Running Head: RANDOMNESS IN PERCEPTION

Using PPT to Account for Randomness in Perception

David Trafimow

Justin A. MacDonald

Stephen Rice

New Mexico State University

Direct Correspondence to:

David Trafimow
Department of Psychology, MSC 3452
New Mexico State University
PO Box 3001
Las Cruces, NM 88003-8001
dtrafimo@nmsu.edu

Abstract

According to many theories of decision making, of which signal detection theory is the most prominent, randomness is the main factor responsible for imperfect performance. These theories imply that correcting for attenuation due to randomness should result in perfect scores as long as the participants use non-extreme decision criteria. Based on a recent advance termed potential performance theory (Trafimow & Rice, 2008), we performed an auditory and a visual detection experiment and corrected the scores for attenuation. Most participants in both experiments tended to perform at a less than perfect level, even after their scores were corrected. The findings demonstrate that there is at least one systematic factor influencing detection that is not included in signal detection theory.

Words: XXXX

Keywords: Potential Performance Theory, Signal Detection Theory, Randomness, Perception

Using PPT to Account for Randomness in Perception

Suppose that participants are presented with visual or auditory stimuli in a detection or discrimination task. If the task is sufficiently difficult, participants will make errors. The large class of statistical decision models, of which signal detection theory is the most prominent member (e.g., Wald, 1950; Swets, Tanner, & Birdsall, 1961), features the notion of random noise as the main factor responsible for the lack of perfect responding. Although the current research is focused on signal detection theory due to its prominence in the literature, the research to be presented is more generally applicable to the notion of randomness as an explanatory factor in perception studies.

Consider a two interval forced choice (2IFC) detection task where the participant attempts to determine the interval that contains the signal stimulus. According to signal detection theory (hereafter, SDT), the observer is given a pair of percepts (one for each interval) and generates a response by identifying the interval that contains the largest percept. SDT is based on the assumption that random noise is included in the percepts given to the observer for classification: distributions of percepts (typically Gaussian) are associated with each stimulus type. Given an unlucky sampling, the percept associated with the signal interval could be less intense than the percept associated with the noise interval, and an error could occur.

Of course, a variety of challenges to the classic equal-variances Gaussian version of SDT have been put forward over the years. Most notably, empirical evidence has accumulated to support the unequal variances version of SDT over the classic equal-variances version (Ratcliff, Shue, & Gronlund, 1992; Swets et al., 1979; Wixted, 2007). Typically, the variance of the target (signal) distribution is estimated to be larger than that of the noise distribution (Mickes, Wixted, & Wais, 2007; Egan, 1958, 1975; Ratcliff et al., 1992). Representing a more fundamental attack on SDT, Balakrishnan (1999c; see also Balakrishnan & MacDonald, 2008) presented evidence suggesting that

the decision criterion does not shift under response bias manipulations. Mueller and Weidemann (2008; see also Benjamin, Diaz, & Wee, 2009) attempted to account for these findings by adding random noise to the placement of the decision criterion. Complicating matters, Wickelgren and Norman (1966) demonstrated that there is no theoretical way to distinguish percept noise from criterion noise within the SDT framework, effectively requiring empirical methods to lend support to the criterion variability account.

Each of these refinements to the classical equal-variances SDT model essentially amount to inserting additional variability (randomness) into the model. Because there might be different types of randomness (criterial, perceptual, or others), different conditions under which different types of randomness matter more or less, and because the variances of the signal and noise distributions may be equal or unequal, it is very difficult to evaluate the basic question of concern: can the SDT framework provide a full and accurate account of decision making behavior? In other words, if we keep adding sources of nonsystematic variability to the model will we eventually capture the complete decision process? The aim of this paper is to answer this question. The usual method would be to add sources of nonsystematic variability incrementally, assessing model fit after each addition. We chose to attack the question from the other direction: if all sources of nonsystematic randomness (variability) were removed from whatever process is employed by the human observer, is the resulting behavior consistent with SDT? If yes, then the continued incremental addition of nonsystematic variation to the SDT model is appropriate. If no, then there is a source of systematic (non-random) variability in the human observer that is missing from the model. If it is really true that imperfect responding is due to randomness, then it implies that future researchers should devote their efforts towards uncovering the different sources of randomness and to formulating models that feature them. In contrast, if there is an unsuspected nonrandom reason for imperfect responding, the

implication is that future researchers should devote their efforts towards discovering what it is and describing it.

Pretend for a moment that it is possible to magically remove all randomness in responding on 2IFC tasks. This includes percept randomness, criterion randomness, and even types of randomness that have not yet been discovered or inserted into the SDT framework. Although we are well aware that it is impossible to achieve this in practice, a recent advance, termed *potential performance theory* (PPT), renders it possible to estimate performance in the absence of random noise (in other words, to correct for randomness). The point of this paper is to identify the corrected performance of the SDT observer and play out the consequences for SDT. Before we can accomplish this it is necessary to briefly describe PPT.

A Brief Description of PPT

According to PPT (Trafimow & Rice, 2008; 2009; also see Rice et al., 2010; 2011a; 2011b, 2012; Trafimow et al., 2011 for further empirical support), there are two factors that influence task performance. One of them is consistency. To understand the role of consistency in PPT, imagine that participants complete a 2IFC detection experiment that consists of two blocks of 50 trials. Further imagine that the trials across blocks can be paired up by stimulus type so that there are 50 pairs of responses for each participant. The correlation coefficient computed from these paired data is a consistency measure. PPT defines randomness in the classical way, as a lack of consistency (see Gulliksen, 1987; Lord and Novick, 1968 for details). Thus, the consistency coefficient can be considered an inverse measure of randomness; less randomness implies a larger consistency coefficient and zero randomness implies that the consistency coefficient will equal unity.

Imagine an extreme example where a person's observed responses were completely random. In this extreme example, the person's expected value would equal chance. For tasks involving two

possible responses that are of concern to PPT, complete randomness implies that, on average, performance should be at the 50% correct level. In addition, the expected value of the consistency coefficient would be zero. Or to think of it in the reverse way, reducing randomness will push consistency coefficients and observed scores higher, providing that the systematic factors at play favor better than chance performance. Assuming that one has a measure of consistency (e.g., there is a consistency coefficient across two blocks of trials), and an observed score for each participant, PPT provides equations that estimate what the person would have scored had he or she been perfectly consistent – the person's *potential score*.

The PPT strategy for obtaining each person's potential score is as follows in an experiment where a stimulus appears in one of two intervals and the participants attempt to choose the correct interval. In this paradigm, the participant can choose one or the other interval and the correct answer can be one or the other interval. Thus, each person's performance across the trials can be summarized by a 2 x 2 table where *a*, *b*, *c*, and *d* refer to the cell frequencies, r_1 and r_2 refer to the row frequencies, and c_1 and c_2 refer to the column frequencies (see Table 1). This table can be converted into a correlation coefficient.

Given that a consistency coefficient has been obtained, the correlation coefficient obtained from the performance table can be adjusted for attenuation due to inconsistency, via the famous formula that originally was derived from classical true score theory (e.g., Spearman, 1904) but also can be derived from more modern theories (see Allen & Yen, 1979; Cohen & Swerdlik, 1999; Crocker & Algina, 1986; Gulliksen, 1987; Lord & Novick, 1968 for reviews). In the usual PPT paradigm, the correlation coefficient derived from the actual performance table is corrected by the consistency coefficient obtained across two blocks of trials to obtain the corrected or "potential" correlation coefficient designated by *R*. In turn, instantiating the corrected correlation coefficient into

PPT equations (see Trafimow & Rice, 2008) estimates the cell frequencies that would be obtained with perfect consistency (i.e., no randomness). Finally, a person's potential score – the estimate of how well the person would have performed in the absence of randomness – is simply the sum of the corrected cell frequencies for hits and correct rejections divided by the number of trials.

Worked PPT Example

Suppose that a person has the following cell frequencies: a = 15, b = 10, c = 10, and d = 15, and so all margin frequencies = 25. The observed score is $\frac{a+d}{a+b+c+d} = \frac{15+15}{15+10+10+15} = .60$ which can be converted into a correlation coefficient as follows: $r = \frac{|ad-bc|}{\sqrt{R_1R_2C_1C_2}} = \frac{(15)(15)-(10)(10)}{\sqrt{(25)(25)(25)(25)}} = .20$. Assuming that there are two blocks of trials, suppose that the person's correlation across blocks – that person's consistency score – equals .55. In that case, it is possible to correct the observed correlation coefficient as follows:

$$corrected\ or\ potential\ correlation = R = \frac{r}{\sqrt{consistency\ coefficient}} = \frac{.20}{\sqrt{.55}} = .27.$$

To obtain the potential cell frequency for cell *A*, one uses the following equation (see Trafimow & Rice, 2008 for a proof): $A = \frac{R\sqrt{R_1R_2C_1C_2}+C_1R_1}{(R_1+R_2)} = \frac{.27\sqrt{(25)(25)(25)(25)}+(25)(25)}{(25+25)} = 15.875$. It follows that the potential frequency for cell B = 25 – 15.875 = 9.125. By similar reasoning, the potential cell frequencies for cells *C* and *D* are 9.125 and 15.875, respectively. Thus, the potential score is $\frac{A+D}{TOTAL} = \frac{15.875+15.875}{50} = .635$, which exceeds the observed score of .60.

To see the importance of consistency, suppose that we had used .15 for the consistency coefficient in the example instead of .55. In that case, the potential correlation would have been .52 (rather than .27), and the potential cell *A* frequency would have been 19. Thus, the potential score would have been .76 (instead of .635). Or, if the consistency coefficient in the example had been set

at .04, the potential cell *A* frequency would have been 25, and the potential score would have been 1.0.

Before continuing, we stress that observed scores and consistency scores are subject to sampling error. Consequently, potential scores are *estimates* of what would happen in the absence of randomness rather than representing absolute truth. This fact implies the interesting consequence that potential scores can exceed 100%. As an illustration, suppose that people would be perfect detectors in the absence of randomness; the population potential score, both within and between people, equals 100%. In addition, suppose that it were possible to take an infinite number of independent samples of *N* trials for a single person. In that case, due to sampling error, we would expect potential scores to be less than 100% for some of the samples but to exceed 100% for other samples, with the mean at 100%. Alternatively, still assuming that all people are perfect detectors in the absence of randomness, suppose that we obtained a single sample for each person, but for an infinite number of people. In that case, we would obtain potential scores that are less than 100% for some people but greater than 100% for other people, with the mean again at 100%.

Predictions of SDT

All that remains is to determine the corrected performance of an observer acting according to the dictates of SDT. In a 2IFC detection task the SDT observer collects a percept for each interval and responds by identifying the interval associated with the largest percept (Green & Swets, 1966). This response strategy is the optimal one if the mean of the signal distribution is greater than the mean of the noise distribution (that is, if *d'* is positive). If the random noise is eliminated in the distributions of signal and noise percepts, then the percept associated with the signal stimulus will always be greater than the percept associated with the noise stimulus, and performance will be perfect. A formal proof of this result can be obtained via the following website:

http://justinmacdonald.net/publications. The website also presents computer simulations that back up the proof by showing that performance, corrected for attenuation due to randomness by PPT, was essentially perfect when averaged across 1,000,000 observers.

In summary, SDT predicts perfect corrected performance if sensitivity is positive. Consequently, researchers can now perform SDT experiments, test the assumption, correct for randomness with PPT, and then determine if the potential scores average out to 100%. If so, this finding would support the SDT assumption that randomness is the only factor that contributes to imperfect performance. But if the average of the potential scores remains significantly lower than perfect, it would show that there is a systematic factor of importance that is thus far missing from SDT.

Experiment 1A

We used an auditory detection task in Experiment 1A. The main difference between Experiment 1A and many such experiments is that we used two blocks of trials, with stimuli of a variety of auditory frequencies, so that trials could be paired across blocks based on frequencies. Consistency coefficients were based on that pairing.

Method

Participants. Twenty-eight undergraduates from the New Mexico State University Psychology Department Subject Pool served as participants for course credit. They exhibited normal hearing (thresholds less than or equal to 25 dB HL) and normal or corrected-to-normal vision. Hearing was tested in both ears at octave intervals from 250 to 8000 Hz.

Stimuli. Target stimuli were 200-ms sinusoids with 20-ms raised cosine onset and offset ramps and were sampled at 44.1 kHz. Stimuli ranged in frequency from 500 to 1000 Hz in 20-Hz intervals and were 40 dB(A) in intensity measured at the output of the headphone. The background

noise stimulus was Gaussian white noise that was 54 dB(A) in intensity measured at the output of the headphone.

Procedure. After signing an informed consent document, participants completed a hearing screening to determine their eligibility for the study. Subsequently, participants were seated in front of an experiment workstation that consisted of a personal computer, a mouse, a 20.1-inch diagonal LCD monitor operating at 1680x1050 resolution, and a pair of AKG K240 circumaural headphones attached to the output of the computer's sound card. The experiment was conducted using custom software. Each participant started by reading an instruction screen that explained the procedure. Participants clicked a button to start each trial, which consisted of a 2IFC detection task. The target stimulus was presented in the left headphone and the background noise was presented in both headphones continuously throughout both intervals in the trial. A visual indicator was presented on the screen during each listening interval, and a 500-ms gap was inserted between intervals. After the second interval, participants were prompted to click one of two buttons to indicate which interval contained the target tone. The experiment consisted of two blocks of 52 trials each (26 target frequencies by 2 intervals) for a total of 104 trials. The order of the stimuli within a block was completely randomized. In addition, the random orders of the block 1 and block 2 stimuli differed for each participant. Participants were allowed to take breaks as necessary between trials and blocks. The experimental session, including paperwork, the hearing screening, and the main experiment took about 25 minutes to complete.

Results

Consistent with the simulations presented earlier, participants with non-positive consistency or *d'* estimates were eliminated from the analysis; this results in the removal of three participants. The mean observed proportion correct was .89 [SD = .01; CI(95%) = .89 ± .0008]. The mean

potential score (corrected for attenuation) was .97 [SD = .05; CI(95%) = .97 ± .004]. Both confidence intervals do not overlap with unity.

At the risk of "data mining" and capitalizing on chance, we noticed that there seemed to be a "break point" at observed performance around 90%. So we divided the sample into two groups – those with observed scores lower than 90% and those with observed scores at or above 90%. There were twelve participants in the lower group and the mean potential score was .93 [SD = .06; CI(95%) = .93 ± .011]. In contrast, there were thirteen participants in the upper group and the mean potential score was 1.00 [SD = .01; CI(95%) = .1.00 ± .002]. The two confidence intervals do not overlap. Figure 1 presents histograms indicating potential scores in both groups and shows that the central tendency for potential scores to be less than perfect in the lower group was not solely attributable to a few outlying low scores.

In summary, participants with observed scores below .90 had potential scores well below perfect (.93 vs. 1.00), thereby indicating that there is a systematic component to the performance of the decision task that SDT does not take into account. But in participants whose observed scores exceeded .90, the SDT prediction is confirmed with impressive precision (1.00 vs. 1.00).

Experiment 1B

Experiment 1A possessed two obvious limitations. First, the sample size was small. Second, the successful split of observed scores below or not below .90 may have capitalized on chance. Experiment 1B addressed both limitations.

Method

Experiment 1B was exactly the same as Experiment 1A but with an independent sample of 89 participants.

Results

Seven participants failed to meet the requirements for inclusion and were removed. The mean observed score was .90 [SD = .10; CI(95%) = .90 ± .002] and the mean potential score was .97 [SD = .07; CI(95%) = .97 ± .002]. Both confidence intervals do not overlap with unity.

As in Experiment 1A, we divided the sample into two groups – those with observed scores lower than .90 and those with observed scores at or above .90. There were 26 participants in the lower group and the mean potential score was .91 [SD = .11; CI(95%) = .91 ± .009]. In contrast, there were 56 participants in the upper group and the mean potential score was .99 [SD = .02; CI(95%) = .99 ± .0007]. As in Experiment 1A, the two confidence intervals do not overlap. Figure 2 presents histograms indicating potential scores in both groups and shows that the central tendency in the lower group was not attributable solely to a few outlying low scores.

Experiment 2

Experiments 1A and 1B pertained to auditory detection and showed that there is a non-random component not accounted for by SDT, at least for participants with observed scores less than .90. Experiment 2 attempts to generalize to visual detection.

As in Experiments 1A and 1B, it was necessary to pair data across blocks of trials. Instead of using auditory frequencies, as in the foregoing experiments, we used 25 different levels of contrast in Experiment 2.

Method

Participants. Twenty-six participants from a large southwestern university participated in the experiment for partial course credit. The mean age was 20.42 (*SD* = 6.79). All participants were tested for normal or corrected-to-normal vision and colorblindness.

Materials and Stimuli. The experimental display was presented on a Dell PC with a 22" monitor using 1024 x 768 resolution with a 65 Hz refresh rate. All experimental stimuli were presented via E-prime 1.1.

Twenty-five display images were created in Photoshop CS2. Each image had a grey background (color code: 7b7b7b). There was a pink dot in the middle of each image measuring 1 degree in visual angle, and varying in contrast to the grey background in such a way as to provide 25 levels of difficulty in detecting the dot. The most difficult image was made up using the following attributes: H = 0; S = 3%; B = 50%; R = 127; G = 123; B = 123; L = 52; A = 2; B = 1; C = 51%; M = 46%; Y = 45%; and K = 9%. The color code was 7f7b7b. The least difficult image was made up using the following attributes: H = 0; S = 25%; B = 64%; R = 163; G = 123; B = 123; L = 56; A = 16; B = 6; C = 36%; M = 53%; Y = 44%; and K = 5%. The color code was a37b7b. A 2IFC method was used to present the stimuli on a Dell E2209Wc monitor, whereby the pink dot was randomly placed in either the first or second interval and participants were asked to pick which interval they detected the dot. There was a blank grey display in the opposing interval. Thus, there were a total of 50 trials per block (25 target displays in the first interval and 25 target displays in the second interval) and two blocks of trials for 100 total trials. As in Experiment 1, trials were randomly ordered within blocks and persons, and the random order differed between the two blocks.

Procedure. After signing a consent form, participants sat comfortably in front of the experimental display. Instructions were given onscreen and participants were free to ask questions until they were confident of the task. They then pressed a key to begin the experiment. The 50 randomly ordered trials, in each block of trials, began with a fixation screen, whereby a black cross was located in the middle of the grey background for 1000 ms. Following this, a mask screen was presented for 200 ms, followed by the first interval display for 1000 ms, followed by another mask

screen for 200 ms, followed by the second interval display for 1000 ms. Lastly, a Choice display asked participants to press F if they detected the dot in the first interval display and J if they detected the dot in the second interval display.

Results and Discussion

Based on the same criteria used in the previous experiments, three participants were removed prior to analysis. The mean observed score was .78 [SD = .08; CI(95%) = .78 ± .006) and the mean potential score was .90 [SD = .09; CI(95%) = .90 ± .007]. Both confidence intervals do not overlap with unity. There were too few participants with observed scores in excess of .90 to make splitting as in the previous experiments worthwhile. Although Figure 3 shows that there was an outlying low potential score, the strong majority of the potential scores failed to reach perfection. Even with the extreme low score removed, the mean potential score nevertheless was only .91 [SD = .06; CI(95%) = .91 ± .005]. Thus, as was true in the low groups in the auditory experiments, there was a systematic component in the visual experiment that SDT does not take into account.

Based on the findings obtained in Experiments 1A and 1B, one might suspect that consistency scores and potential scores in Experiment 2 might correlate. For example, it could be that people who use a better detection strategy also use it more consistently. We tested for this and did not obtain a significant correlation (r = .07, p = .71). In one sense, the lack of a correlation is a limitation because it does not support a potentially interesting hypothesis. In another sense, the lack of a correlation is convenient because it suggests that the consistency scores were not biased, though we hasten to add that the presence of a correlation would not necessarily indicate bias.

Obviously, performance tended to be better in Experiments 1A and 1B than in Experiment 2. This should not be taken to indicate that detection via vision is more difficult, in general, than is detection via audition. It seems more likely that our particular auditory detection task was too easy,

on average, which is what necessitated the split for observed scores at or above .90 whereas our vision detection task was better calibrated to the abilities of the participants.

Supplementary Results, All Experiments

Though tangential to our main analyses, there are three additional issues that warranted supplementary analyses. The first set of analyses pertained to learning across blocks of trials. Because the responses were dichotomous (i.e., the participants chose one interval or the other), learning across blocks would cause consistency coefficients to be underestimates, thereby causing PPT to overcorrect, and so potential scores would be overestimates. Because the potential scores were too low, rather than too high, an argument that we overcorrected strengthens, rather than weakens, our case that SDT is missing a systematic factor of importance. In addition, Trafimow and Rice (2011) showed that a great deal of learning is necessary to substantially influence potential scores. Nevertheless, we tested directly for learning across blocks of trials.

The second set of analyses pertained to the effect of each trial on the subsequent one. If each trial influenced the subsequent one, this effect would be a candidate for the systematic factor missing from SDT. In contrast, if this is not so, it would suggest that researchers should look elsewhere.

The third set of analyses addressed time-order error (e.g., Hairston, 2007; Hellström, 1985; Yeshuran, Carrasco, & Maloney, 2008). Participants may have been systematically biased to choose one of the intervals more often than they should have. If so, the proportion correct when Interval 1 was chosen should have differed from the proportion correct when Interval 2 was chosen.

Learning Across Blocks of Trials

In Experiment 1A, the mean proportions correct were .89 and .91 in blocks 1 and 2, respectively. In Experiment 1B, the mean proportions correct were .90 and .94 in blocks 1 and 2, respectively. In Experiment 2, the mean proportions correct were .78 and .77 in blocks 1 and 2,

respectively. The difference in Experiment 1B was statistically significant but it was not statistically significant in Experiments 1A and 2. In general, differences across blocks were too small and unreliable to plausibly explain our main findings (see Trafimow & Rice, 2011 for a detailed discussion of this issue).

Sequential Effects

We assessed sequential effects in two ways, via correlations and via clustering analyses. To analyze sequential effects with correlations, we formed an *N* by *N*+1 matrix for each participant to assess whether the response on trial *N* predicted the response on trial *N*+1. To analyze sequential effects with clustering, we computed the adjusted ratio of clustering index (ARC), proposed by Roenker, Thompson, and Brown, (1971), for each participant. When ARC equals zero it indicates chance clustering and when ARC equals 1 it indicates perfect clustering.

Mean sequential correlations were -.04, .06, and -.04 for Experiments 1A, 1B, and 2, respectively. Mean ARC scores were .02, .11, and -.01 for Experiments 1A, 1B, and 2, respectively. None of these effects were statistically significant.

Yet another way to assess sequential effects is to search for runs in the data using the Wald-Wolfowitz runs test (Wald & Wolfowitz, 1940). This test determines whether successive responses are independent by counting up the number of runs (identical responses in sequence) in a set of responses and comparing this count to the number of runs expected if successive responses were independent. Significant results in either direction indicate sequential effects: positive z-scores indicate a number of runs greater than chance whereas negative *z*-scores indicate fewer than chance runs. We tested this possibility by conducting the Wald-Wolfowitz Runs Test on the individual data for each participant. In the absence of systematic sequential effects, we would expect 5% of the *z*-scores to be statistically significant (in other words, the Type I Error rate is equal to our significance

level, or 0.05). We will call these "extreme"), and the question is whether significantly more than 5% of the *z*-scores are extreme. We found that 14%, 2%, and 17% of the *z*-scores were extreme in Experiments 1A, 1B, and 2, respectively, so there was no clear pattern of a systematic tendency to have more than 5% extreme *z*-scores, across the experiments. Therefore, it is not plausible that our findings, across the three experiments, can be explained via sequential effects. But to be absolutely sure, we performed PPT analyses with extreme participants taken out, and the findings did not change substantially.

Time-Order Error

Participants could have been biased to choose inappropriately one of the time intervals over the other. To test this possibility, we analyzed the proportion of correct responses given either that Interval 1 or Interval 2 was chosen using the method given in Yeshuran et al. (2008). In Experiment 1A, the mean proportions were .896 and .909 ($X^2(1) = 1.26$, $p = .26$), respectively; in Experiment 1B, they were .921 and .900, respectively ($X^2(1) = 12.72$, $p < .001$); and in Experiment 3 they were .738 and .731, respectively ($X^2(1) = 0.21$, $p = .65$). Thus, the difference was extremely small in all three experiments, the difference was in a different direction in Experiment 1A than in the other two experiments, and it was statistically significant only in Experiment 1B. Therefore, no plausible case can be made that time-order error accounts for our findings.

Discussion

The auditory detection experiments suggest that no additional systematic model components are necessary for SDT to account for participants with observed performance at or above .90. For these participants, it appears that almost all errors really were due to random noise, as the potential scores near perfection indicate. In contrast, for participants with observed scores below .90, potential scores tended to be well below perfection, thereby indicating the existence of one or more systematic

factors underlying their failures that is not included in the SDT model. The existence of unmodeled systematic factors is not restricted to auditory detection tasks; the vision detection experiment also resulted in potential scores well under perfection.

There are potential counterarguments to address. For example, learning could have served as the unmodeled systematic factor. But as we pointed out earlier, there are two problems with this argument. Most important, the supplementary analyses contradict that significant learning took place; performance was approximately equal in both blocks of trials in all of the experiments. Second, though rendered less important by the supplementary analyses, the dichotomous nature of the task implies another problem with invoking learning as the missing systematic factor. Consider that had the trials been continuous, learning could have occurred without influencing the consistency coefficients, if a constant was added to each item. But when the trials are dichotomous, learning forces consistency coefficients to decrease (Trafimow & Rice, 2011). In turn, if consistency coefficients are decreased, overcorrection results, and potential scores are too high. Recalling that the present issue for SDT is that the potential scores are too low rather than too high, the implication is that learning fails to plausibly explain the findings.

One might also argue that PPT produces a biased estimate of corrected performance due to its reliance on the classic correction for attenuation formula. Zimmerman and Williams (1997) performed computer simulations and reported that the attenuation formula produces overcorrected estimates of true performance in some situations, most notably when consistency (reliability) is low. We find this argument ineffective for the following reasons, any one of which is sufficient. First, as we pointed out in connection with the learning argument, overcorrection, if it occurred, merely makes our argument stronger rather than weaker; again, from the SDT perspective, potential scores were too low rather than too high. Second, in the direct PPT simulations that Trafimow and Rice

(2011) performed, PPT's estimates were demonstrated to be effectively unbiased estimates of true corrected performance. Third, the present simulations using a SDT paradigm similarly supported a lack of overcorrection. Fourth, there is much disagreement about the validity of the conclusions of Zimmerman and Williams (see Hunter & Schmidt, 2004 for a review).

A more general criticism, from the point of view of the formula for correcting for attenuation, is that there is sampling error in the consistency coefficients as well as in the observed performances. Consequently, PPT could be plagued with the problem of variance. However, both the simulations performed by Trafimow and Rice (2011), and those performed here, resulted in low standard deviations given the numbers of trials per block that we used in the present experiments. Relatedly, there is the issue of whether we had a sufficient number of participants for low standard errors. In fact, our standard errors were very low, thereby resulting in very small confidence intervals in all three experiments. Consequently, concerns related to variance cannot be used validly to dismiss our findings.

Yet another potential criticism is that others have published objections to SDT and so the present findings constitute only an incremental advance. However, PPT allows for an assessment of SDT that cannot be obtained through other approaches. Any model (SDT included) will inevitably fail to obtain a perfect fit to empirical data. The task of the explanatory modeler is to implement theory-driven modifications to the model to account for a greater proportion of variability in the data. In the case of SDT, there are many potential modifications to the model, some systematic in their effects and others random. The application of PPT to SDT allows the modeler to determine whether there are systematic effects missing from the model, thereby potentially eliminating a category of model modifications from consideration. For example, the analyses presented in this paper demonstrate that additional systematic components are unnecessary to account for the

performance of participants with observed scores greater than or equal to 90%. Given that SDT fails to provide a perfect fit to the data for these subjects, therefore, the PPT analysis demonstrates that the addition of random model components (e.g., criterial noise, additional variability in the stimulus distributions, or any other nonsystematic component) will be sufficient to model these data accurately. However, the PPT analyses lead to the opposite conclusion for participants whose observed scores were below 90%: the addition of any number of nonsystematic components to the SDT model will not lead to an accurate representation of the decision-making process.

Systematic Error

Figures 1-3 make it obvious that there is at least one source of systematic (nonrandom) error that prevents perfect performance. We stress that although PPT can identify systematic error, it cannot identify precise mechanisms by which such error comes about. Therefore, it is necessary to speculate about these mechanisms, and eventually perform appropriate experimental manipulations that can be used in conjunction with PPT analyses to test them. The present findings at least help to eliminate some possibilities.

Consider that researchers often split the decision process into two parts: the collecting of information about the stimulus to generate a percept (the sampling procedure) and the mapping of percepts to responses (the decision procedure). Systematic error could be introduced during either or both parts of the decision-making process. Since we did not collect information on the sampling behavior of our participants we cannot shed any light on possible sources of systematic error in the sampling process. We can, however, speculate about sources of systematic error in the decision process. Adopting the SDT framework, suppose that each trial causes participants to shift their criterion levels and this shift, in turn, influences responses on the following trial. Or imagine a priming effect whereby the mere fact of having responded a certain way on trial *N* increases the

probability of a similar response on trial *N*+1. Or perhaps it is simply easier to press the same key on trial *N*+1 as was pressed on trial *N*. Had such sequential processes occurred, we should have been able to discern their effects by computing sequential correlations, ARC scores, or the Wald-Wolfowitz runs analyses. However, the analyses did not support the plausibility of these processes. In addition, the sequential analyses render classes of models that depend on sequential effects in response selection as inadequate explanations for our findings (e.g., Brown, Steyvers, & Hemmer, 2007; Treisman & Williams, 1984).

Because systematic sequential effects were insufficient to plausibly explain our main findings, it seems worthwhile to consider random sequential effects. For example, perhaps the decision criterion level changed randomly on each trial. A substantial amount of random criterion shifting could have had deleterious effects on observed performance. Could it have influenced potential scores? This is not plausible because potential scores should have corrected for randomness of any sort, including random criterion shifts, as we saw in the upper groups of participants in Experiments 1A and 1B.

Given that others have obtained sequential effects in the decision process (Treisman & Williams, 1984; Taylor & Lupker, 2001; Stewart, Brown, & Chater, 2002), why were they missing in the present experiments? One possible answer to this question lies in considering that in some SDT experiments, the same stimulus is presented (or not presented) repeatedly. But in our experiments, the stimuli differed with respect to frequency (Experiments 1A and 1B) or contrast (Experiment 2) and so they were not the same. This dissimilarity across trials might have been a factor in prevented us from obtaining the sequential effects that others have obtained.

Although our main goal was to test whether systematically including additional sources of randomness could render SDT as an adequate account of the perceptual process, the present research

suggests an additional gain that might be equally important. Specifically, consider again not only that we obtained no systematic sequential effects, but we also obtained no learning across blocks, nor were there discernible effects pertaining to time order error. In short, the biases others have uncovered in 2IFC paradigms seem to be absent in the present version of that paradigm. For researchers who wish to study biases, our version of the 2IFC paradigm would be a poor choice because it eliminates the phenomena of interest. But if biases are not of primary interest, and merely add complexity that is difficult for researchers to handle, the present version of the 2IFC paradigm seems idea for eliminating that complexity. Even those researchers who do not wish to use PPT to address the issue of randomness in responding likely will find our paradigm nevertheless preferable to other 2IFC paradigms, because our version removes the layers of biases that otherwise would constitute obstructions that are difficult or impossible to surmount. Thus, with the usual biases removed, our version of the 2IFC paradigm allows researchers to penetrate directly to the heart of that which they wish to investigate, though with an important limitation.

The limitation is that because our paradigm uses many frequencies (or contrasts) and the present analyses were carried out across all of them, there is no way for researchers to obtain definitive information about stimuli at any one particular frequency (or contrast). We see two levels at which this limitation can be addressed, depending on the aims of the researcher; these are the paradigmatic and theoretical levels. First, it might be possible to change the paradigm so that if the researcher is interested in particular frequencies, these can be presented more often than frequencies in which the researcher is not interested. How many repetitions of particular frequencies would it take for the usual biases to emerge? At present, we have no idea. But we see nothing to prevent researchers from performing systematic research to find out. If it turns out that it is possible to have many repetitions without biases emerging, we would have a paradigmatic solution to the limitation.

An alternative paradigmatic solution would be if it turns out to be possible to present many blocks of trials without biases emerging. In this case, even if one could not present a disproportionately high number of trials at the frequencies of interest, it might be possible to have a large number of blocks, so that the frequencies of interest are represented a sufficient number of times for analysis. It also might be possible to combine aspects of both solutions (e.g., present the frequencies of interest a disproportionate number of times, but not so disproportionately that biases emerge, and also present multiple blocks of trials, but not so many blocks that biases emerge).

At the theoretical level, we would argue that there is no reason why perception or decision theories have to be at the level of particular stimuli of interest. If researchers were to propose general theories that apply to classes of stimuli, rather than proposing models that apply to particular stimuli, there would be less reason for concern with sensitivity at the level of a particular type of stimulus, and so the foregoing limitation would be less problematic. From our perspective, although we realize that researchers are used to measuring sensitivity, the SDT notion of *d'* is incorrect because it assumes that randomness is the only relevant factor, an assumption that all three experiments disconfirm. We would prefer a theoretical shift whereby researchers would become concerned with potential scores rather than with *d'*, in which case the limitation we identified would again be reduced in importance.

In addition to eliminating biases, our version of the 2IFC paradigm confers an additional advantage of increased ecological validity over other 2IFC paradigms. In normal living, people are unlikely to hear the same sounds, or see the same sights, repeatedly except under unusual circumstances. In most circumstances, people hear sounds with a variety of frequencies, see sights with a variety of contrasts, and generally experience a larger degree of stimulus variation than in usual 2IFC paradigms. We hasten to add that we are not arguing that ecological validity should be

the main focus of perception researchers. On the contrary, we believe theory or model testing to be extremely important. However, if all else is equal, we also believe that ecological validity should at least be a consideration. And as we pointed out in the previous paragraph, all else is not equal; our version of the 2IFC paradigm has important advantages from an internal validity perspective too, in that it allows researchers to remove the layers of bias that provide fodder for alternative explanations.

It is possible to speculate further about the systematic component of perception uncovered here. It might be that high performers (observed score $\geq .90$) approach the SDT ideal where the only contributing factor to insensitivity really is random noise, and hence the correction for attenuation works as it is supposed to and renders potential scores near perfection. In contrast, lower performers might use a single systematically flawed strategy, switch between multiple strategies where at least one of them has a systematic flaw, or otherwise undergo a systematically flawed decision process. For example, perhaps low performers attempt to employ conscious strategies that actually interfere with the normal process of perception that depends mostly on unconscious processes. Obviously, future research is needed to test these possibilities but we believe that PPT will play an important role in performing this research.

Although we focused on SDT because of its prominence, there are many models that feature the notion of random noise as the main factor responsible for the lack of perfect responding. The present findings constitute a problem for these as well as for SDT. Lacking sufficient space to discuss these models properly, we merely note that some kind of systematic component will have to be added to them and that this systematic component cannot be redundant with the decision criterion notion in SDT.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Balakrishnan, J. D. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors*, *40*, 601-623.

Balakrishnan, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods*, *3*, 68-90.

Balakrishnan, J. D. (1999c). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception & Performance*, 25, 1189-1206.

Balakrishnan, J. D., & MacDonald, J. A. (2008). Decision criteria do not shift: Reply to Mueller and Weidemann (2008). *Psychonomic Bulletin & Review, 15,* 1022-1030.

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

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science, 18,* 40 - 45.

Cohen, R. J., & Swerdlik, M. E. (1999). *Psychological testing and assessment: An introduction to tests and measurements* (4th ed.). Mountain View, CA: Mayfield.

Crocker, L., & Algina, J. (1986). *Introductions to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* New York: Academic Press.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 25,* 500-513.

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

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale: NJ: Lawrence Erlbaum Associates, Publishers.

Hairston, I. S., & Nagarajan, S. S. (2007). Neural mechanisms of the time-order error: An MEG study. *Journal of Cognitive Neuroscience*, *19*, 1163-1174.

Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. Psychological Bulletin, 97, 35–61.

Hunter, J.E. & Schmidt, F.L. (2004). Methods of meta-analysis: Correcting error and bias in research synthesis (2nd ed.). London: Sage Publications.

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers, Volume 1* (J. Worrall & G. Currie, Eds.). Cambridge, United Kingdom, UK: Cambridge University Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Mickes, L., Wixted, J. T., & Waisd, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulleting & Review, 14,* 858-865.

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

Ratcliff, R., Shue, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518-535.

Rice, S., Geels, K., Hackett, H., Trafimow, D., McCarley, J. S., Schwark, J. & Hunt, G. (2012). The harder the task, the more inconsistent the performance: A PPT analysis on task difficulty. *Journal of General Psychology, 139*(1), 1-18.

Rice, S., Geels, K., Trafimow, D. & Hackett, H. (2011a). Our students suffer from both lack of knowledge and consistency: A PPT analysis of test-taking. *US-China Education Review, 1*(6), 845-855.

Rice, S. & Trafimow, D., & Hunt, G. (2010). Using PPT to analyze sub-optimal human-automation performance. *Journal of General Psychology*, *137*, 310-329.

Rice, S., Trafimow, D., Keller, D., Hunt, G. & Geels, K. (2011b). Using PPT to correct for inconsistency in a speeded task. *The Journal of General Psychology, 138*, 12-34.

Roenker, D. K., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, *76*, 45-48.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72-101.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, memory, and Cognition, 28,* 3-11.

Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., & Freeman, B. A. (1979). Assessment of diagnostic technologies. *Science, 205,* 753-759.

Swets, J. A., Tanner, W. P., Jr, & Birdsall, T. G. (1961). Decision Processes In Perception. *Psychological Review*. *68*(5), 301-340.

Taylor, T. E., & Lupker, S. J. (2001). Sequential effects in naming: A time-criterion account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 117-138.

Trafimow, D., & Rice, S. (2008). Potential Performance Theory: A general theory of task performance applied to morality. *Psychological Review, 115*(2), 447-462.

Trafimow, D., & Rice, S. (2009). Potential Performance Theory (PPT): Describing a methodology for analyzing task performance. *Behavior Research Methods, 41*(2), 359-371.

Trafimow, D., & Rice, S. (in press). Using a sharp instrument to parse apart strategy and consistency: An evaluation of PPT and its assumptions. *Journal of General Psychology*.

Trafimow, D., Hunt, G. Rice, S. & Geels, K. (2011). Using potential performance theory to test five hypotheses about meta-attribution. *Journal of General Psychology, 138*, 1-13.

Treisman, M. & Williams, T. C. (1984). Theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68-111.

Wald, A. (1950). *Statistical decision functions*. Oxford, England: Wiley.

Wald, A. and Wolfowitz, J. (1940), "On a test whether two samples are from the same population," *Annals of Mathematical Statistics, 11,* 147-162.

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316-347.

Wixted, J. T. (2007). Spotlighting the probative findings: Reply to Parks and Yonelinas (2007). *Psychological Review, 114*, 203-209.

Yeshuran, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*, 1837-1851.

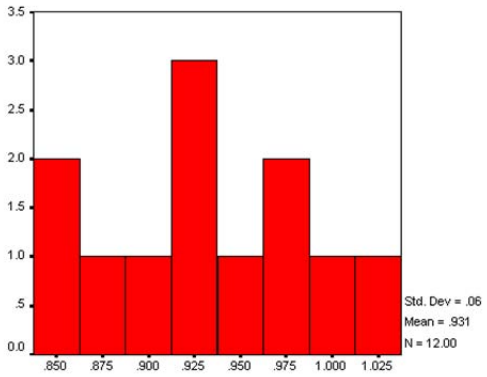
Zimmerman, D. W., & Williams, R. H. (1997). Properties of the Spearman correction for attenuation for normal and realistic non-normal distributions. *Applied Psychological Measurement*, *21*, 253-270.

Table 1. A 2 (individual choice) x 2 (correct choice) frequency table where *a*, *b*, *c*, and *d* indicate the actually observed cell frequencies, r_1 and r_2 are the row frequencies, and c_1, and c_2 are the column frequencies.

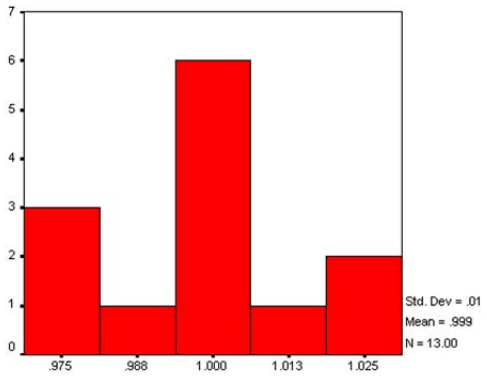
	Individual's Choice		
Correct Choice	Interval 1	Interval 2	Row Margin
Interval 1	*a*	*b*	r_1
Interval 2	*c*	*d*	r_2
Column Margin	c_1	c_2	

Histograms for Potential Scores in Experiment 1A

The Under .90 Group

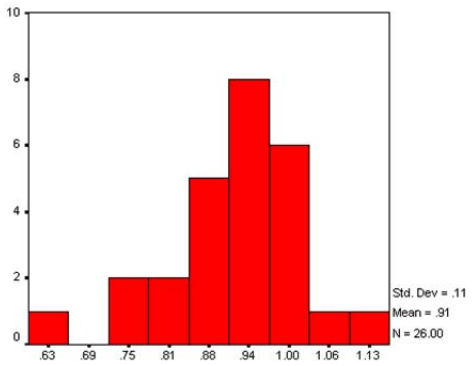


The .90 and Over Group

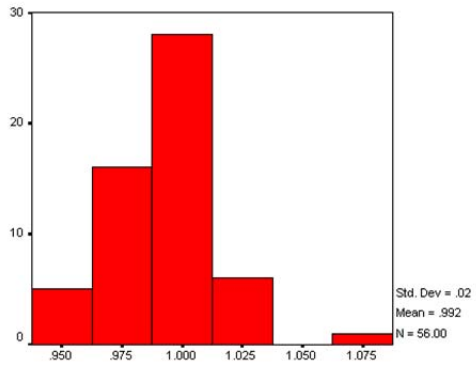


Histograms for Potential Scores in Experiment 1B

The Under .90 Group



The .90 and Over Group



Histograms for Potential Scores in Experiment 2