses are accurate and complete.** If the project scope or design is later changes, I will resubmit for review. I will obtain written approval from the Authorized Representative of all non-LSU institutions in which the study is conducted. I also understand that it is my responsibility to maintain copies of all consent forms at LSU for three years after completion of the study. If I leave LSU before that time the consent forms should be preserved in the Departmental Office.

Screening Committee Action: Exempted ✓ Not Exempted ____ Category/Paragraph 2
Signed Consent Waived?: Yes / [✓ No]
Reviewer Mathews **Signature** **Date** 1/18/13

Vita

Bryan A. Franks graduated from the University of New Mexico in 2012 with a Bachelor of Arts degree in Psychology. Immediately after graduation, he was awarded a competitive research internship funded by the National Science Foundation (NSF) at Clemson University. He entered his first semester as a Ph.D. student at Louisiana State University in the Fall 2012 semester under the supervision of Dr. Jason L. Hicks. His primary research interests involve understanding how people make decisions about their memory.