2006

d' is not appropriate for contrasting yes-no and forced-choice recognition

Moses Michael Langley
Iowa State University

d' is not appropriate for contrasting yes-no and forced-choice recognition

Moses Michael Langley

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Psychology

Program of Study Committee:
Anne M. Cleary, Major Professor
Veronica J. Dark
William S. Robinson

Iowa State University

Ames, Iowa

2006

UMI Number: 1439845

UMI Microform 1439845

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

TABLE OF CONTENTS

ABSTRACT

The equal-variance signal detection (EVSD) model predicts superiority of two-alternative forced-choice (2AFC) detection over yes-no (YN) detection by a factor of $\sqrt{2}$. To make balanced comparisons between detection in these tasks, the equation for calculating 2AFC detection involves a division by $\sqrt{2}$. While the detection literature confirms this prediction (Wickelgren, 1968), the prediction sometimes fails when the model is extended to discrimination tasks (Creelman & Macmillan, 1979). Nevertheless, this model has been widely used in recent years to contrast discrimination in YN and 2AFC tasks. Three experiments tested the $\sqrt{2}$ prediction under conditions that previous research suggests are theoretically ideal for the use of EVSD in discrimination measurement; the $\sqrt{2}$ prediction failed across all three experiments. The present results challenge previous assertions that the EVSD model may be appropriate for discrimination under the present circumstances. The implications of these findings for the study of discrimination are discussed.

CHAPTER 1. INTRODUCTION

Characterizing mental processes can be challenging, especially when the nature of such processes must be inferred from the observation of human behavior. Thus, psychologists take great care in selecting empirically and theoretically sound statistical measures in their studies of cognition. Among recognition memory researchers, a commonly used statistical measure is *d'* (d-prime). d' stems from signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 1991, 2004), and some memory researchers have used this theory as a framework for describing how people discriminate between old and new events (e.g., Parks, 1966). On many occasions, however, memory researchers find that their empirical observations differ from what SDT predicts. Consequently, there has been a growing concern within the memory literature about the use of SDT and d'.

This study focuses on two SDT assumptions that are problematic for the theory's use in recognition memory research: 1) the equal-variance assumption and 2) the √2 assumption. These particular assumptions are important to consider because if made inappropriately they can result in the miscalculation of discrimination ability.

Past research has shown that the equal-variance assumption is often (but not always) violated for discrimination data. However, very little research has explicitly examined the status of the √2 assumption. The √2 assumption is a very important assumption for the numerous studies that have used d' to contrast performance in yes-no (YN) and two-alternative forced-choice (2AFC) discrimination tasks (e.g., Bastin & Van der Linden, 2003; Cook, Marsh, & Hicks, 2005; Deffenbacher, Leu, & Brown, 1981; Green & Moses, 1966; Holdstock et al., 2002; Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000; Kroll, Yonelinas, Dobbins, & Frederick, 2002; Mayes et al., 2002; Nolde, Johnston, & Raye, 1998; Westerberg, Paller, Holdstock, Mayes, & Reber, 2006; Yonelinas, Hockley, & Murdock, 1992). If the √2 assumption is violated, then d′ will not allow for balanced comparison to be made between YN and 2AFC discrimination as SDT claims that it will. Moreover, if the √2 assumption is violated, then the d′ value for 2AFC discrimination may be inaccurate.

The aim of the present study, therefore, was to examine the √2 assumption under optimal conditions for the use of d′ (when the equal-variance assumption is met). This manuscript presents its findings within the context of the ongoing debate over the level of discrimination observed in YN and 2AFC recognition tasks. Thus, the following discussion begins with an exploration of SDT and how it describes performance in two discrimination tasks.

1.1 The YN Task

One widely used task in gauging discrimination performance is the YN task. The YN task requires discrimination between old and new test-items, whereby participants positively endorse a given test-item as old by responding “yes” (it was studied) and reject a new test-item by responding “no” (it was not studied). The accuracy of a participant’s performance is gauged by the proportion of correctly endorsed old test-items and the proportion of incorrectly endorsed new test-items.

1.2 The 2AFC Task

A second widely used task in gauging discrimination performance is the 2AFC task. In 2AFC, two items are presented during each test trial; one item is old and the other item is new. Participants in this task are instructed to endorse the old item (left or right item, or top or bottom item). The accuracy of a participant’s performance can be determined strictly by the proportion of correctly endorsed test-items.

1.3 YN, 2AFC, and the Equal-Variance Signal Detection (EVSD) Model

According to the EVSD model (as applied to discrimination tasks), participants in a memory experiment can discriminate old from new items on the basis of a particular test-item's signal strength. Participants compare the stimulus's signal strength to a *decision criterion*. This criterion is the minimum level of strength the stimulus must have before participants will endorse it as previously encountered. When the signal satisfies the decision criterion, a "yes" response is given, which indicates that a participant recognizes the stimulus as studied (see Figure 1). For example, such a signal could be the amount or degree of familiarity a participant experiences when s/he encounters a stimulus.

Of course, some stimuli are inherently more familiar than others (regardless of their study status). Therefore, in the EVSD model, the decision variable representing the strength of a specific test-item must vary over a range of values. The model assumes the distribution of these values is Gaussian, and that there are separate distributions for old and new stimuli (Figure 1).

In YN discrimination, SDT refers to the presentation of a studied test-item as a "signal trial," and participants are to respond "yes" during such trials. SDT theory refers to the presentation of an unstudied test-item as a "noise trial," and participants are to respond "no" during these trials. A *hit* occurs when participants respond "yes" on a signal trial. The proportion of hits a participant makes is called the *hit rate,* and this equals the proportion of the signal distribution that satisfies the decision criterion. Similarly, a *false alarm* occurs when a participant response "yes" to an unstudied test-item. The proportion of false alarms made (the *false alarm rate*) equals the proportion of the noise distribution that satisfies the decision criterion.

The hit and false alarm rates reflect two factors, response criterion (or response bias) and the degree of overlap between the signal and noise distributions (i.e., discrimination); SDT's acclaim stems from its ability to separate the two. Researchers measure discrimination (d') from the hit and false alarm rates, and this value corresponds to the distance between the mean of signal and the noise distributions (Figure 1). A d' value of zero indicates no discrimination between studied and unstudied test-items, while values reliably greater than zero reflect successful discrimination between old and new stimuli.

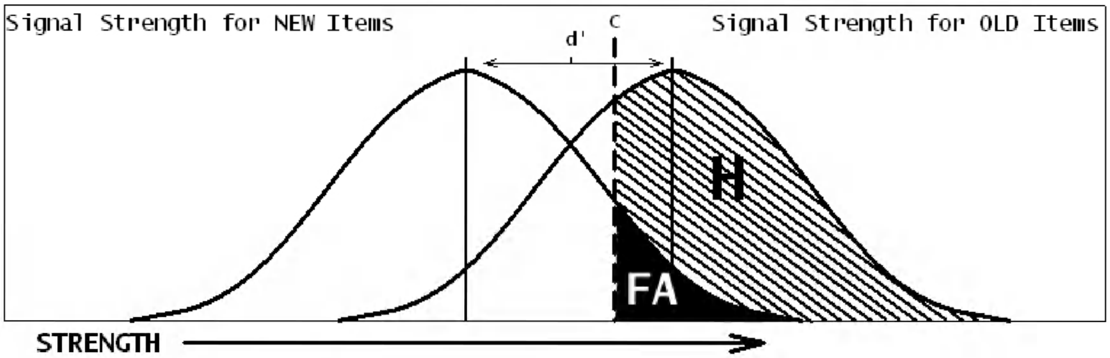


Figure 1. The Gaussian distribution on the left represents unstudied items (noise distribution) and the Gaussian distribution on the right represents studied items (signal distribution). The black region (false alarms, or FA) represents the proportion of the noise distribution that has satisfied *C*, the response criterion. The shaded region (hits, or H) represents the proportion of the signal distribution that has satisfied the response criterion.

In order to separate the influence of the response criterion (i.e., *C*) from discrimination (i.e., d'), the two cannot influence each other. According to the EVSD model, response bias does not affect discrimination if two assumptions are met: 1) the signal distribution and the noise distribution are normal, and 2) the signal distribution and the noise distribution have the same standard deviation (or have equal variances). The consequence of violating either assumption is that discrimination will vary with response criterion.

Numerous studies (e.g., Ratcliff, Sheu, & Gronlund, 1992; Slotnick, Stanley, Dodson, & Shimamura, 2000; Smith and Duncan, 2004) have demonstrated that the data emerging from tests of overall recognition and source memory often violate the equal-variance assumption, especially under YN testing conditions (Smith & Duncan). This has led some researchers to the view that d′ should not be used under such conditions and others to generally assume an unequal-variance signal detection (UVSD) model when examining memory from a detection theory perspective (e.g., Heathcote, 2003, Slotnick & Dodson, 2005, Smith & Duncan).

There are conditions, however, in which equal-variances are observed. For example, some studies have shown satisfaction of the equal-variance assumption when discrimination is based on familiarity in the absence of recollection (assuming a dual-process model; e.g., Cleary, 2005; Dobbins, Kroll, & Qiang, 1998; Yonelinas, 1994). Such demonstrations suggest that d′ may be appropriate for discrimination under these conditions. When equal-variances *are* present, SDT claims that participants' performance in 2AFC can actually be predicted based on their performance in a YN task (this assumes the stimuli are the same between the two tasks; the equations for deriving d' from both YN and 2AFC data can be found in the Appendix) (Green & Swets, 1966). This prediction is possible because of assumptions SDT makes about how YN and 2AFC decision processes relate.

According to SDT, YN discrimination is based on the judgment of a single item's signal strength; the stronger the signal, the further apart the theoretical old and new variance distributions are. The difference between the means of the old and new distributions is the statistic d'_{YN}, and this value is proportional to:

$$d'_{YN} = \frac{\mu_1 - \mu_2}{\sigma} = \frac{\mu_1 - \mu_2}{1} = \mu_1 - \mu_2$$

That is, d′ for YN discrimination (d'_{YN}) is proportional to the difference between the means of the old and new distributions (μ_1, μ_2) divided by the standard deviation of the distributions (which are assumed to be of equal-variance and a value of one). For a 2AFC task, however, it is assumed that discrimination is based on a comparative judgment of two items. Thus, the mean of the difference distribution (which is derived from the *differences* between the two sets old and new distributions) will be equal to two times the distance between the old and new distributions for YN (or two times d'_{YN}). Because the variances themselves are additive when calculating a difference distribution, the variance of the difference distribution also doubles to a value of two, which leaves the standard deviation (the square-root of the variance) between the difference distributions for each item on the 2AFC task at √2. Therefore, the performance advantage 2AFC has over YN *should* be a factor of √2 (Deffenbacher et al., 1981):

$$d'_{2AFC} = \frac{2(\mu_1 - \mu_2)}{\sqrt{\sigma^2 + \sigma^2}} = \frac{2(\mu_1 - \mu_2)}{\sqrt{1^2 + 1^2}} = \frac{2(\mu_1 - \mu_2)}{\sqrt{2}} = \sqrt{2}(\mu_1 - \mu_2) = \sqrt{2}(d'_{YN})$$

That is, d'_{2AFC} should equal $\sqrt{2}(d'_{YN})$. To compensate for the √2 advantage the equation for d'_{2AFC} involves a division by √2, thus allowing for balanced comparisons to be made between YN and 2AFC performance (Macmillan & Creelman, 2004). The justification for the d'_{2AFC} equation, therefore, hinges critically on the fulfillment of the √2 assumption. If the assumption is violated, then the √2 division will compensate 2AFC performance with a value that is not representative of the advantage it actually has over YN (assuming an advantage exists).

What is more, accepting the √2 assumption implies that one is accepting an additional assumption: Participants in a 2AFC task are using both items presented at test in making their old-new decision (Macmillan & Creelman, 2004). Hence forth, this will be referred to as the *optimal performance assumption*, and the d' value derived for 2AFC performance under this assumption will be referred to as *optimal* $d'_{\boldsymbol{2AFC}}$, or merely $d'_{\boldsymbol{2AFC}}$. The alternative assumption is that participants in a 2AFC task do not always use both available pieces of information (i.e., both stimuli) in making their old-new decision. This will be referred to as the *non-optimal performance assumption*, and the d' value derived for 2AFC performance under this assumption will be referred to as *non-optimal* $d'_{\boldsymbol{2AFC}}$.

Under the non-optimal performance assumption, participants are essentially performing a YN task given that their decision is based on only one of the two test items. Therefore, when participants are performing non-optimally, the $\sqrt{2}$ division is not warranted (Macmillan & Creelman, 2004). The traditional calculation of d'_{2AFC} assumes optimal performance (and involves the division by $\sqrt{2}$), which is why the non/optimal performance assumption and the $\sqrt{2}$ assumption are important. If 2AFC performance was non-optimal and the calculation of d'_{2AFC} still assumed optimal performance, then the d'_{2AFC} value obtained would be greatly underestimated; this is a sentiment also discussed by Deffenbacher, Leu, and Brown (1981). As it turns out, the equation for calculating non-optimal d'_{2AFC} is the same as the measure $\boldsymbol{d'_s}$ for 2AFC presented by Creelman and Macmillan (1979); thus, $d'_{s,\,2AFC}$ may be used interchangeably with non-optimal d'_{2AFC}.[1]

As is the case for the equal-variance assumption, there is evidence to suggest that the $\sqrt{2}$ assumption does not hold under all conditions. As its name implies, detection theory was originally conceived to describe performance in detection (not old/new discrimination) tasks.[2] A review of studies contrasting YN and 2AFC detection reveals a long and consistent history of confirmation for the $\sqrt{2}$ relationship in detection (e.g., Wickelgren, 1968). However, upon extension of SDT to old/new discrimination tasks (e.g., Parks, 1966), it was found that 2AFC had an advantage closer to a factor of two than to a factor of $\sqrt{2}$ (e.g., Deffenbacher et al., 1981; Creelman & Macmillan, 1979).[3] Unfortunately, these demonstrations of SDT's failure to predict the relationship between YN and 2AFC *discrimination* have been largely overlooked in subsequent research. The many studies that have contrasted YN and 2AFC discrimination explicitly state that the $\sqrt{2}$ adjustment was made for 2AFC performance (e.g., Bastin & Van der Linden, 2004; Cook, Marsh, & Hicks, 2005; Khoe et al., 2000; Kroll et al., 2002; Smith & Duncam, 2004) or indicate that the adjustment was made by referring to the 2AFC d′ equation in Macmillan and Creelman

(1991) or Green and Swets (1966) (e.g., Holdstock et al., 2002; Westerberg, Paller, Holdstock, Mayes, & Reber, 2006). If it is true that the 2AFC advantage in discrimination is greater than it is in detection, it is possible that the d'_{2AFC} is not appropriate for measuring discrimination. In other words, if the true 2AFC advantage is near two for discrimination, as suggested by Creelman and Macmillan (1979), then memory researchers have been overestimating (by undercompensating with a division by 1.41 instead of 2.0) 2AFC discrimination relative to YN discrimination.

Over the past five decades, many studies have directly contrasted discrimination in YN and 2AFC recognition tasks using d′. A review of this literature reveals that the findings are quite mixed. On one hand, some studies suggest that discrimination does not differ between YN and 2AFC discrimination after the $\sqrt{2}$ adjustment (e.g., Greene & Moses, 1966; Khoe et al., 2000, Yonelinas et al., 1992), a sentiment that provides converging support for the predictions of SDT as they apply to old/new discrimination. A number of other studies suggest that discrimination is superior in the 2AFC task even after the $\sqrt{2}$ adjustment (e.g., Aggleton & Shaw, 1996; Creelman & Macmillan, 1979; Deffenbacher et al., 1981), a sentiment that converges on the idea that the YN/2AFC relationship for discrimination differs from the relationship found in the detection literature. On the other hand, some research suggests that the YN/2AFC relationship for discrimination is variable and depends on the stimuli and encoding conditions (Cook, Marsh, & Hicks, 2005). Further still, some research suggests that d'_{YN} is an unreliable measure of discrimination and can lead to inflated measures of YN discrimination (Kroll et al., 2002). These latter two findings neither support the idea of equivalent discrimination between tasks nor the idea of superior 2AFC discrimination. Rather, as do the studies demonstrating superior 2AFC discrimination, they point to a failure of SDT to generalize from detection tasks to discrimination tasks.

The lack of consensus in the literature implies that YN and 2AFC discrimination may not share the same relationship as SDT describes for YN and 2AFC detection. If this is true, it may have significant implications for the study of recognition memory, especially for recognition examined under 2AFC testing conditions. Because the d'_{2AFC} equation itself assumes a $\sqrt{2}$ relationship with d'_{YN} (and optimal 2AFC performance) a failure of the $\sqrt{2}$ prediction would almost certainly mean a failure of the measure d'_{2AFC} for discrimination.

It was thus the aim of the present study to examine the $\sqrt{2}$ assumption in YN and 2AFC discrimination (as indexed by d′). It seemed most appropriate to examine this assumption under conditions known to satisfy the other assumptions of d′. Previous research has shown that many of the assumptions underlying d′ are satisfied when discrimination is familiarity-based (e.g., Dobbins, Kroll, & Qiang, 1998; Yonelinas, 1994). Therefore, the present study sought to isolate familiarity-based recognition for the purpose of examining the $\sqrt{2}$ assumption.

CHAPTER 2. EXPERIMENT 1

In many instances, dual-process theorists have observed a relative sparing of familiarity (relative to recollection) in populations with mild to server memory impairments (for a review, see Yonelinas, 2002). Consequently, one method that researchers have used to examine recognition-familiarity has been to contrast populations with presumed recollection deficits to healthy young adults. However, an alternative to using individuals with recollection deficits to examine familiarity-based recognition is to create circumstances experimentally that minimize the potential for studied stimuli to be recollected. For example, Gardiner, Gregg, Mashru, and Thaman (2001) showed that impoverished encoding conditions (e.g., divided attention) lead to subsequent recognition based largely on familiarity (and less on recollection). This idea stemmed from prior work (Gregg & Gardiner, 1994, also see Gardiner & Gregg, 1997) showing that when rapidly presented study stimuli are used to create impoverished encoding conditions, subsequent recognition is based primarily on stimulus familiarity.

Coincidentally, very recent research has also taken this approach to isolating familiarity in picture recognition (Langley, Ramos, & Cleary, 2006). In a variation of the recognition without perceptual identification (RWPI) paradigm (Cleary & Greene, 2005), Langley et al. presented participants with masked, 60 ms presentations of line-drawings during an encoding phase. Following the presentation of each drawing, participants were asked to name them; participants were, on average, only able to identify these drawings by name 45% of the time. During a subsequent recognition test, participants were presented with a series of pictures; half of these pictures had been presented briefly during the encoding phase and half were new. It was found that participants could discriminate reasonably well between old and new test pictures even when they were not identified during the encoding phase. In line with the logic of Gregg and Gardiner (1994), it was presumed that poorly encoded stimuli (i.e., the unidentified pictures) would have a low likelihood of being recollected and, therefore, were likely recognized on the basis of their familiarity.

Although it has been used to isolate familiarity, the RWI procedure not been used to address the question of familiarity's role in YN versus 2AFC recognition. For several reasons the use of such a method may be diagnostic. First, unlike some methods of isolating familiarity, the present method does not rely on estimates of recollection and familiarity that are based on introspective reports (i.e., remember-know judgments). Despite the frequency of its use, some recent research suggests that "know" judgments do not provide an accurate familiarity estimate (e.g., Gardiner, Gregg, & Karayianni, 2006).

Second, the present method does not involve estimates of recollection and familiarity that are computationally derived (via process-dissociation procedures). As a consequence, it avoids making assumptions of independent or mutually exclusive processes. Third, the procedure itself is designed to minimize the contribution of recollection in “normal” participants and therefore does not require special populations (e.g., amnesiacs, elderly) to examine familiarity-based memory. Fourth, Cleary (2005) has shown that data indicative of old/new discrimination in the absence of recollection satisfies the equal-variance assumption. Thus, because the RWPI procedure may provide a unique and potentially insightful method for isolating discrimination ability that satisfies the assumptions of EVSD, the present study employed this method to examine the relationship between YN and 2AFC discrimination.

2.1 Method

2.1.1 Participants

Twenty-two Iowa State University undergraduates participated in exchange for credit in their introductory psychology course.

2.1.2 Materials

The experiment was administered via Dell OPTIPLEX GX260 and Dell Dimension 8100 computers. E-prime v1.1 software and 19 inch Dell CRT monitors set at 85Hz and a resolution of 1024 x 768 were used for stimulus presentation. Ninety-six black and white line-drawings from the 260 Snodgrass and Vanderwart (1980) picture set were chosen for use in the present experiment; each line-drawing was 245 x 245 pixels. The 96 line-drawings were divided into two equal sets for the purposes of counterbalancing the stimuli across the different recognition tasks. A mask used in prior, related work (Langley, Cleary, Kostic, & Woods, 2006; Langley, Ramos, & Cleary, 2006) was use here as a forward and backward mask; the mask was 255 x 255 pixels (see Figure 2). The order in which participants completed the two tasks was counterbalanced and all stimuli were randomly assigned to study-test block; the two stimulus sets were also counterbalanced across task and task order.

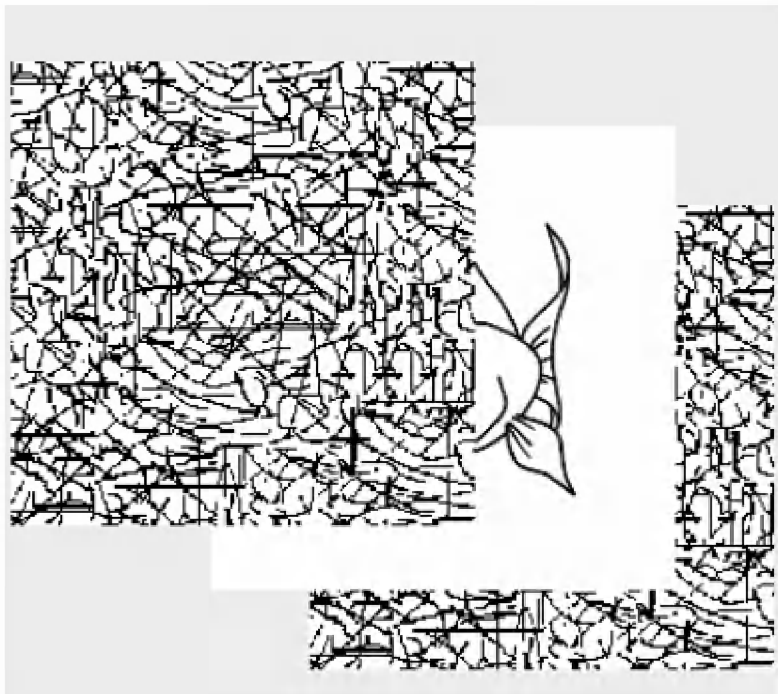


Figure 2. Pictures in the encoding phase were presented for 48ms (Experiment 1a) or 60ms (Experiment 1b) and forward and backward masked. The forward and backward masks (picture above) were presented for 93ms each.

2.1.4 Design and procedure

During the experiment, each participant completed two tasks, a YN task and a 2AFC task. Participants completed three blocks of each task. Within each block, participants viewed an eight-picture study-list. Each study-list was followed by a 16-picture test-list (eight old, eight new).

Prior to the initial study-list, participants read instructions instructing them a) to identify each drawing presented in the study-list and b) that their memory for thee drawings would be subsequently tested. To ensure neutral encoding in each task, participants learned the nature of the test at its onset (after encoding). During each study-phase (across both tasks), a fixation cross appeared for 300 ms central to where the study-list drawings appeared. Study-list pictures were presented for 48 ms in the upper left corner of the monitor. Each line-drawing was forward and backward masked, with each mask presentation lasting 93 ms (see Figure 2).

Immediately following each masked presentation, participants completed a perceptual identification (PI) task. Participants entered their identification response by typing the name of the picture into a dialog box that was present on the computer monitor. After an identification response was made, the next masked line-drawing was presented. If participants could not identify the picture, they were asked to guess. Because each line-drawing was not presented until participants made an attempt to identify the line-drawing that preceded it, the progression of the study-list was participant-paced (or self-paced).

The YN test. In the YN task, 16 pictures composed each test-list. Each test consisted of eight studied and eight unstudied pictures. The study-status (studied or unstudied) of each stimulus was determined randomly at the onset of the experiment (for each participant). Each test-drawing appeared in the upper left-hand corner of the monitor. After two seconds a dialog box appeared asking the participant to judge whether the drawing had appeared briefly in the study-list. The test drawing remained on the screen for 2 s. Participants entered their response (“1” for yes and “2” for no) and pressed the “enter” key to advance to the next briefly presented study-list picture.

The 2AFC test. In the 2AFC task, each of the eight studied pictures was presented along with an unstudied picture during the memory test. That is, eight picture pairs (for a total of 16 test pictures) were presented for participants to judge on each test. The pictures appeared side by side in the upper sector of the monitor. Here, participants chose which of the two pictures (“1” = left, “2” = right) had been presented briefly at study. The studied picture appeared unsystematically but equally often on the left side as on the right side.

2.2 Results

This section first considers performance on the basis of task (YN vs 2AFC). For each task, performance is further conditionalized on the identification status of study-list stimuli. Unless otherwise stated, all values of d' here and in Experiments 2 and 3 are reliably greater than zero at the .05 level. The symbol $\tilde{\tau}$ is used to signify power ($1 - \beta$).

2.2.1 YN recognition performance

Of the 24 briefly the average proportion of identified study pictures was .091 (*SD* = 0.08). Mean discrimination (d'_{YN}) of study pictures from new pictures was 0.42 (*SD* = 0.54, $\tilde{\tau}$ = .97), 95% C.I. = 0.18, 0.66.

Discrimination d'_{YN} between identified study pictures and new pictures was 1.22 (*SD* = 0.74, $\tilde{\tau}$ = 1.0), 95% C.I. = 0.90, 1.56. This finding is not surprising, as participants tend to easily recognize identifiable pictures in this paradigm (Langley et al., 2006).

For pictures participants did not identify during the PI study-task, mean d'_{YN} equaled 0.29 (*SD* = 0.53, $\tilde{\tau}$ = .80), 95% C.I. = 0.05, 0.53. This result demonstrates that even when participants are not successful in a PI task, participants can still recognize unidentified stimuli during subsequent testing. This finding replicates Langley et al.'s (2006) demonstrated of such picture recognition in a YN task.

2.2.2 2AFC recognition performance

The average proportion of identified study pictures was .104 (*SD* = 0.121) and did not differ from the identification rate observed in the YN recognition task, $t(21) = -0.57$, *SE* = 0.02, $p = .58$. Mean discrimination (d'_{2AFC}) between unidentified study pictures and new pictures was 0.34 (*SD* = 0.51, $\tilde{\tau}$ = 1.0), 95% C.I. = 0.11, 0.56.

Mean d'_{2AFC} for pictures identified during the PI study-task was 1.07 (*SD* = 0.79, $\tilde{\tau}$ = 1.0), 95% C.I. = 0.72, 1.42. Mean d'_{2AFC} for unidentified pictures was 0.18 (*SD* = 0.44, $\tilde{\tau}$ =

.58), 95% C.I. = -0.01, 0.38, and this value was only marginally reliable (t = 1.92, p = .07). In general, these results replicate the pattern found in the YN task.

2.2.3 Comparing YN and 2AFC recognition performance

Overall recognition. Discrimination in the YN and 2AFC tasks did not differ, $t(21)$ = 0.61, *S.E.* = 0.13, p = .55 ($\tilde{r}$ = .09). A post-hoc sample size analysis (Lenth, 2006) revealed that given the present data, a sample size of approximately 120 would be necessary to detect a 0.2 difference between the means with alpha at .05.[4]

Identified pictures. Discrimination did not differ between tasks for identified pictures, $t(21)$ = 0.89, *SE* = 0.18, p = .39 ($\tilde{r}$ = .13). However, identification rates were very low in the present experiment (i.e., 9-10%). It is therefore difficult to interpret the discrimination measures obtained because it is based on performance for only three pictures on average. In addition, performance for these items was near ceiling (M = .97, SD = .09); at this level of performance d' can produce unreliable estimates of performance.

Unidentified pictures. Discrimination did not differ between tasks for unidentified pictures, $t(21)$ = 0.77, *SE* = 0.14, p = .45 ($\tilde{r}$ = .12).

Unpublished data from this paradigm demonstrate that discrimination increases and variability decrease with the increase of presentation duration (for both identified and unidentified pictures). With less variability and greater values of d' in both tasks, differences in discrimination between the task (if they exist) should be more easily discernible and require fewer participants to detect than would be necessary under the present conditions. Therefore, in Experiments 2 and 3, the presentation duration of masked study-list pictures was increased to 60-67 ms and the sample size was increased.

2.3 Discussion

Despite low statistical power in the between task contrasts, the results of Experiment 1 at least indicate that the RWPI effect likely extends to forced-choice testing conditions.

The results of the present experiment also confirm the finding of prior studies (e.g., Cleary & Langley, 2006; Langley et al., 2006; Lloyd, Westerman, & Miller, in press), namely that variations of the recognition without identification effect can occur as reliably in YN recognition as in ratings recognition task. However, due to low statistical power, the results of Experiment 1 cannot address the main question of this study: What is the relationship between YN and 2AFC recognition performance? The data of interest are the d' values in each recognition task for overall performance and performance for stimuli that were and were not identified during the PI study-task. As stated, differences may exist between performance in the YN and 2AFC tasks for overall recognition and recognition of identified pictures. However, it is likely that the present experiment did not possess the statistical power necessary to detect these differences. Improvements to the procedure took place in subsequent experiments; therefore, further discussion regarding performance in 2AFC and YN performance is withheld until the discussion of Experiment 2.

CHAPTER 3. EXPERIMENT 2

Experiment 2 sought to increase statistical power. This was done by increasing the presentation duration. Evidence suggests that increasing the presentation duration will increase d' in both YN and 2AFC procedures. For example, Yonelinas et al. (1992) attempted to vary memory trace strength for study-list words by manipulating presentation duration from 50 ms – 1,600 ms. For both YN and 2AFC, d' was found to increase with longer presentation durations. Moreover, unpublished data from this paradigm suggests that discrimination increases linearly with increasing presentation duration. Thus, the prediction is that discrimination in Experiment 2 will mirror the pattern of discrimination found in Experiment 1 but with higher levels of discrimination in each task. Increasing the presentation duration should magnify differences in discrimination between tasks in the present experiment relative to Experiment 1 (if differences exist). If the numerical trend increases linearly between experiments across both tasks, then discrimination may be higher in the YN task (a finding that is atypical in the literature).

3.1 Method

3.1.2 Participants, materials, and procedure

Fifty-seven Iowa State University undergraduates participated. The materials were identical to those used in Experiment 1. With one exception, the procedure in the present experiment was identical to the procedure of Experiment 1. Here, the presentation duration of each study-list picture was increased from 48 ms to 60 ms.

3.2 Results and Discussion

Table 1 illustrates that increasing the presentation duration resulted in higher discrimination for overall recognition and for recognition of identified and unidentified pictures. Figure 3 shows change in discrimination as a function of identification rates with fixed presentation durations (60 ms). As Figure 3 illustrates, Experiment 2 replicated the findings of aforementioned unpublished RWPI data and Yonelinas et al. (1998), demonstrating a linear increase in the magnitude of discrimination with increasing stimulus identifiablity.

3.2.1 YN recognition performance

Identification. The average proportion of identified study pictures was .44 (SD = 0.22), a level of identification attained in prior RWPI studies (Langley et al., 2006).

Discrimination. For overall picture recognition, d'_{YN} was 1.45 (SD = 0.80, $\tilde{r}$ = 1.0), 95% C.I. = (1.24, 1.67). d'_{YN} for identified pictures was 2.53 (SD = 0.72, $\tilde{r}$ = 1.0), 95% C.I. = (2.34, 2.73), and d'_{YN} for unidentified pictures was 0.75 (SD = 0.64, $\tilde{r}$ = 1.0), 95% C.I. = (0.58, 0.92). Like Experiment 1, all d' values reported in Experiment 2 are reliably greater than zero (alpha of .05) unless otherwise stated.

Table 1.

Experiment 2: Hit and False Alarm Rates in Yes-No (YN) and 2-Alternative-Forced-Choice (FC) Recognition

	Overall Picture Recognition				**Identified Pictures**				**Unidentified Pictures**			
Test Format	Hits		False Alarms		Hits		False Alarms		Hits		False Alarms	
	M	*SE*	M	*SE*	M	*SE*	M	*SE*	M	*SE*	M	*SE*
YN	.65	*.02*	.19	*.02*	.93	*.01*	.19	*.02*	.41	*.03*	.19	*.02*
FC	.75	*.02*	.25	*.02*	.90	*.02*	.10	*.02*	.61	*.02*	.39	*.02*

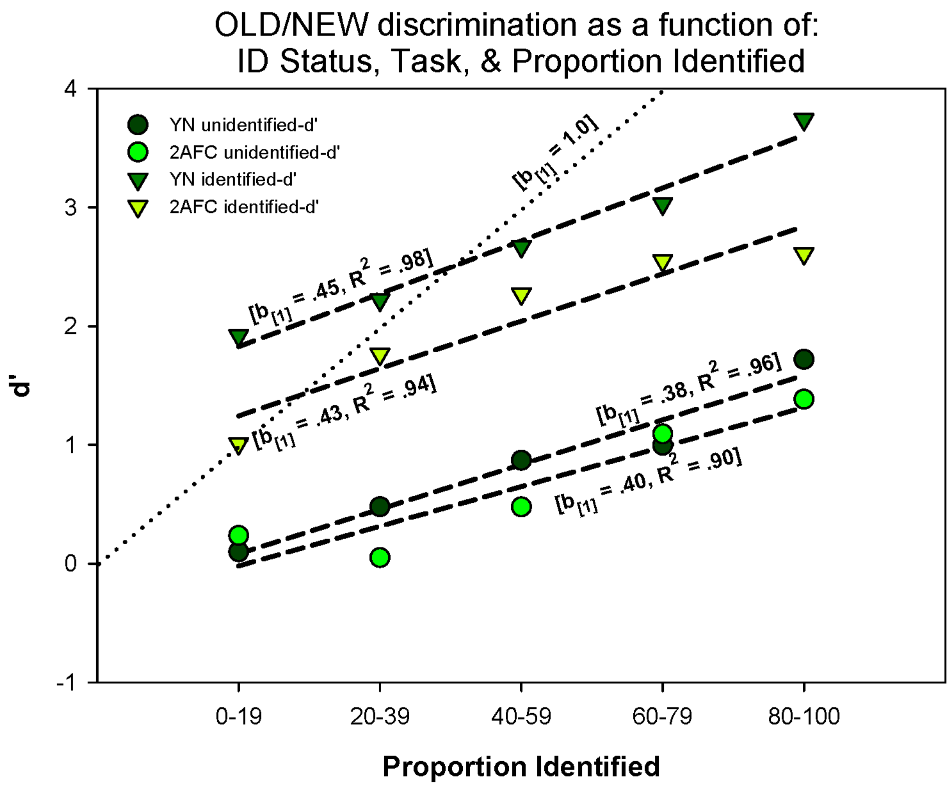


Figure 3. The above graph shows how old/new discrimination (d') is influenced by participants' ability to identify the picture stimuli during the encoding phase. Old/new discrimination for identified and unidentified test pictures is considered separately for each of the two recognition tasks. Dark colors represent YN; light colors represent 2AFC. Triangles represent performance for identified pictures; circles represent performance for unidentified pictures.

3.2.2 2AFC recognition performance

Identification. The average proportion of identified study pictures was .44 (*SD* = 0.23); this identification rate did not differ from the identification rate in the YN task, $t(56) = 0.00$, *SE* = 0.02, $p = 1.0$, 95% C.I. = (-0.05, 0.05).

Discrimination. d'_{2AFC} for overall recognition was 1.09 (*SD* = 0.82, $\tilde{r} = 1.0$), 95% C.I. = (0.87, 1.31). d'_{2AFC} equaled 2.00 (*SD* = 0.79, $\tilde{r} = 1.0$), 95% C.I. = (1.79, 2.20) for identified pictures and 0.46 (*SD* = 0.65, $\tilde{r} = 1.0$), 95% C.I. = (0.29, 0.63) for unidentified pictures.

3.2.3 Comparing YN and 2AFC recognition performance

Under the optimal performance assumption, d'_{YN} was reliably greater than d'_{2AFC} for overall recognition [$t(56) = 3.57$, $SE = 0.1$, $p = .001$, 95% C.I. = (0.16, 0.56), $\tilde{r} = .94$], recognition of identified pictures [$t(56) = 5.37$, $SE = 0.10$, $p < .001$, 95% C.I. = (0.33, 0.74), $\tilde{r} = 1.0$], and recognition of unidentified pictures [$t(56) = 2.73$, $SE = 0.11$, $p < .01$, 95% C.I. = (0.08, 0.50), $\tilde{r} = .85$]. This finding replicates the trend found in Experiment 1 and suggests that discrimination is not equal between tasks, at least not under the present conditions and not when indexed by d′. The present data pattern is inconsistent with studies showing that discrimination is equivalent between tasks (e.g., Khoe et al., 2000) when d′ is used as a statistical measure and also inconsistent with the predictions of SDT.

As a whole, the relevant memory literature suggests that 2AFC recognition should be superior to YN recognition if performance between tasks is found to differ at all. In the present experiments, d' measures of old/new discrimination did not adhere to the predicted pattern. Under all conditionalizations of the data, traditional d' computations show YN recognition to have a reliable advantage over 2AFC recognition (see Table 1). Other published studies have shown this trend (e.g., Kroll et al., 2002; Yonelinas et al., 1992). However, the authors of these studies argued that the trend was the result of confounds, such as the covariance of response criterion and discrimination. To determine whether variance of response criterion with discrimination could also explain the trend in the present data, the YN test advantage (the difference between d'_{YN} and optimal d'_{2AFC}) was plotted against YN response criterion as suggested by Kroll et al.

As Figure 4 illustrates, there was no notable correlation between the YN test advantage and response criterion ($R^2 = 0.01$; $F(1, 55) = 0.73$, $MSR = 0.16$, $p = .40$) for overall recognition. The same held for recognition of both identified and unidentified pictures ($R^2 = 0.04$, $p = .14$; $R^2 = 0.001$, $p = .80$), implying that sensitivity did not vary with response criterion as it did in the study of Kroll et al. (2002). The conclusion, therefore, is that the advantage of YN over 2AFC seen here is not the result of d' being an inappropriate measure due to variation of discrimination with YN criterion.

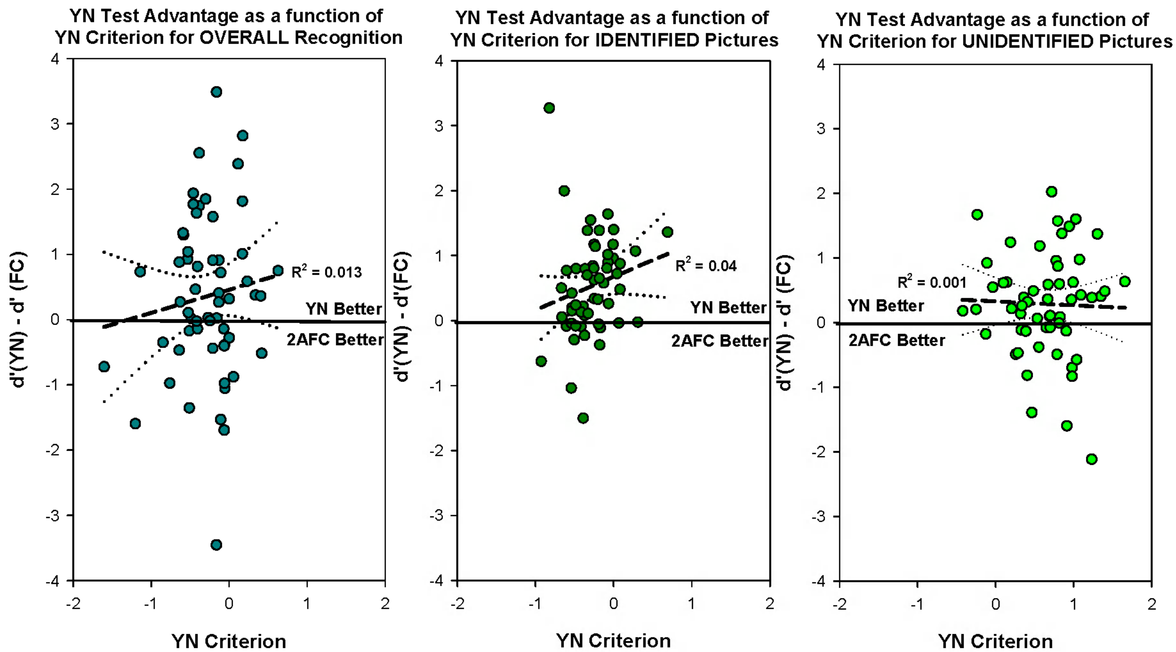


Figure 4. Shown here is the YN test advantage ($d'_{YN} - d'_{2AFC}$) as a function of YN response criterion. Negative data points (those below the zero line) represent participants showing a 2AFC test advantage. Test advantage against criterion for overall recognition is shown in the left panel. The middle panel shows performance only for pictures that were identified at study, while the right panel shows performance for pictures were not identified at study. Linear best fit (R^2) lines are shown as black dashed lines; 95% C.I.s are shown as dotted lines.

3.2.4 Evaluating the Empirical YN/2AFC Discrimination Relationship

Although there was no correlation between the YN test advantage and response criterion, there was a correlation present between YN and 2AFC discrimination; SDT did not specifically predict this correlation, however. SDT predicts that when d'_{2AFC} is corrected for the advantage it supposedly has over d'_{YN}, performance is the same provided that the old and new variance distributions are equal and that discrimination does not vary with response bias. If this were true under the present conditions, the data points in Figure 5 would lie along the diagonal; no such relationship is apparent. As shown in the left panel of

Figure 5, participants in the present experiment tended to perform better on YN than 2AFC for overall recognition. This relationship is linear with a small but reliable correlation between the d'_{YN} and d'_{2AFC} (R^2 = .32; $F(1, 55) = 25.44$, $MSR = 0.45$, $p < .01$). The task to task linear relationship was also present for identified pictures (R^2 = .25, $F(1, 55) = 18.30$, $MSR = 0.40$, $p < .001$) but not for unidentified pictures (R^2 = 0.05, $F(1, 55) = 2.83$, $MSR = 0.40$, p = .10 ; see Figure 5). While the reliable correlations merely show that as performance in one task increased performance in the other did also, they are important because they demonstrate that a) overall discrimination in the two tasks does have a linear relationship, but b) the relationship is not the numerical relationship SDT predicts.

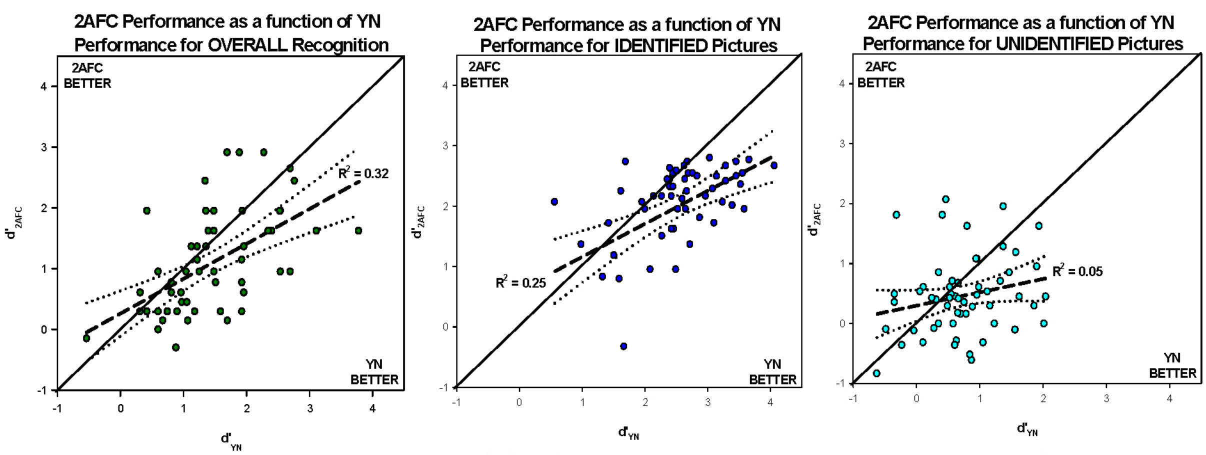


Figure 5. Here, individual participants' performances on the two memory tasks are plotted against each other. Data points to the left of the diagonal represent participants who performed better on 2AFC and points to the right of the diagonal represent those who performed better on YN. Overall recognition (left panel), recognition for identified pictures (middle panel), and recognition for unidentified pictures (right panel) are considered separately. The black dashed lines represents the linear regression best fit (R^2), the dotted lines represent 95% C.I.s.

Directly assessing the √2 prediction first requires the calculation of 2AFC discrimination prior to the √2 adjustment (i.e., $d'_{s,\,2AFC}$). Second, $d'_{s,\,2AFC}$ must be divided by the corresponding d'_{YN} value; this computation produces the empirical advantage that 2AFC discrimination holds over YN discrimination. $d'_{s,\,2AFC}$ for overall discrimination was 1.55 (*SD* = 1.16), 95% C.I. = (1.24, 1.85). For identified pictures, $d'_{s,\,2AFC}$ was 2.81 (*SD* = 1.12), 95% C.I. = (2.52, 3.10), and 0.65 (*SD* = 1.0), 95% C.I. = (0.41, 0.90) for unidentified pictures.

When divided by its respective d'_{YN} value, the 2AFC advantage for overall discrimination was 1.07 (95% C.I. = 0.83, 1.36). The 2AFC advantage for the discrimination of identified pictures was 1.11 (95% C.I. = 0.97, 1.26) and was 0.87 (95% C.I. = .50, 1.33) for the discrimination of unidentified pictures (See Figure 6a). None of the 95% confidence intervals surrounding the 2AFC advantage values contain the 2AFC advantage value that SDT predicts (√2 or 1.41). The confidence intervals therefore show that the √2 assumption was violated for the present experiment under all conditionalizations of the data.

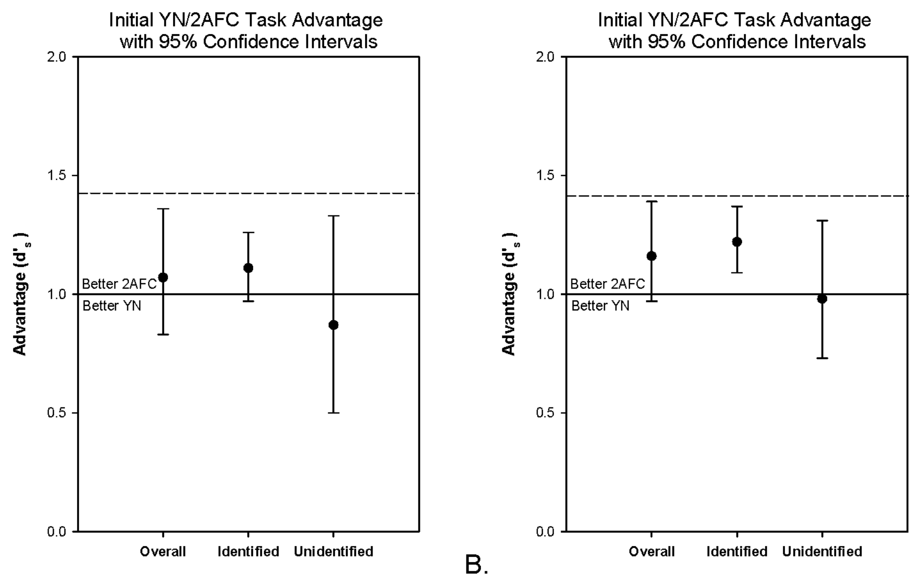


Figure 6. The initial FC advantage over YN was calculated by dividing $d_{s,\ 2AFC}$ by $d_{s,\ YN}$. Figure 6a represents the data from Experiment 2 and Figure 6b represents the data from Experiment 3. The dashed lines represent 1.41 (or the square root of 2), or the level of advantage SDT predicts FC should have over YN. Error bars represent 95% confidence intervals.

Of particular significance is the finding that the √2 assumption was violated for the discrimination of new pictures and those that participants did not identify during encoding. The scenario most likely to produce confirmation of the √2 prediction is a scenario in which other d′ assumptions are met. Past research has shown that a) poorly encoded/unidentified stimuli may be subsequently recognized on the basis of familiarity (e.g., Gregg & Gardiner, 1994; Langley et al., 2006) and b) data implying such recognition-familiarity meets the equal-variance assumption (e.g., Cleary, 2005; Yonelinas, 1994). In instances when old/new discrimination satisfies the equal-variances, d′ should be a fit statistic to measure performance (Yonelinas, 1994). Thus, in the present experiment the data most likely to generate support for the √2 assumption is the data pertaining to unidentified pictures; however, this data does not justify the √2 assumption in the present experiment.

Contrasts of d'_{YN} and $d'_{s,\ 2AFC}$ found no reliable differences in discrimination between the two tasks for unidentified pictures [$t(56) = 0.75$, $SE = 0.13$, $p = .46$], further implying that performance between tasks was equivalent prior to the √2 adjustment. This result is important because not only did SDT's √2 prediction fail under these conditions, but the more general prediction of superior 2AFC performance also failed. Experiment 2 therefore shows that regardless of whether the traditional d' assumptions are satisfied, satisfaction of the √2 assumption did not occur for the present data.

3.2.5 Possible Explanations for the YN Advantage and √2 Assumption Failure

While the present study has thus far demonstrated circumstances in which the √2 prediction fails, it has not produced an explanation for *why* the prediction failed. Regression analysis on the YN advantage and response criterion has ruled out response bias as a factor contributing to the failure of the √2 prediction. An alternative explanation for the failure of the SDT prediction is that participants encoded more information in the YN task than in the 2AFC task. Yonelinas et al. (1992) reported YN superiority over 2AFC across multiple experiments only to find that the YN superiority was due to participants' specific encoding strategies that they developed for each task. Yonelinas et al. argue that if participants have prior knowledge of the type of memory test they are to receive during within-subject experiments, then they may adapt their level of encoding at study to suit the demands of the specific memory test. According to Yonelinas et al., because YN tests are perceived by participants as more difficult, participants intentionally encode more information about study-list items prior to these tests. When controlled for through the randomization of test type order such that the type of test could not be predicted, these researchers found equivalent performance between the two tasks.

While the order of test type was not randomized for each participant in the present experiment, it is unlikely that encoding strategy is the reason for the superiority of YN performance. Participants in the present study had no knowledge about the nature of the memory test they received, nor did they know that there would be multiple types of memory tests. Instead, participants completed three blocks of one test and three blocks of the other (counterbalanced across participants). For participants to develop a strategy capable of influencing performance to the degree needed to explain the present results, they would have needed to adopt a new strategy immediately after the first test-phase of the second test-type (again, because they did not know that a new test type would occur until it was upon them—*after* encoding of the first corresponding study-list). Moreover, this hypothetical encoding rationale assumes that participants developed a specific encoding strategy for the *initial* test type they received. For this to be the case participants would also have needed to immediately engage in the new encoding strategy for the remaining two blocks; this strategy would have needed to be sufficiently distinct and effective for it to influence performance. In contrast to the three blocks of each test type that participants completed here, Yonelinas et al. had their participants complete ten blocks of each test type, a number that clearly would have allowed sufficient time for participants to develop and execute encoding related strategies. The view here is that participants' potential task specific encoding strategies make for an unlikely account of the present data.

An additional possibility is that the $\sqrt{2}$ assumption is only satisfied when the equal-variance assumption is also satisfied. Unfortunately, the question of whether or not the variance distributions were equal in Experiment 2 cannot be answered with the present data. It is therefore possible that the old/new distributions of the data corresponding to unidentified pictures were not equal as previously assumed.

In order to answer this question, a receiver operating characteristic (ROC) curve must be generated.[5] SDT predicts that equal variances should give rise to an ROC slope of 1.0 (approximately) (Macmillan & Creelman, 2004). If the slope is less than 1.0, which is often the case for YN data (Smith & Duncan, 2004), the variances are assumed to be unequal and d' an unfit sensitivity measure (e.g., Kroll et al., 2002). Determining whether the data corresponding to unidentified pictures violated the equal-variance assumption may provide an explanation for why the $\sqrt{2}$ prediction failed.

CHAPTER 4. EXPERIMENT 3

There are two methods by which one can construct and ROC. The first method involves systematically biasing participants' response criteria such that participants produce a level of discrimination for each response criterion (see VanZandt, 2000). The second method of generating an ROC is through the use of confidence ratings. In this method, participants provide numerical ratings indicating how confident they are in their old/new discrimination judgment. The cumulative probabilities of participants choosing each rating constitute the construction of the ROC.[6] Of the two methods, construction of an ROC curve via confidence ratings is generally the simplest method. Moreover, ROCs based on confidence ratings are a frequently used method of generating ROCs from human behavioral data.

Experiment 3 was, therefore, a replication of Experiment 2 with the addition of confidence ratings to each memory test. Confidence ratings will allow for the construction of ROCs, which, in turn, will allow for the examination of the equality of the old/new variance distributions. Upon isolating the data that meet the equal-variance assumption, the assessment of the $\sqrt{2}$ assumption can occur.

4.1 Method

4.1.1 Participants

One-hundred-fifty-eight (79 in each of two conditions) Iowa State University undergraduates were randomly assigned to either the YN condition or the 2AFC condition. Participants received credit in their introductory psychology course for participating.

4.1.2 Materials

The materials in this experiment were identical to those of Experiments 1 and 2 with the following exception. The computer monitors used in Experiment 3 were set at 60Hz rather than 85Hz, which consequently required an increase in the presentation duration from 60 ms to 67 ms.[7]

4.1.3 Design and Procedure

The procedure of Experiment 3 was identical to the procedure of Experiment 2 with the following exceptions. First, to make the memory test more difficult, the length of the study-list increased from eight to 24 pictures. Second, due to the increase in the number of studied pictures (the number of total stimuli remained the same), the number of study-test blocks was reduced from three to two. Third, the design was moved from within-subjects to a between-subjects design. Fourth, participants provided confidence ratings following their recognition decision.

The rating scale was from one to three. For the YN task, a rating of “1” indicated the highest level of confidence (and “3” the lowest) for both the “yes” response and the “no” response. For the 2AFC task, a rating of “1” indicated the highest level of confidence (and “3” the lowest) for both the “left” response and the “right” response.

Table 2.

Experiment 3: Hit and False Alarm Rates in Yes-No (YN) and 2-Alternative-Forced-Choice (FC) Recognition

	Overall Picture Recognition				**Identified Pictures**				**Unidentified Pictures**			
Test Format	Hits		False Alarms		Hits		False Alarms		Hits		False Alarms	
	M	*SE*	M	*SE*	M	*SE*	M	*SE*	M	*SE*	M	*SE*
YN	.69	*.02*	.23	*.01*	.94	*.01*	.25	*.02*	.39	*.02*	.23	*.01*
FC	.76	*.01*	.24	*.01*	.90	*.01*	.10	*.01*	.59	*.01*	.41	*.01*

4.2 Results and Discussion

Table 2 shows the mean hit and false alarm rates for all levels of performance in both tasks; Table 3 shows the mean d′ and d_a values in both tasks. All discrimination are reliably greater than zero (i.e., $p < .05$) unless otherwise stated.

4.2.1 YN Performance

Identification. The average proportion of identified study pictures was .52 (*SD* = 0.21), or approximately 25 pictures.

Discrimination. For overall recognition d'_{YN} was 1.33 (*SD* = 0.73) [95% C.I. (1.17, 1.50)] and d_a was 1.19 (*SD* = 0.68) [95% C.I. (1.04, 1.34)]. These values differed reliably with $d'_{YN} > d_a$ ($p < .001$) (Table 3). For identified pictures d'_{YN} was 2.38 (*SD* = 0.60) [95% C.I. (2.21, 2.54)] and d_a was 2.47 (*SD* = 0.55) [95% C.I. (2.32, 2.62)]. These values differed reliably with $d_a > d'_{YN}$ ($p < .05$) (Table 3). For unidentified pictures d'_{YN} was 0.52 (*SD* = 0.53) [95% C.I. (0.40, 0.64)] and d_a was 0.44 (*SD* = 0.55) [95% C.I. (0.32, 0.56)]. These values differed reliably with $d'_{YN} > d_a$ ($p < .01$). Table 3 also shows the criterion values corresponding to d′ and d_a. The only instance in which criterion was correlated with discrimination was for the d′ and *C* estimates of performance for identified pictures (Figure 7), $r^2 = .18$ ($p < .05$).

ROC analysis. The mean slopes of the zROCs for overall discrimination and discrimination of identified pictures were, respectively, 0.63 (*SD* = .22) [95% C.I. (.59, .68)] and 0.45 (*SD* = .30) [95% C.I. (.35, .51)], and both were reliably less than 1.0 ($p < .01$). The mean slope of the zROC for unidentified pictures, however, approximated 1.0, with a value of 0.94 (*SD* = .31) [95% C.I. (.88, 1.01)], suggesting that, unlike overall YN recognition and YN recognition of identified pictures, YN recognition of *unidentified* pictures is well described by a equal-variance signal detection model, the statistic d′ (Figure 8), and may be based primarily on familiarity.

Table 3.

Experiment 3: Sensitivity & Criterion Measures in Yes-No (YN) and 2-Alternative-Forced-Choice (FC) Recognition

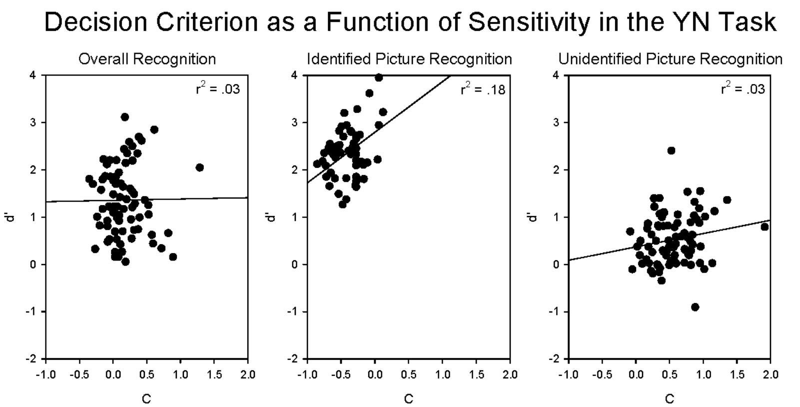
	Overall Picture Recognition					
	Equal-variance		***zROC Intercept***		***Unequal-variance***	
	d'	C	d'_2	C_2	d_a	C_a
Test Format	M/SD	M/SD	M/SD	M/SD	M/SD	M/SD
YN	1.33/.7	.14/.8	1.08/.6	--	1.19/.7	.13/.2
FC	1.09/.6	--	1.04/.6	--	1.07/.6	--

	Identified Pictures					
	Equal-variance		***zROC Intercept***		***Unequal-variance***	
	d'	C	d'_2	C_2	d_a	C_a
Test Format	M/SD	M/SD	M/SD	M/SD	M/SD	M/SD
YN	2.38/.6	-.39/.2	2.29/.6	--	2.47/.6	-.28/.2
FC	2.08/.7	--	2.14/.9	--	2.33/1.0	--

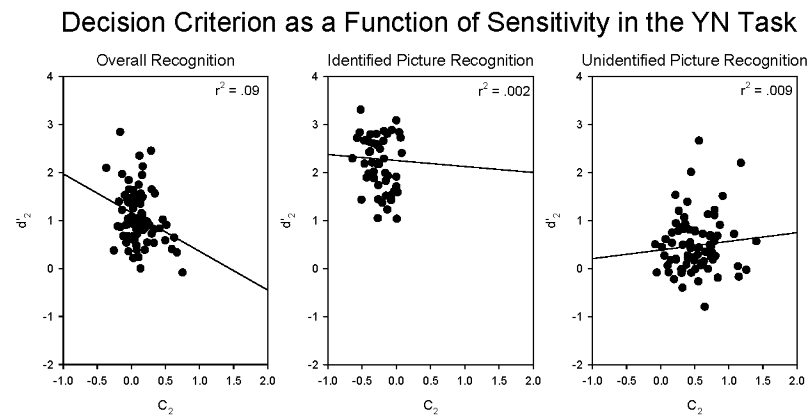
	Unidentified Pictures					
	Equal-variance		***zROC Intercept***		***Unequal-variance***	
	d'	C	d'_2	C_2	d_a	C_a
Test Format	M/SD	M/SD	M/SD	M/SD	M/SD	M/SD
YN	.52/.5	.55/.3	.48/.6	--	.44/.6	.53/.3
FC	.36/.5	--	.44/.7	--	.39/.6	--

Figure 7.

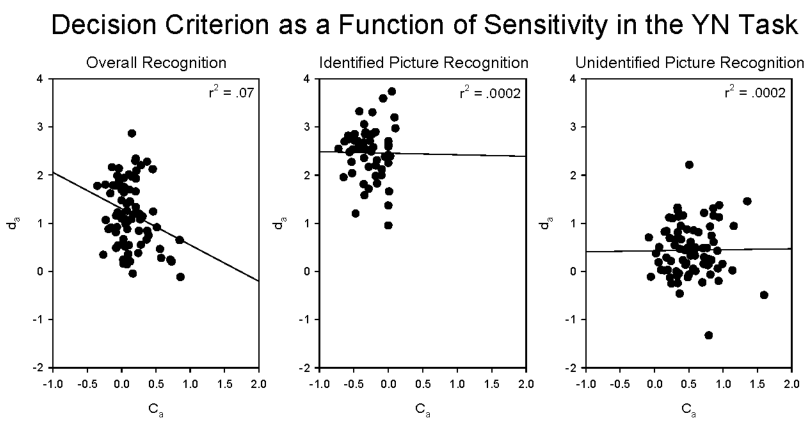
Experiment 2: Criterion and Sensitivity in the Yes-No (YN) Task



A.



B.



C.

Figure 7. The criteria statistics on the x-axis corresponding to its sensitivity statistics on the y-axis are shown above.

4.2.2 2AFC performance

Identification. The average proportion of identified study pictures was .48 (*SD* = 0.22), or approximately 23 pictures; this identification rate did not differ reliably from that of the YN condition (t = 1.07, p = .29).

Discrimination. For overall discrimination, d'_{2AFC} was 1.09 (*SD* = 0.64) [95% C.I. (0.94, 1.23)] and $d_{a,\,2AFC}$ was 1.07 (*SD* = 0.68) [95% C.I. (0.93, 1.21)]. These measures differed reliably with $d'_{2AFC} > d_{a,\,2AFC}$ ($p < .001$) (See Table 3). Although these values differ statistically, they for all practical purposes identical.

For identified pictures d'_{2AFC} was 2.08 (*SD* = 0.71) [95% C.I. (1.89, 2.27)] and $d_{a,\,2AFC}$ was 2.33 (*SD* = 0.99) [95% C.I. (2.06, 2.59)]. These values differed reliably with $d_{a,\,2AFC} > d'_{2AFC}$ ($p < .01$). For unidentified pictures d'_{2AFC} was 0.36 (*SD* = 0.54) [95% C.I. (0.24, 0.48)] and $d_{a,\,2AFC}$ was 0.39 (*SD* = 0.61) [95% C.I. (0.26, 0.53)]. d'_{2AFC} and $d_{a,\,2AFC}$ did not differ for unidentified pictures.

ROC analysis. The mean slopes of the zROCs for overall picture recognition and recognition of identified pictures in the forced-choice task were, respectively, 0.95 (*SD* = .37) [95% C.I. (.86, 1.03)] and .80 (*SD* = .47) [95% C.I. (.67, .92)]. The slope for overall recognition approximated 1.0, but the slope for recognition of identified pictures did not ($p < .001$). The mean slope of the zROC for unidentified pictures also approximated 1.0, with a value of 1.04 (*SD* = .58) [95% C.I. (.91, 1.17)], suggesting that, like overall 2AFC recognition and YN recognition of unidentified pictures, 2AFC recognition of *unidentified* pictures is well described by a familiarity-like equal-variance signal detection process and the statistic d′ (Figure 8).

Figure 8.

Experiment 2: ROCs & zROCs in the Yes-No (YN) and the 2-Alternative-Forced-Choice (FC) Tasks

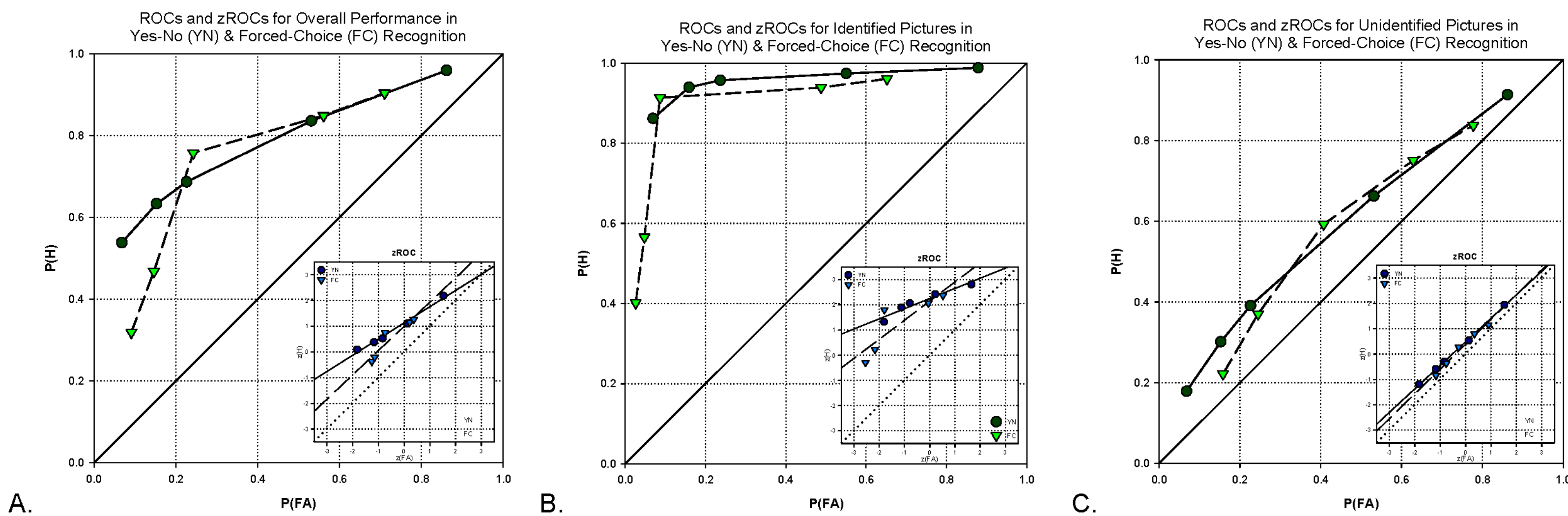


Figure 8. The diagonal line on the ROC graph represents chance performance. The diagonal line on the zROC graph represents a slope of 1.0. In all cases, circles represent YN data and triangles represent 2AFC data.

4.2.3 Comparing YN and 2AFC recognition performance

Overall YN discrimination was greater than 2AFC discrimination according to the d′ statistic [t = 2.48, SE = 0.11, p = .01], but not the d_a statistic [t = 1.38, SE = 0.1, p = .17]. This was also the case for the discrimination of identified pictures, with d′ implying greater discrimination in the YN task than in the 2AFC task [t = 2.46, SE = 0.12, p = .02]; discrimination as indexed by d_a suggested no difference between tasks [t = 0.92, SE = 0.15, p = .36]. The discrimination of unidentified pictures also followed this patter, with superior YN discrimination indexed by d′ [t = 2.03, SE = 0.09, p = .04], but not by d_a [t = 0.52, SE = 0.09, p = .60]. In the present experiment, discrimination as indexed d′ by replicated the pattern found in Experiment 2 for all conditionalizations of the data; discrimination as indexed d_a, however, showed no reliable differences in performance between tasks under any circumstances.

4.2.4 Evaluating the Empirical YN/2AFC Discrimination Relationship

ROC analyses show that the equal-variance assumption was satisfied for the data corresponding to unidentified pictures (Cleary, 2005); this was true in both the YN and the 2AFC tasks and is a finding that permits the assessment of the √2 assumption under optimal conditions. The $d'_{s,\,2AFC}$ value for overall discrimination was 1.54, (SD = 0.91) 95% C.I. = (1.33, 1.74). For identified pictures, $d'_{s,\,2AFC}$ was 2.92 (SD = 1.0), 95% C.I. = (2.68, 3.21), and 0.51 (SD = 0.77), 95% C.I. = 0.33, 0.68 for unidentified pictures. When divided by its respective d'_{YN} value, the 2AFC advantage for overall discrimination was 1.16 (95% C.I. = 0.97, 1.39). The 2AFC advantage for the discrimination of identified pictures was 1.22 (95% C.I. = 1.09, 1.37) and was 0.98 (95% C.I. = .73, 1.31) for the discrimination of unidentified pictures (Figure 6b). As in Experiment 2, none of the 95% confidence intervals surrounding the 2AFC advantage values contain the 2AFC advantage value that SDT predicts (√2 or 1.41). Thus, this analysis strongly suggests that the √2 prediction has failed for the present experiment under all conditionalizations of the data. Once again, of particular significance is the finding that the √2 prediction failed for the discrimination of new pictures and those that participants did not identify during encoding. ROC and zROC analysis confirmed that although the data corresponding to unidentified pictures met the equal-variance assumption in both tasks, the √2 assumption was still not met. This finding implies that even under conditions that are theoretically optimal for the use of d′, this statistic may still violate important assumptions that preclude its use in memory research.

From a dual-process perspective, asymmetries in the ROCs for overall recognition and recognition of identified pictures can also imply the contribution of recollection or recall to recognition performance, which suggests that an equal-variance model assuming a continuous familiarity-type memory process may be inadequate for describing the observed performance. In the same vein, the pattern observed here supports the claim that YN recognition tasks engender recollection processes. However, YN recognition performance for unidentified pictures produced a fairly symmetric ROC curve and a zROC slope approximating 1.0. This finding is consistent with the argument that recognition without identification effects are based primarily on familiarity (e.g., Cleary, 2005) and, more specifically, that the recognition of stimuli encoded under impoverished conditions results from familiarity-based memory (e.g., Gregg & Gardiner, 1997; Langley, Ramos, & Cleary, 2006).

With regard to forced-choice recognition, a somewhat different pattern of results were found via ROC analysis. Overall picture recognition and recognition of unidentified pictures produced symmetric ROCs and zROCs with slopes approximately equal to 1.0. These observations generally imply that SDT adequately described forced-choice recognition when it is not approaching ceiling performance. Moreover, from a dual-process perspective these findings support the claim that 2AFC may engender familiarity-based recognition (or engender recollection to a lesser degree). When contrasted with YN performance, the present findings could be interpreted as evidence for the differential reliance of recognition tasks on the processes subserving recognition memory.

What is clear from the present observations is that the equal-variance assumption is not satisfied for all conditionalizations of the data, especially for YN recognition, and this latter finding replicates the findings of many previous studies. Moreover, the $\sqrt{2}$ assumption was also unsatisfied, as evident by the reliable superiority of d'_{YN} over d'_{2AFC} across all Experiment 3 data. The failure of these assumptions precludes the use of d′ as a reliable measure of old/new discrimination under the present experimental conditions. Consequently, the d′ estimates of YN and 2AFC cannot be relied up on to assess the performance relationship between the two tasks. To determine whether the $\sqrt{2}$ relationship between YN and 2AFC recognition is actually present, we must turn to alternative SDT statistics. Discrimination measures obtained from the SDT statistics d_a indicated no reliable differences between YN and 2AFC recognition. This measure of discrimination is essentially in agreement with the prediction of SDT in that performance between the two tasks did not differ statistically following the $\sqrt{2}$ adjustment. Therefore, the present findings support the view that when a SDT approach is taken in measuring discrimination, the UVSD model and the d_a statistic are the most appropriate.

CHAPTER 5. GENERAL DISCUSSION

The present study contributes to the growing concern in the literature regarding the use of d′ in the measurement of recognition memory. The focus here was on a) the use of d′ in the comparison of memory performance in YN and 2AFC tasks and b) the $\sqrt{2}$ assumption underlying d′. Experiments 2 and 3 demonstrate the failure of d′'s theoretical underpinnings as they apply to discrimination tasks.

To briefly review the main findings, in Experiment 1 it was shown that the RWPI effect extended to a forced-choice recognition task and that recognition without identification in YN recognition extended to conditions involving the impoverished encoding of pictorial stimuli. Experiment 2 replicated these findings and YN discrimination was observed to be superior to 2AFC discrimination when d′ was used to index recognition. This pattern of results showed SDT's failure to predict the relationship between YN and 2AFC discrimination and also implies that the equation given for estimating sensitivity in 2AFC discrimination is flawed. Further, it was argued that the greater d'_{YN} value was not the result of the variation of response criterion with discrimination or participants engaging in specific encoding strategies.

Experiment 3 replicated the failure of SDT to predict the relationship between YN and 2AFC discrimination in most cases. Regarding overall recognition, ROC analyses showed the violation of the equal-variance assumption, which has been shown previously under YN testing conditions (e.g., Smith & Duncan, 2004). This, in conjunction with the violation of the $\sqrt{2}$ assumption, implies that d′ is an unfit discrimination measure for the present data. Interestingly, even when the variances were observed to be roughly equal in both tasks, the $\sqrt{2}$ prediction still failed (e.g., as was the case for YN recognition of unidentified pictures). So, the issue is not that the violation of the equal-variance assumption caused the $\sqrt{2}$ violation. Rather, the two violations co-occurred in a seemingly independent fashion.

Because the data were found to be in violation of the equal-variance assumption, the unequal-variance SDT statistic, d_a, was used to estimate discrimination; it satisfied the $\sqrt{2}$ assumption (and the variance assumption by definition). With regard to the relationship between YN and 2AFC episodic recognition, the use of d_a here suggests that 2AFC does in fact roughly maintain a 1.41 (i.e., √2) advantage over YN, and when 2AFC performance is adjusted downward, performance in each tasks is statistically equivalent.

The present study illustrates the potential distortion in discrimination measurement that can occur when the process of verifying theoretical assumptions is overlooked. Here, discrimination in the YN task appeared to be superior to discrimination in the 2AFC task when d′ was used to estimate performance. When the assumptions underlying this performance measure were examined, however, it became clear that d′ was inappropriate. When a more theoretically justified performance measure was used (d_a), the pattern of results changed. That is, rather than a data pattern indicating that YN discrimination was superior to 2AFC discrimination, a pattern indicating equivalent discrimination between the two tasks was observed. By all accounts it appears as though YN performance in the present study is inflated by d′'s measure, and there are two main factors suggesting that this is the case. First, t-tests contrasting d'_{YN} and $d_{a,\,YN}$ showed $d'_{YN} > d_{a,\,YN}$ for overall picture recognition and for recognition of unidentified pictures. When the old and new variance distributions are equal, these two estimates should be equivalent. When the variances are unequal, d′ provides a distorted estimate of performance, and in this case the distortion is in the form of a heightened value. Second, $d'_{YN} > d'_{2AFC}$ was observed for all conditionalizations of the data. When variances are equal, SDT predicts that the measures will be equal. Given that the variances are not equal, d_a must be used, and when it is $d_{a,\,YN} = d_{a,\,2AFC}$ (as SDT predicts). Thus, the use of d′ appears to result in inflated measures of discrimination. Kroll et al. (2002) also found that d'_{YN} produced inflated performance measures for YN discrimination (but for different reasons). They too recommended abandoning the use of d′, though without acknowledging the failed √2 prediction that they also observed.

While the recognition task used in the present study is not considered "standard," the results are similar to those emerging from more typical recognition procedures. It has been over 25 years since Deffenbacher et al. (1981) and Creelman and Macmillan (1979) acknowledged that d′"s √2 relationship was not present for YN and 2AFC discrimination tasks. Since these initial studies, however, memory researchers have continued to use the d′ statistic in measuring and contrasting YN and 2AFC recognition, as if the predictions had been confirmed all along. This course of events could be especially damaging for memory research performed under 2AFC testing conditions, given that the accuracy of the equation for the widely used SDT sensitivity estimate, d'_{2AFC}, depends on the accuracy of the √2 prediction.

Further support for the findings presented here can be seen by looking to the memory literature itself. Most, if not all, of the studies comparing YN and 2AFC discrimination have assumed the validity of the √2 prediction and therefore adjusted downward the estimate of 2AFC discrimination. If this were the correct action to take, estimates of discrimination in YN and 2AFC would be statistically equivalent, as SDT predicts. Only a handful of studies find equivalent performance when performance is measured with d′, and of these studies some only find equivalence under specific experimental conditions (e.g., Basin & Van der Linden, 2003). Upon close inspection, many studies are an example of the √2 prediction failing, as confirmation of this prediction only occurs when performance is found to be equivalent after adjusting 2AFC performance.

A question that the present study raises is how to address the fact that the equation for d'_{2AFC} may be flawed when it comes to measuring performance in discrimination tasks. As mentioned, the majority of studies comparing YN and 2AFC performance have done so using d′. If the equation for d'_{2AFC} is in fact inappropriate for discrimination tasks, then the results of decades of memory research that have either a) used d′ to estimate 2AFC recognition performance alone or b) used d'_{2AFC} in a contrast with d'_{2AFC} may be in question. As an example of the consequences an inappropriate equation could have on the outcome of a study, consider the study of Cook, Marsh, and Hicks (2005). Like in the studies of Kroll et al. (2002) and the present study, Cook et al. observed an instance of superior YN performance over 2AFC as indexed by d′. Specifically, these researchers observed overall YN discrimination to be superior to overall 2AFC discrimination for uncommon words; overall performance was observed to be equivalent between the two tasks for very common words. The researchers went on to calculate estimates of recollection and familiarity in each task and asserted that recollection was higher in YN for uncommon words than it was in 2AFC. They suggested that one possible explanation for this finding was that performance on YN tests was superior to that of 2AFC tests when the stimuli were particularly recollectable (as uncommon words are known to be), again because YN recognition engenders recollection more so than 2AFC recognition.

An alternative interpretation of those data is that because the stimuli were highly recollectable, participants in the 2AFC were performing "*non-optimally*." When participants perform *non-optimally* in a 2AFC task, they do not use all of the information available at test (i.e., both words) in making their recognition judgment. Thus, the participants essentially convert the 2AFC test to a YN test. This is most likely to occur when confidence and recollection are high, and under these conditions the $\sqrt{2}$ correction for 2AFC may not be warranted (Macmillan & Creelman, 2004). Perhaps coincidentally (but perhaps not), when the $\sqrt{2}$ correction of overall 2AFC recognition is undone in the Cook et al. study, the new d'_{2AFC} estimate (i.e., $d'_{s,\,2AFC}$) is nearly identical to the reported d'_{YN} estimate. This suggests that the pre-adjusted d'_{2AFC} had no initial advantage over the d'_{YN} estimate, raising the question once again of whether one should attempt to compensate for an advantage that is not initially present. If the data had been measured this way, the conclusion would not have been one of YN discrimination surpassing 2AFC discrimination for highly recollectable stimuli, but rather one of equivalent overall performance for highly recollectable stimuli *and* highly familiar stimuli (i.e., common words). What might otherwise be described as null effects if *optimal* performance had not been assumed, the implications of the Cook et al. study rest squarely on their claim that YN performance is superior to 2AFC performance for highly recollectable stimuli. Their claim of YN superiority can also be challenged simply by acknowledging that the √2 assumption failed for their data pertaining to uncommon words, thus making their estimate of 2AFC discrimination unreliable. In addition, because the variance assumption was not assessed for their data, there is little basis to argue that d′ is an appropriate index for YN or 2AFC.

The example above should illuminate the potential problems that may arise from the prospect of d'_{2AFC} as a flawed discrimination performance estimate. The degree to which the d'_{2AFC} equation relies upon the accuracy of the √2 assumption is a matter that should be debated, as it is certain that contrary views on this issue exist. For the time being, however, it can be said with some certainty that both the equal-variance assumption and the √2 assumption are frequently violated by what is found empirically and that the ramifications of such violations of theory must be understood. The present study has highlighted these issues and was undertaken with the goal of initiating dialog on this important issue.

REFERENCES

Aggleton, J. P., & Shaw, C. (1996). Amnesia and recognition memory: A re-analysis of psychometric data. *Neuropsychologica, 34*, 51-62.

Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology, 17*, 14-24.

Cleary, A. M. (2005). ROCs in recognition with and without identification. *Memory, 5*, 472-483.

Cleary, A. M., & Greene, R. L. (2005). Recognition without perceptual identification: A measure of familiarity? *Quarterly Journal of Experimental Psychology*, *58A,* 1143-1152.

Cleary A. M., & Langley, M. M. (*under revision*). Recognition without cued-recall: A tool for assessing the remember-know procedure (and vice versa).

Cleary, A. M., Langley, M. M., & Seiler, K. R. (2004). Recognition without picture identification: Geons as components of the pictorial memory trace. *Psychonomic Bulletin & Review, 11,* 903-908.

Cook, G. I., Marsh, R. L., & Hicks, J. L. (2005). Revisiting the role of recollection in item versus forced-choice recognition memory. *Psychonomic Bulletin & Review, 12*, 720-725.

Creelman, C. D., & Macmillan, N. A. (1979). Auditory phase and frequency discrimination: A comparison of nine procedures. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 146-156.

Deffenbacher, K. A., Leu, J. R., & Brown, E. L. (1981). Memory for faces: Testing methods, encoding strategy, and confidence. *American Journal of Psychology, 94,* 13-26.

Dobbins, I. G., Kroll, N. E. A., & Qiang, L. (1998). Confidence-accuracy inversions in scene recognition: A remember-know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345-355.

Gardiner, J. G., & Gregg, V. H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review, 4,* 474-479.

Gardiner, J. G., Gregg, V. H., Mashru, R. & Thaman, M. (2001). Impact of encoding depth of perceptual effects in recognition memory. *Memory & Cognition, 29*, 433-440.

Gregg, V. H., & Gardiner, J. M. (1994). Recognition memory and awareness: A large effect of study-test modalities on “know” responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology, 6,* 131-147.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Huntington, NY: Robert E. Krieger.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1210-1230.

Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O’Reily, R. C., & Norman, K. A. (2002). Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus, 12,* 341-351.

Khoe, W., Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Knight, R. T. (2000). The contribution of recollection and familiarity to yes-no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologica, 38*, 1333-1341.

Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measure of recognition memory. *Journal of Experimental Psychology: General, 131,* 241-254.

Lenth, R. V. (2006). Java Applets for Power and Sample Size [Computer software]. Retrieved *May 10th, 2006,* from http://www.stat.uiowa.edu/~rlenth/Power.

Langley, M. M., Ramos, A. L., & Cleary, A. M. (*under revision*). Recognition-familiarity for pictures: Effects of incongruity in size, orientation, and modality.

Lloyd, M. E., Westerman, D. L. & Miller, J. K. (in press). Familiarity from orthographic information: Extensions of the recognition without identification effect. *Memory & Cognition.*

Lockhart, R. S. (2000). Methods of memory research. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 45-57). New York: Oxford University Press.

Macmillan, N. A., & Creelman, C. D. (1991, 2004). *Detection theory: A user's guide*. New York: Cambridge University Press.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380-387.

Nolde, S. F., Johnston, M. K., & Raye, C. L. (1998). The role of prefrontal cortex during tests of episodic memory. *Trends in Cognitive Science, 2*, 399-406.

Parks, T. E. (1966). Signal-detection theory of recognition-memory performance. *Psychological Review, 73*, 44-58.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition, 33,* 151-170.

Slotnick, S. D., Stanley, K. A., Dodson, C. S., & Shimamura, A. P. (2000). Analysis of signal detection and threshold models of source memory. *Journal of Psychology:*

Learning, Memory, and Cognition, 26, 1499-1517.

Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30,* 615-625.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6,* 174-215.

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

Viemeister, N. F. (1970). Intensity discrimination: Performance in three paradigms. *Perception & Psychophysics, 8*, 417-419.

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5*, 102-122.

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

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

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345-355.

APPENDIX

Discrimination

$$d'_{YN} = \Phi^{-1}(H) - \Phi^{-1}(F)$$
(Equal-variance assumed)

$$d_{a\ YN} = [2/(1 + s^2)]^{1/2} * [\Phi^{-1}(H) - s(\Phi^{-1}(F))]$$
(Unequal-variance assumed)

$$d'_2 = \Phi^{-1}(H) - s(\Phi^{-1}(F)$$

$$\text{Optimal } d'_{2AFC} = [\Phi^{-1}(P_{correct}) - \Phi^{-1}(P_{incorrect})]/\sqrt{2}$$
(Equal-variance assumed)

$$d'_{s,\ 2AFC} = \text{Non-optimal } d'_{2AFC} = 1/[\Phi^{-1}(P_{correct}) - \Phi^{-1}(P_{incorrect})]$$
(Equal-variance assumed)

$$d_{a\ 2AFC} = \sqrt{2}\ \Phi^{-1}(P_{correct})$$
(Unequal-variance assumed)

Response Bias

$$c = -[\Phi^{-1}(H) + \Phi^{-1}(F)]/2$$

$$c_a = [-((\sqrt{2}) * (s))/((1 + s)^{1/2} * (1 + s))] * [\Phi^{-1}(H) + \Phi^{-1}(F)]$$

$$c_2 = [-s/(1 + s)] * [\Phi^{-1}(H) + \Phi^{-1}(F)]$$

AUTHOR NOTE

There are many to thank for their contributions to this manuscript. First, the author would like to thank Joshua A. Woods, Bogdan N. Kostic, Kaeli Sampson, Kate Lyon, and Joe Williams for their assistance in data collection. The author would also like to thank Amy L. Ramos, Ken J. Malmberg, Douglas G. Bonett, Veronica J. Dark, and William S. Robinson for their discussion and advice on various aspects of this work. The data presented here was collected primarily in Anne M. Cleary's Human Memory Laboratory and partially in Alison L. Morris's Psycholinguistics Laboratory, and many thanks are due for the use of their lab resources.

FOOTNOTES

1 For the YN task, d' is equal to d'_s.

2 The difference between detection and discrimination rests with the demand that each task places on memory (Viemeister, 1970). For YN detection, participants merely decide whether the stimulus is a signal or a non-signal. For example, a participant may be asked to respond “yes” to three second tones and “no” to everything else. Here, all that is required from memory is the temporal characteristic of the signal. For YN discrimination, however, participants must respond “yes” to a number of to-be-remembered tones (signals) and “no” to new tones (non-signals). Here, a participant must be able to recollect all of the to-be-remembered tones so that they can be differentiated from those that are new. Obviously, the discrimination task is more demanding of memory than detection tasks and thus is likely more difficult.

3 Deffenbacher et al. (1981) report a 1.61 advantage of 2AFC over YN while Creelman and Macmillan (1979) report a ~2 advantage. Due to the scarcity of studies that have explicitly acknowledged the numerical advantage 2AFC had over YN (prior to the $\sqrt{2}$ correction), a formal and rigorously tested value for the 2AFC advantage in discrimination tasks has not been established. What can be said with certainty, however, is that SDT does not provide a valid theoretical prediction for the relationship between YN and 2AFC *discrimination* tasks.

4 The value 0.20 represents the smallest difference between d' values that is of theoretical interest. This value is somewhat arbitrary, but not entirely. Previous data emerging from the present procedure has shown that differences between d' values ranging from 0.20-0.40 often become reliable with a high sample size. Differences below this range generally do not become reliable, and differences above this range are generally reliable with few participants. To err on the side of caution, the low end

of said range was used as the smallest theoretically interesting difference between d' values.

5 For discussion of receiver operating characteristics, the reader is referred to Macmilliam and Creelman (2004, p. 51-77).

6 For in depth reviews on construction of ROCs from confidence ratings, see Malmberg (2002) and Van Zandt (2000).

7 One of two 19 inch CRT monitors capable of performing at 85Hz was stolen prior to the onset of Experiment 3. The replacement monitor was of lower standard, and was capable of performing at 60Hz. Therefore, all monitors used in Experiment 3 were set at 60Hz.