Perception & Psychophysics
2006, 68 (4), 643-654

Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'

MICHAEL F. VERDE
University of Plymouth, Plymouth, England

and

NEIL A. MACMILLAN and CAREN M. ROTELLO
University of Massachusetts, Amherst, Massachusetts

Signal detection theory offers several indexes of sensitivity (d', A_z, and A') that are appropriate for two-choice discrimination when data consist of one hit rate and one false alarm rate per condition. These measures require simplifying assumptions about how target and lure evidence is distributed. We examine three statistical properties of these indexes: accuracy (good agreement between the parameter and the sampling distribution mean), precision (small variance of the sampling distribution), and robustness (small influence of violated assumptions on accuracy). We draw several conclusions from the results. First, a variety of parameters (sample size, degree of discriminability, and magnitude of hits and false alarms) influence statistical bias in these indexes. Comparing conditions that differ in these parameters entails discrepancies that can be reduced by increasing N. Second, unequal variance of the evidence distributions produces significant bias that cannot be reduced by increasing N—a serious drawback to the use of these sensitivity indexes when variance is unknown. Finally, their relative statistical performances suggest that A_z is preferable to A'.

Performance in a discrimination task depends on two factors: the available evidence and the rules whereby that evidence is applied to a decision. Suppose that an observer's goal is to discriminate between targets and lures (for example, between target-present and target-absent trials in visual detection, or between old and new items in recognition memory) and that an experimental manipulation increases the rate of correct target identifications. Has the manipulation led subjects to become more sensitive in their discriminations, or has it only made them more willing to claim that a target was present on any given trial? The ability of signal detection theory (SDT) to model this critical distinction between sensitivity and response bias has made it an invaluable tool in psychophysics, perception, memory, and other domains (Green & Swets, 1966; Macmillan & Creelman, 2005).

A useful construct motivated by SDT is the receiver operating characteristic (ROC), which plots the hit rate (H) versus the false alarm rate (F) at different degrees of response bias as sensitivity is held constant. ROCs can be constructed efficiently with a rating design: For each test probe, the subject responds with some level of confidence that it is a target or lure, each level corresponding to a different degree of response bias. The ROC connects (F, H) points calculated by cumulating response proportions from the most conservative to most liberal decision rules. Discrimination is accurate to the extent that the hit rate exceeds the false alarm rate, and as the difference between hits and false alarms increases, the ROC moves closer to the upper left corner (Figure 1). An important virtue of the ROC function was demonstrated by Green (1964): The area under the curve equals the proportion correct obtained by an unbiased observer in a two-alternative forced choice task. This area is thus a pure measure of sensitivity, uncontaminated by response bias. Its value can be estimated by using the ROC points to form a series of trapezoids and adding their areas (Pollack & Hsieh, 1969).

Many researchers find it difficult or impractical to gather rating data or to manipulate response bias in multiple experimental conditions, as is required to obtain ROCs. A (very common) alternative task provides subjects with only two response choices, "target" or "lure," and the resulting data consist of a single hit and false alarm rate per condition. Such data constitute a single point on the ROC curve. Any *single-point* sensitivity measure has implications for the overall shape of the ROC curve on which it lies and the form of the underlying distribution of evidence. A number of measures can be applied to data from the two-response discrimination task, and in this article we compare three of them: d', A_z, and A'.

This research was supported by National Institutes of Health Research Grant MH60274-02 to C.M.R. and N.A.M. We are grateful to Michael Hautus and an anonymous reviewer for their helpful comments on an earlier draft of this article. Correspondence concerning this article should be addressed to M. F. Verde, School of Psychology, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK (e-mail: michael.verde@plymouth.ac.uk).

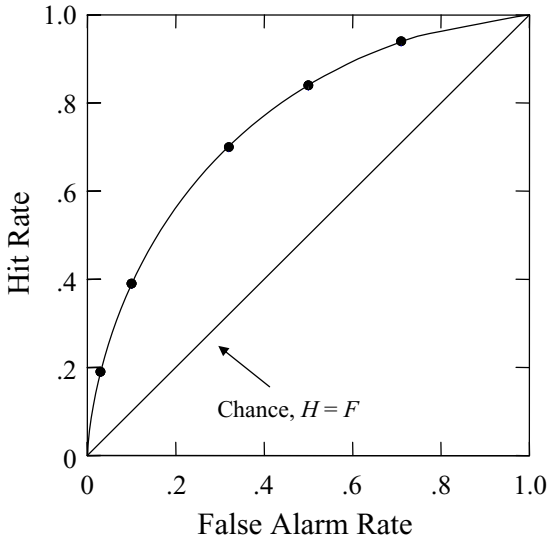

Figure 1. A hypothetical ROC. Each point on the curve represents a different level of confidence.

The first two of these indexes are derived from explicit SDT assumptions. The distributions for targets and lures are Gaussian with equal variance, and d' is defined as the standardized distance between the means. The statistic A_z equals the area under the (equal-variance Gaussian) ROC curve that contains (F, H), and is a monotonic transformation of d'. Many empirical studies support the Gaussian assumption, but many fewer are consistent with equal variance (Swets, 1986).

A heavily used alternative to d' (or A_z) is A', a geometric approximation of the area under the ROC curve (Pollack & Norman, 1964). The popularity of A' is due in large part to the claim (by its inventors and many users since) that it is nonparametric, although this claim has been shown to be false (Macmillan & Creelman, 1996). In fact, A' makes strong assumptions about the forms of the underlying distributions, which resemble equal-variance logistic distributions at low sensitivity and rectangular distributions at high sensitivity. For example, Figure 2 shows one pair of evidence distributions that are consistent with $A' = .9$.

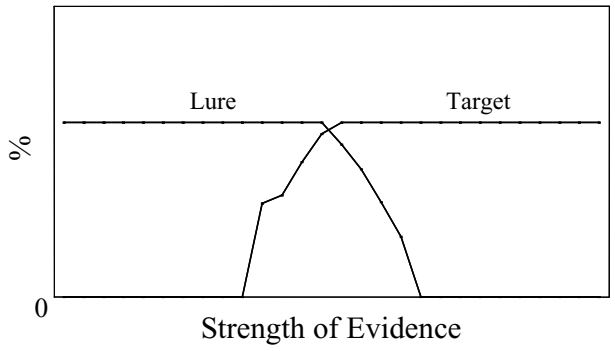

Figure 2. Hypothetical evidence distributions implied by A'.

Although the distributional assumptions have not been precisely specified, A' does imply symmetric ROCs and equal-variance underlying distributions.

Statistical Properties of Sensitivity Estimators

All single-point measures are fallible—that is, they entail assumptions that are sometimes wrong. If some such measure must be used, however, a choice among them can be made on statistical grounds. Every sensitivity measure is a statistic and thus has a sampling distribution with properties that depend on sample size and the model parameters (in this case, the true degree of discriminability). Each sensitivity statistic is an *estimator* of some model parameter and can be evaluated for three standard properties of estimators: accuracy (good agreement between the parameter and the sampling distribution mean), precision (small variance of the sampling distribution), and robustness (small influence of violated assumptions on accuracy). We examined these properties of d', A_z, and A' by systematically varying sample size and true discriminability.

Previous studies have looked at these issues in a more limited way. Miller (1996) and Kadlec (1999) examined the accuracy and precision of d' given the standard model of equal-variance Gaussian evidence distributions. Miller considered only performance by an unbiased observer, whereas Kadlec varied criterion location (response bias) as well as sensitivity. Miller noted that with small sample sizes the sampling distribution of d' is neither Gaussian nor unimodal and can produce extremely biased estimators; in this respect his calculations constituted an important advance over previous methods (Gourevitch & Galanter, 1967) that assumed normal sampling distributions. Both Miller and Kadlec found that statistical bias is most extreme when true discriminability is very high. Neither author examined characteristics of the area measures A_z and A', nor did they address the robustness question.

Donaldson (1993) did evaluate the robustness of d' and A' over a portion of ROC space and concluded that d' is more robust than A' when the variance of evidence distributions is equal, but that A' is more robust when it is not. If true, this would make A' an attractive alternative in domains where unequal variance is the rule (for example, recognition memory; Donaldson, 1996; Macmillan, Rotello, & Verde, 2005; Ratcliff, Sheu, & Gronlund, 1992; Verde & Rotello, 2003). Although Donaldson (1993) considered A_z to be an appropriate standard against which to evaluate estimated values of A', he did not examine its statistical properties.

Calculational Method

To evaluate the statistical bias[1] of an index, we compared it against true discriminability, computed from an underlying ROC defined by the model parameters. For the SDT measures d' and A_z, it is helpful to consider an ROC curve in which the z scores of F and H are plotted as coordinates to form a zROC (Figure 3). The slope s of the zROC equals the ratio of the lure and target distribution standard deviations, so equal-variance Gaussian distributions imply a zROC that is linear with unit slope. The stan-

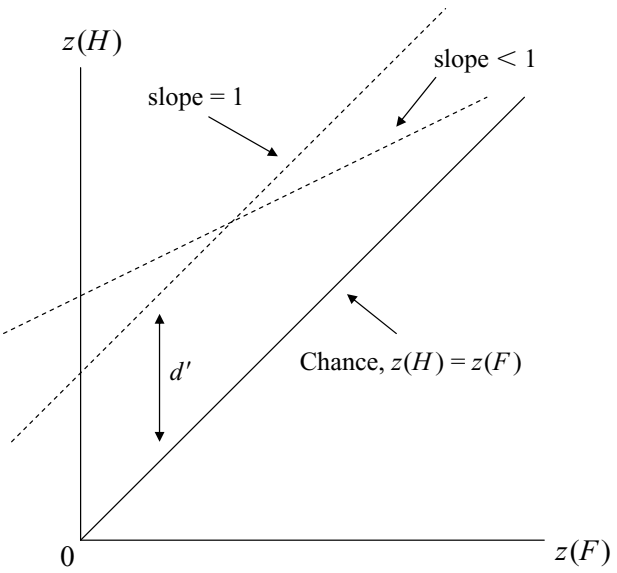

Figure 3. ROC curves on *z*-coordinates. If the underlying distributions are Gaussian, *z*ROC curves are straight lines with slope *s* equal to the ratio of lure and target distributions.

dardized distance between the means of the target and lure distributions, d', equals the vertical distance between the *z*ROC and the chance line; because these lines are parallel, this value can be derived from any point on the *z*ROC:

$$d' = z(H) - z(F). \quad (1)$$

For Gaussian distributions of unequal variance, the *z*ROC is not parallel to the chance line; it is steeper if the target distribution has smaller variance than the lure distribution, and shallower otherwise. Because the vertical distance to the chance line varies along the ROC curve, a decision must be made about the point at which "sensitivity" is to be defined or, equivalently, how the two standard deviations are to be combined. We follow Donaldson (1993) in adopting d_a (Simpson & Fitter, 1973), which measures the mean difference in units of the root mean square of the two standard deviations.

$$d_a = \left(\frac{2}{1+s^2}\right)^{1/2}\left[z(H) - sz(F)\right]. \quad (2)$$

This statistic is equivalent to d' when $s = 1$; thus, d' can be thought of as a special case of d_a (the equal-variance case). When evaluating the accuracy and robustness of d', true discriminability is computed in terms of d_a.

The index A_z, the area under the best-fitting ROC curve derived from equal-variance Gaussian distributions, is simply related to d_a. In fact, this relation provides another justification for using d_a rather than some other distance measure.

$$A_z = \Phi\left(\frac{d_a}{\sqrt{2}}\right) = \int_{-\infty}^{\frac{d_a}{\sqrt{2}}}\left(\frac{1}{\sqrt{2\pi}}\right)e^{-z^2/2}\,dz. \quad (3)$$

When calculating A_z for a single (F, H) observation, d' is substituted for d_a in this equation.

The geometric approximation of ROC area, A', was described by Pollack and Norman (1964) as the average of the maximum and minimum areas of ROCs containing the point (F, H). Subsequent work has shown this to be not exactly true (Smith, 1995; Zhang & Mueller, 2005), but the exact rationale for using A' is unimportant given how widely it is adopted. The computational formula is

$$A' = \frac{1}{2} + \frac{(H-F)(1+H-F)}{4H(1-F)} \text{ if } H \geq F \quad (4A)$$

and

$$A' = \frac{1}{2} + \frac{(F-H)(1+F-H)}{4F(1-H)} \text{ if } H < F. \quad (4B)$$

When evaluating the accuracy and robustness of the area-based indexes, A_z and A', true discriminability is computed in terms of A_z. Note that because A' cannot accommodate different values of s, it cannot be used to index true discriminability. Another reason to favor A_z as a standard is that empirical ROCs are generally consistent with the Gaussian model.

In order to examine the statistical properties of d', A_z, and A', we systematically varied sample size and true discriminability. Sample size N, which equaled the number of targets and the number of lures, was set to 8, 16, 32, 64, 128, 256, and 512. True discriminability can be described by an (F, H) point in ROC space and the slope of the *z*ROC that passes through this point. We surveyed all of ROC space representing above-chance performance (i.e., the area above the major diagonal in Figure 1).[2] We allowed F and H to take on the values .01, .1, .2, . . . , .9, and .99. For the *z*ROC slope, we included the standard equal-variance case ($s = 1$), as well as four cases of unequal variance ($s = 0.6$, 0.8, 1.2, and 1.5). Results from a representative subset of these parameter values are discussed below; a complete treatment can be found in the Psychonomic Society online archive.

For a given (F, H) point, the sampling distribution of each sensitivity index was constructed according to the method described by Miller (1996). Each sampling distribution has three parameters: N, H, and F. The observed number of hits (N_h) and the observed number of false alarms (N_f) are binomial random variables. For N_h,

$$P\left(N_h = k\right) = \binom{N}{k}H^k(1-H)^{N-k},\ k = 0,1,\ldots,N. \quad (5)$$

The analogous distribution for N_f involves the parameters N and F.

The product of the N_h and N_f distributions is the sampling distribution of $(\hat{F}, \hat{H})$, where $\hat{F} = N_f/N$ and $\hat{H} = N_h/N$.[3] This discrete distribution has $(N + 1)^2$ possible values. The sampling distributions of $\hat{d}'$, $\hat{A}_z$, and $\hat{A}'$, computed by applying Equations 1, 3, and 4, respectively, to the $(\hat{F}, \hat{H})$ distribution, were used to find $E(\hat{d}')$, $E(\hat{A}_z)$, $E(\hat{A}')$, and standard errors.

A difficulty that arises in calculating d' and A_z is that Equations 1 and 3 are undefined when $\hat{H}$ and $\hat{F}$ take on values of 0 or 1. This problem can be addressed by discarding, replacing, or transforming those cases. For example, one can simply discard observations for which $\hat{N}_h$ or $\hat{N}_f$ takes on values of 0 or N, normalizing the remaining observations so that their probabilities sum to 1. As is often done in psychophysics (Macmillan & Kaplan, 1985), one can replace $\hat{N}_h$ and $\hat{N}_f$ values of 0 with 0.5 and values of N with $(N - 0.5)$. Finally, one can transform all of the observed hit and false alarm rates so that $\hat{N}_h = (\hat{N}_h + 0.5)/(N + 1)$ and $\hat{N}_f = (\hat{N}_f + 0.5)/(N + 1)$, referred to as the *log-linear rule* because of its association with log-linear analysis. With regard to $\hat{d}'$ bias, Miller (1996) found the discarding and replacement corrections to be about equally successful. Hautus (1995) described simulations that favored the log-linear rule over the replacement correction. However, Kadlec (1999) noted that those simulations sometimes involved unrealistic parameter settings; her own simulations suggested that the two corrections performed about equally well. We implemented all three corrections in our computations for $\hat{d}'$ and $\hat{A}_z$ and found that no single correction was always best at minimizing statistical bias and standard error; the winner varied with sample size, location in ROC space, and s. However, the log-linear model transformation seemed the best choice overall, and the results presented here use this correction. We refer those interested in a more detailed comparison of the correction methods to the online database.

Accuracy

The statistical bias of a sensitivity index (the inverse of accuracy) is the difference between the value of a parameter and the expected value of its estimator. We compared the distance measure $\hat{d}'$ with the parameter d_a and the area measures $\hat{A}_z$ and $\hat{A}'$ with the parameter A_z. Figure 4 displays statistical bias when $s = 1$ for sample sizes N of 16, 64, and 256. Each panel shows separate dashed-line functions for each level of H as F is varied. Note that only above-chance performance is shown, so that each H function terminates at $H = F$ (in the figures, the different termination points make it easier to distinguish between the functions).

Figure 4 (top row) shows statistical bias for $\hat{d}'$, $E(\hat{d}') - d_a$ (because $s = 1$, $d_a = d'$ in this case). The effect of bias in all cases is to underestimate true sensitivity. The influ-

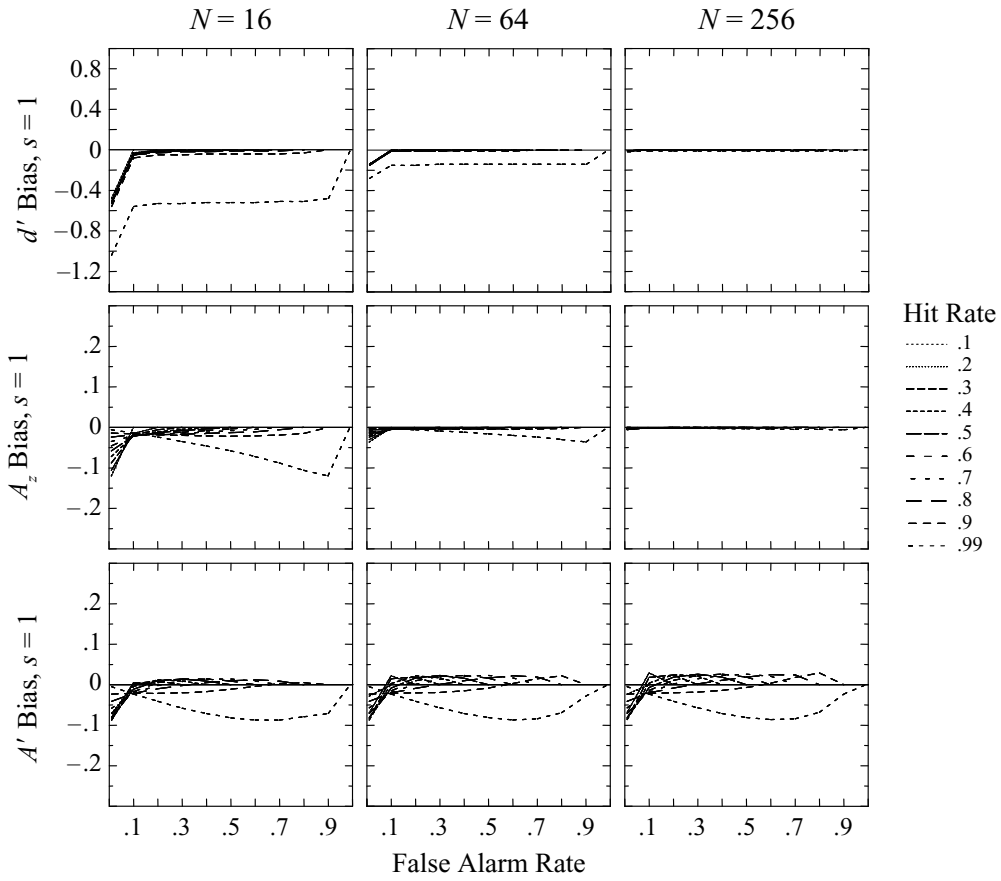

Figure 4. Statistical bias (expected value of the estimator minus the parameter value) of three sensitivity measures. Rows are d', A_z, and A'; columns are numbers of trials, N = 16, 64, and 256. Each parametric value of H is plotted as a function of F.

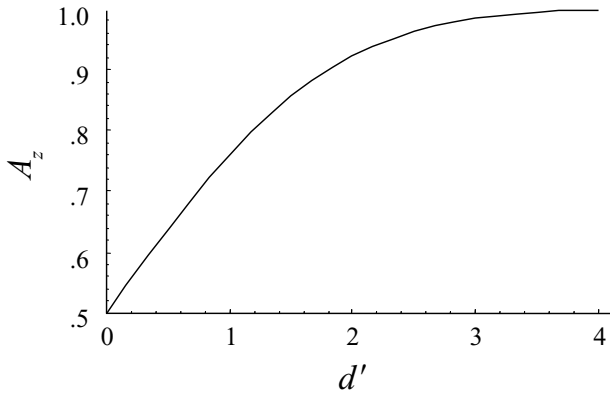

Figure 5. The transformation from d' to A_z (Equation 3).

ence of sample size is considerable; although bias can be quite large when $N = 16$, it is minimal when $N = 256$. Several factors contribute to this pattern of bias. One factor is that the number of observations (N) constrains the possible values of $\hat{d}'$ that can be obtained in a sample. For example, when $N = 1$, there are only two possible values of $\hat{F}$ (0 or 1), two possible values of $\hat{H}$ (0 or 1), and three possible values of $\hat{d}'$. It is difficult to make generalizations about the effect of this factor on bias except to note that the effect becomes smaller as N grows larger. A second factor is that the parametric values of $F = .01$ and $H = .99$ are much more likely to lead to underestimation bias than are values toward the middle of the probability scale. For example, the sampling distribution of $F = .20$ will produce observations that are well above or below .20. However, the sampling distribution of $F = .01$ cannot produce observations much below .01 due to a floor effect. The sampling distribution of $H = .99$ suffers from an analogous ceiling effect. These effects tend to reduce $\hat{d}'$ and are partly responsible for the extreme bias observed when $F = .01$ or $H = .99$.

A third factor is that the log-linear rule correction of $\hat{N}_f$ and $\hat{N}_h$ constrains the minimum $\hat{F}$ and maximum $\hat{H}$ that can be obtained in a given sample. This constraint is driven by the value of N. When $N = 16$, minimum $\hat{F}$ is $[0.5 / (16 + 1)] = .029$ and maximum $\hat{H}$ is $[(16 + 0.5) / (16 + 1)] = .971$. When $N = 256$, on the other hand, minimum $\hat{F}$ is .002 and maximum $\hat{H}$ is .998. In other words, as N increases, minimum $\hat{F}$ and maximum $\hat{H}$ converge to 0 and 1, respectively. This effect has two consequences. First, remember that d' takes on extreme values as F approaches 0 or H approaches 1. The transformation moderates the values of F and H, reducing d' and thus producing the observed underestimation of true sensitivity. The effect of the transformation, and thus the underestimation problem, is reduced as N grows larger. Second, because sampled values of $\hat{N}_f = 0$ and $\hat{N}_h = 1$ are extremely likely when $F = .01$ and $H = .99$, respectively, the correction has its greatest effect on bias in these areas of ROC space.

Different correction methods place different constraints on obtainable $\hat{F}$ and $\hat{H}$. Thus, methods other than the log-linear rule produce somewhat different patterns of bias; this can be seen in previous studies of d' accuracy

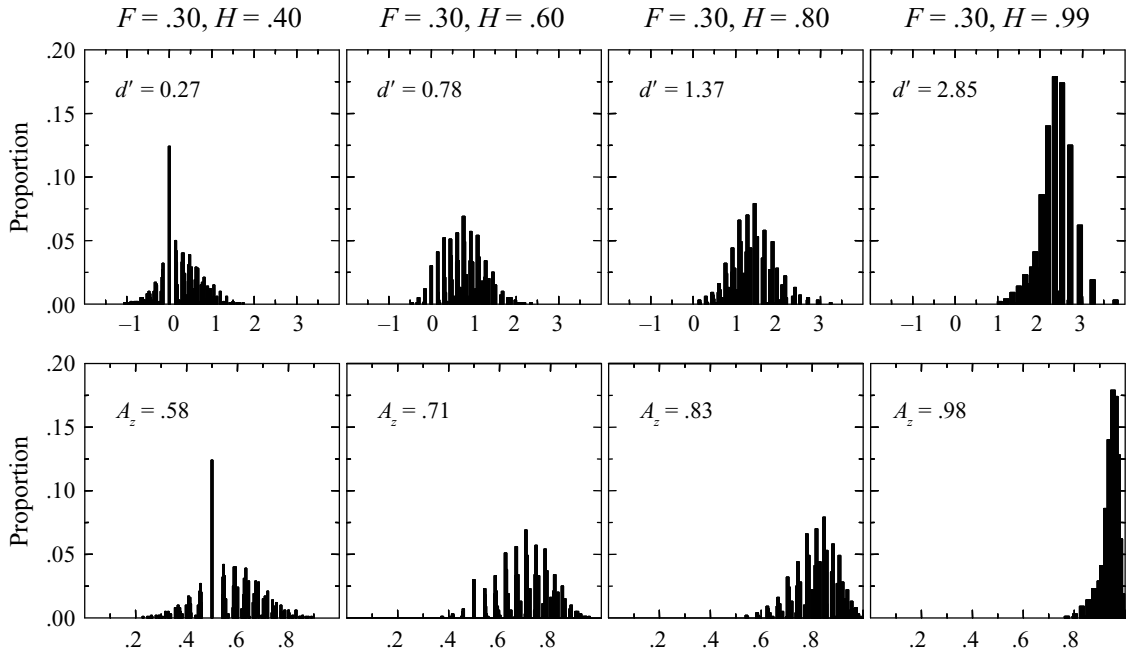

Figure 6. Sampling distributions of d' (top) and A_z (bottom). The sampling distributions are shown for several parametric values of F and H. The value of d' or A_z inset in each panel is calculated from F and H and represents the true discriminability being estimated by the sample.

by Miller (1996) and Kadlec (1999) (see also the online database for this article). However, the other factors that affect bias lead to similarities across all of the studies. Both Miller and Kadlec noted that greater statistical bias is found with very high levels of true discriminability, and Kadlec further noted that statistical bias increases when the decision criterion deviates from that of an optimal observer (in other words, when response bias is greater). Our findings are consistent with these observations but support the more comprehensive point that it is the specific location in ROC space that determines the magnitude of statistical bias.

Figure 4 (middle row) shows statistical bias for $\hat{A}_z$, $E(\hat{A}_z) - A_z$. All estimates are again biased low, and the large bias for $N = 16$ all but disappears for $N = 256$. As with $\hat{d}'$, bias is greatest for extreme values of H and F. However, the patterns of bias in general are somewhat different. Figure 4 (bottom row) shows statistical bias for $\hat{A}'$, $E(\hat{A}') - A_z$. Both positive and negative biases occur for this index. Notably, accuracy does not improve as sample size increases, because with increasing N, $\hat{A}'$ converges on a model with different underlying assumptions from A_z (Macmillan & Creelman, 1996).

On what metric should the accuracy of these indexes be compared? Donaldson (1993) calculated percent error for each statistic:

$$\text{percent error } d' = \frac{\hat{d}' - d_a}{d_a} \times 100 \qquad (6A)$$

and

$$\text{percent error } A' = \frac{\hat{A}' - A_z}{A_z} \times 100\,. \qquad (6B)$$

Donaldson found A' to be superior by this measure, and an examination of Figure 4 confirms his result: Using the calculations above, extreme error (for $N = 16$, $H = .99$, $F = .9$) reaches about 50% in d', but only about 10% in A'. However, this comparison is problematic, because one measure (d') is on a distance scale, and the other (A') is on an area scale; it is more informative to compare the two area measures A' and A_z. The two comparisons can be expected to produce different results because the transformation from d' to A_z is nonlinear, as shown in Figure 5: A given percent change in A_z does not lead to the same percent change in d'. As a consequence, the sampling dis-

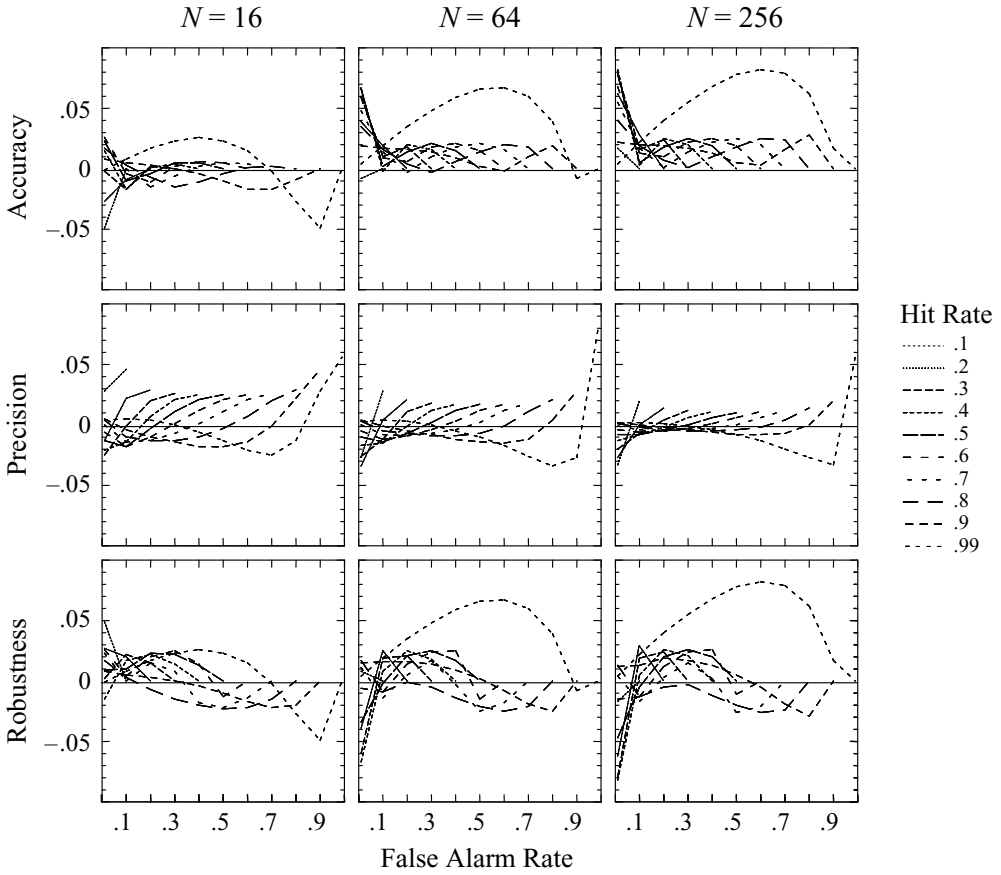

Figure 7. Comparisons of A' and A_z. The ordinate represents differences in absolute values of A' and A_z accuracy (top), precision (middle), and robustness when $s = 0.6$ (bottom). Positive values indicate that A_z is superior, negative values that A' is superior.

tributions of d' and A_z are quite different: Figure 6 shows that the discrepancy is particularly great at high levels of sensitivity.

When A' and A_z are compared, A' loses its advantage. Figure 7 (top row) compares accuracy of A' and A_z in terms of absolute bias, $| E(\hat{A}') - A_z | - | E(\hat{A}_z) - A_z |$. Positive values mean that A_z is more accurate, negative values that A' is preferable. For $N = 16$ neither statistic is clearly superior, but for larger N the more accurate index is A_z, especially for extreme values of H and F.

Precision

Figure 8 displays standard error of the sensitivity indexes as a function of sample size ($N = 16$, 64, and 256) and the true hit and false alarm rates. Standard error is determined by the parameters of the sampling distribution and is unaffected by the value of s.[4] Increasing the sample size reduces standard error. Miller (1996) examined the variance of $\hat{d}'$ and observed a complex trend in which variance increases as $\hat{d}'$ increases, but for smaller values of N variance decreases as $\hat{d}'$ approaches perfect performance. Miller identified two factors at work: the spread of the $\hat{d}'$ distribution as discriminability departs from zero and the narrowing of the distribution as it hits the limit of the maximum possible value of $\hat{d}'$. These factors interact to produce the trends evident in Figure 8 for the two SDT indexes. For $\hat{d}'$, standard error generally increases as true discriminability increases (in other words, for a given F, standard error increases as H increases), but for $\hat{A}_z$ and $\hat{A}'$ the opposite is true. This curious difference seems to be related to the differential impact of the two factors Miller identified on the distance and area scales, as can be seen in Figure 6. Visual inspection of that figure suggests that for $\hat{d}'$, the spread of the distribution increases as H increases and F remains constant. However, when $H = .99$ the distribution becomes more compact, because it presses against the maximum attainable value of d'. Note that the reversal does not occur with larger N (this can also be seen in Miller's data); as N increases, so too does the maximum obtainable d', reducing its limiting effect. For $\hat{A}_z$, on the other hand, the limit imposed by its maximum obtainable value seems to have an influence from the start, so that the distributions become more compact immediately as performance increases above chance.

Figure 7 (middle row) compares A' and A_z in terms of the difference in their standard errors [standard error A' − standard error A_z]. Where values are positive, $\hat{A}_z$ is more precise than $\hat{A}'$ (and vice versa). Which index has superior

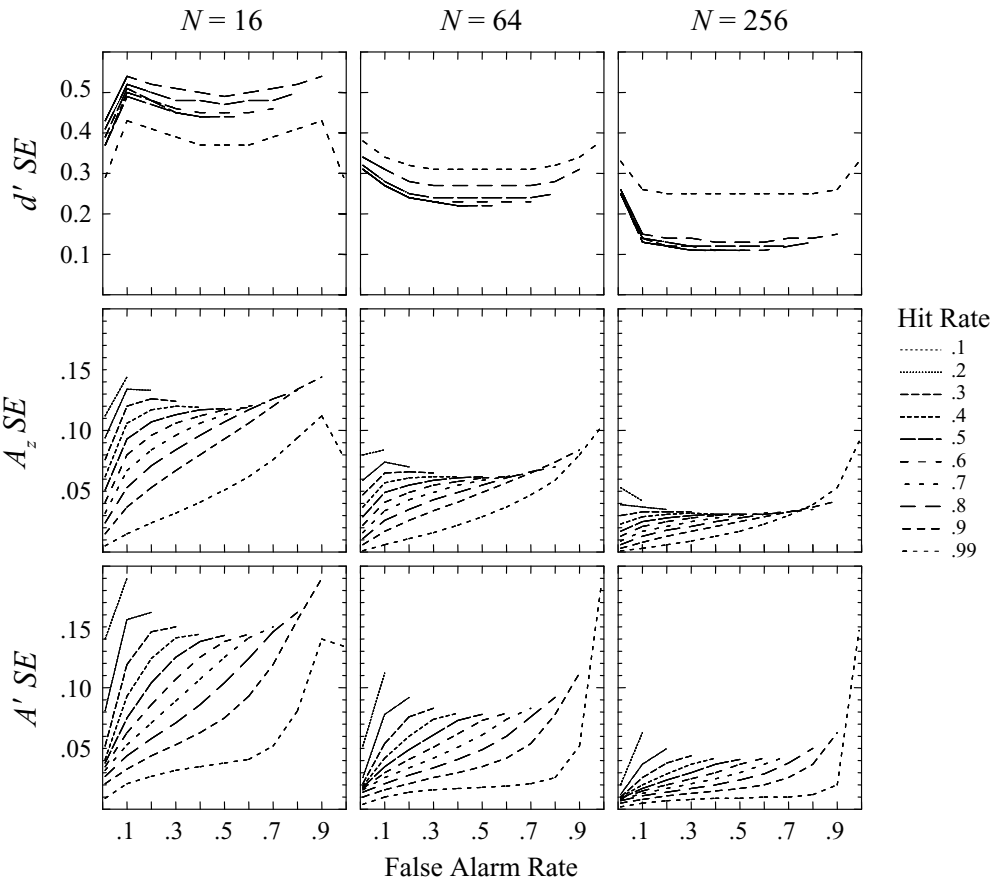

Figure 8. Standard error of the estimator of three sensitivity indexes. Rows are d', A_z, and A'; columns are numbers of trials, $N = 16$, 64, and 256.

precision depends on the location in ROC space, and this is true for all values of N. In general, $\hat{A}_z$ is somewhat more precise, especially for larger values of N.

Robustness

Robustness refers to accuracy of the sensitivity index when underlying assumptions are violated. We focus here specifically on violation of the assumption that the evidence distributions have equal variance (i.e., $s \neq 1$), because all three indexes make this assumption. The ratio of lure and target standard deviations, s, was set to 0.6, 0.8, and 1.2, and we again considered sample sizes $N = 16$, 64, and 256 (additional values of s and N are included in the online database).

Figure 9 shows the statistical bias of d' under conditions of unequal variance. When $s < 1$, d' generally becomes more positively biased for smaller values of H and F (left side of ROC space) and more negatively biased for larger values of H and F (right side of ROC space). The reverse is true when $s > 1$. Bias due to unequal variance can be significant, over 20% when H and F are very large or small. Moreover, the problem is not much alleviated by increasing N, because d' converges on an incorrect model of the underlying distributions.

The systematic bias produced by incorrectly assuming equal variances is illustrated in Figure 10. In this figure, lines A, B, C, and D are z-transformed ROCs consistent with underlying Gaussian distributions. The slopes equal the ratio of the lure to the target standard deviations; for line A, the variances are equal ($s_1 = 1$), whereas for lines B, C, and D, the variances are unequal ($s_2 < 1$). An experimenter who calculates d' from a single point implicitly assumes that the true ROC is of unit slope, like line A. If the true ROC is line B and point p_1 is observed, then the calculated d' equals true d_a. The agreement is, however, entirely fortuitous. If a point to the left of p_1 (such as p_2) is observed, d' produces a value larger than true d_a; for points to the right of p_1 (such as p_3), d' produces a value smaller than true d_a. This is exactly the pattern displayed in Figure 9.

Figure 11 shows A_z bias, and Figure 12 shows A' bias when $s \neq 1$. The patterns and conclusions to be drawn about these indexes are similar to those for d'. Donaldson (1993) calculated the percent error of d' and of A' (Equations 6A and 6B) as true discriminability and s varied and concluded that A' is more accurate in the majority of cases when $s \neq 1$. The present survey covers a larger region of ROC space and includes the effect of sample size. More

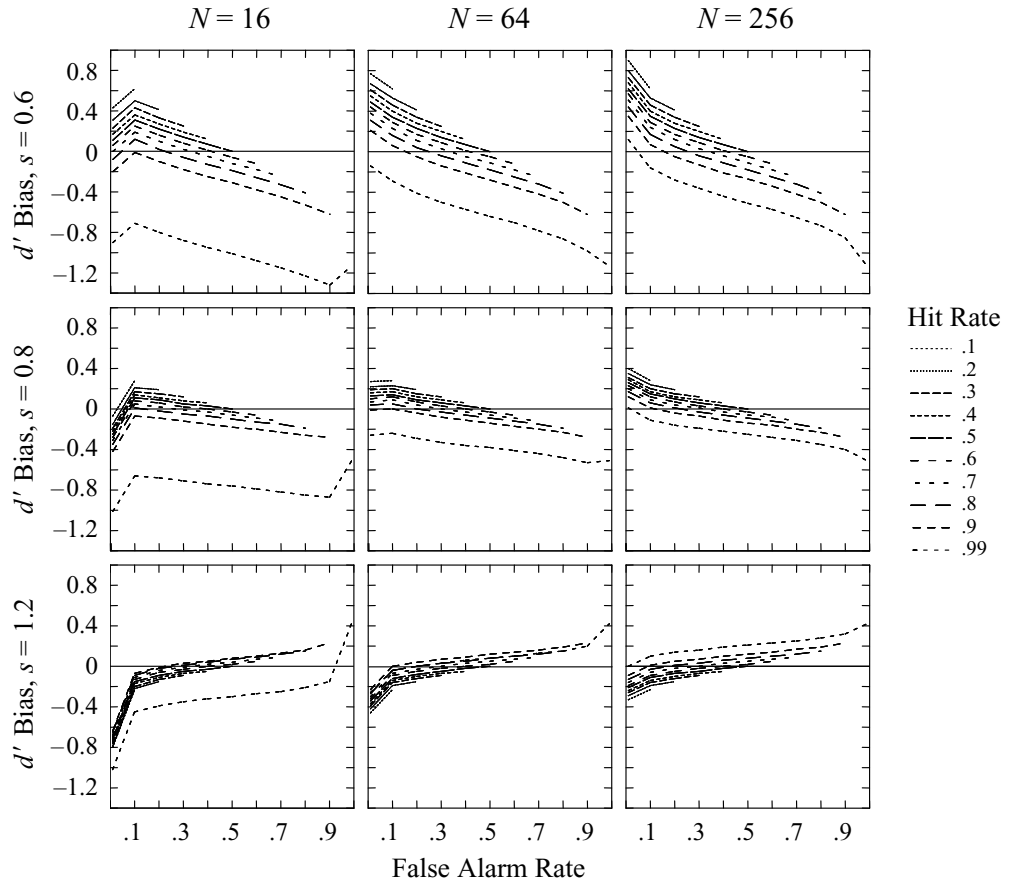

Figure 9. Statistical bias of d' if zROC slope ≠ 1. Rows are slopes of 0.6, 0.8, and 1.2; columns are numbers of trials, $N = 16$, 64, and 256.

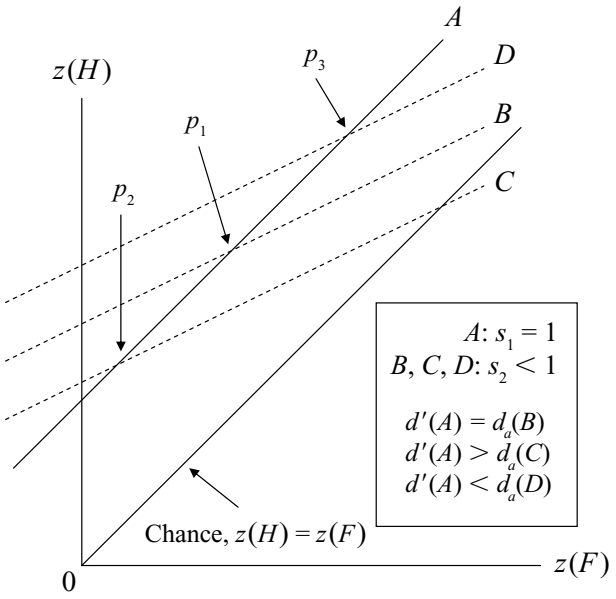

Figure 10. Unequal variance and the *z*ROC. Four hypothetical *z*ROCs: *A* represents equal variance, with slope $s_1 = 1$; *B*, *C*, and *D* represent unequal variance with slope $s_2 < 1$. If the underlying *z*ROC has slope s_2, then *d′* will underestimate sensitivity for any point on *A* to the right of p_1 and overestimate sensitivity for any point to the left of p_1.

importantly, as we suggested earlier, directly comparing percent error of d' and A' is problematic because the distance and area scales are not comparable. A comparison of the robustness of A' and that of the SDT area measure A_z is shown in Figure 7 (bottom row), which plots the difference in absolute bias of the two indexes, $|E(\hat{A}') - A_z| - |E(\hat{A}_z) - A_z|$, when $s = 0.6$. Contrary to Donaldson's conclusion, neither statistic is clearly superior; each is more accurate in some regions of ROC space (although A_z has the advantage over a slightly larger portion of the space).

Implications

Signal detection theory is a standard tool for analyzing performance in many domains. In SDT terms, discrimination sensitivity is determined by the nature of target and lure evidence distributions. One should ideally construct ROCs that provide detailed information about these underlying distributions, but it is not always feasible to gather the data required for ROCs. The alternative is to use a two-response task that provides only one hit and false alarm rate per condition and relies on sensitivity indexes like d', A_z, and A' that make simplifying assumptions about the underlying distributions. The present findings offer investigators several lessons to consider when designing experiments and analyzing data that rely on these indexes.

The need for reasonable sample sizes is something one keeps in mind with any statistic. Miller (1996) cautioned that a difference in statistical bias between conditions that differ in N can confound any d' comparison between them, and Macmillan, Rotello, and Miller (2004) raised the same point about several statistics abstracted from ROC curves. The present findings show that accuracy and precision of A_z and A' can also vary greatly between conditions that differ in sample size. Dealing with this problem is usually a simple matter of designing an experiment such that N is equated across conditions. A more complex problem is that bias and standard error also depend on underlying discriminability. Differences in discriminability may be inherent in the phenomenon under investigation: It may be of interest to compare a hard to an easy condition, or to compare overall discrimination judgments to a subset of those judgments. Such comparisons are analogous to comparing performance in different locations of ROC space. Of course, as long as the locations are not too far apart, the problem can be minimized by ensuring that sample size is reasonably large.

Violation of the equal-variance assumption is a problem that is sometimes acknowledged, but the consequences of such violation have not been well documented. Our findings reveal that unequal variance produces systematic positive bias in one region of ROC space and negative bias in the opposite region, the regions depending on the value of s. Over much of ROC space, this bias is significant (for example, when $N = 256$ and $s = 0.6$, A_z bias often exceeds 10% and can be much higher) and is not much reduced by increasing N, which only leads the index to converge on the wrong model of the underlying distributions. The assumption of equal variance is made by all the single-point indexes we have considered; if incorrect, this assumption can lead to serious errors that cannot be eliminated by computational adjustment or correction.

Recent issues in the memory literature illustrate how systematic statistical bias can pose serious difficulties for theoretical interpretation. In some circumstances, memory illusions (the false belief that something was previously encountered) seem to be the product of changes in decision criterion rather than changes in the actual quality of memory (McDermott & Watson, 2001; Niewiadomski & Hockley, 2001; Verde & Rotello, 2003; Whittlesea, 2002). According to SDT, criterion change has no effect on the characteristics of the evidence distributions, whereas a change in discrimination sensitivity does imply a change in distributional characteristics. Thus, one should be able to claim that a memory illusion that affects the sensitivity index is not solely a product of criterion placement. A problem with this interpretation arises from the observation that evidence distributions in recognition memory typically have unequal variance (Ratcliff et al., 1992). In their investigation of the "revelation effect" illusion, Verde and Rotello (2003) observed consistent effects on d' (calculated from overall H and F) but no effect on d_a (calculated from empirical ROCs). They argued that the illusion in fact affected only criterion placement, but that the change in criterion combined with unequal variance led to systematic effects on d'. The lesson is that under conditions of unequal variance, a change in decision criterion alone can affect a sensitivity index like d', even when sensitivity has not actually changed. Distinguishing between changes in criterion (i.e., response bias) and sen-

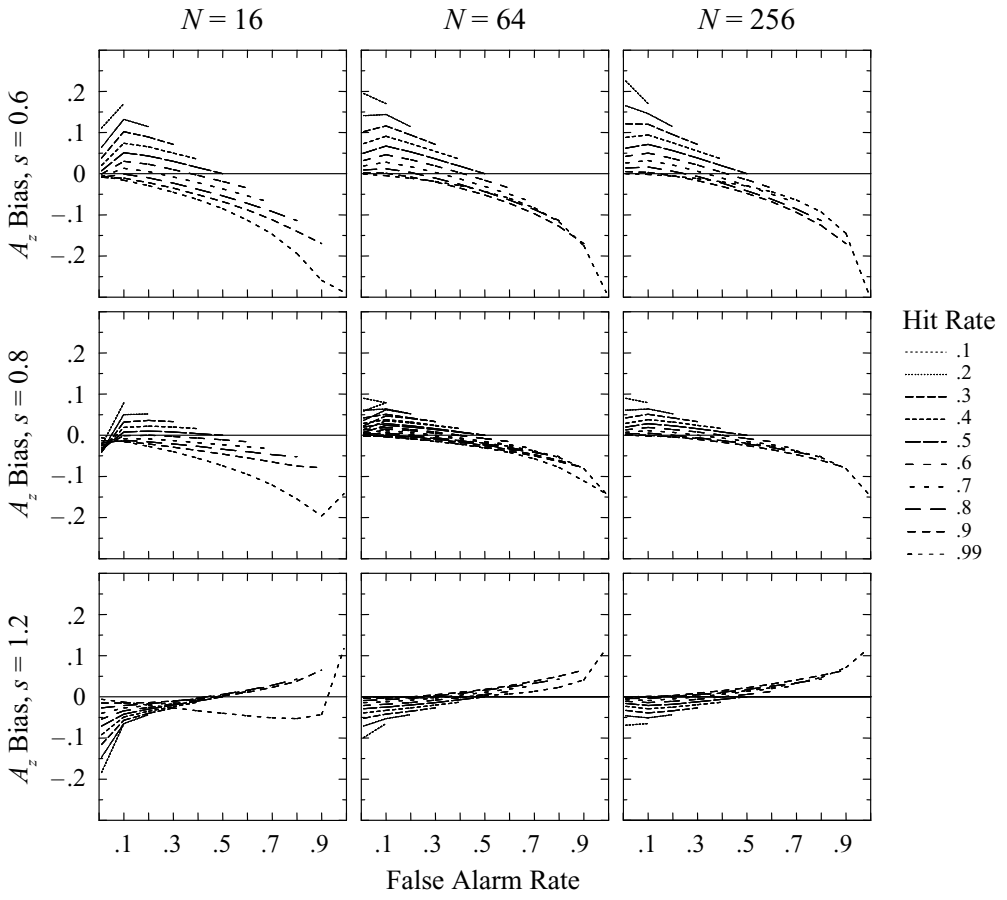

Figure 11. Statistical bias of A_z if zROC slope ≠ 1. Rows are slopes of 0.6, 0.8, and 1.2; columns are numbers of trials, N = 16, 64, and 256.

sitivity is a theoretically important matter in any domain. Unfortunately, if equal variance cannot be assumed but one must rely on a sensitivity index based on a single H and F, then making this distinction is difficult.

If one must rely on a single-point index, which is the best choice? A' has had many proponents over the years. Much of this popularity seems to derive from the mistaken belief that A' is nonparametric (Macmillan & Creelman, 1996). The convenient property that A' can accommodate H and F values of 0 and 1 has also been noted. Finally, Donaldson (1993) has suggested that A' seems to be more robust than d'.

Our results lead us to the conclusion that A_z, the area under the normal–normal ROC curve going through the (F, H) point, is the preferred index on several grounds. The distributional assumptions entailed by A' are as specific but far less commonly justified than the normality assumption of A_z (Macmillan & Creelman, 1996; Macmillan et al., 2005; Pastore, Crawley, Berens, & Skelly, 2003). The use of corrections like the log-linear transformation for d' and A_z solves the in-principle problem of infinite d' (when F or H takes on values of 0 or 1). With regard to the claim of greater robustness made by Donaldson (1993), the present findings allow more detailed conclusions. To avoid comparing percent error d' with percent error A' (which is problematic due to the nonlinear relationship between the distance and area scales), we compared A_z with A', both of which are in units of area, and found A_z to be clearly more accurate under conditions of equal variance except when N is small. With regard to precision and robustness (accuracy under conditions of unequal variance), the picture is less clear; each index does better in different regions of ROC space. However, as N grows large, A_z tends to gain the advantage. Based on these statistical properties alone, A_z seems to be the better choice, especially when the variance ratio is unknown. There seems to be little statistical justification for choosing A' over competing indexes.

All the limitations of single-point measures can, of course, be circumvented by collecting ROC curves (see Macmillan et al., 2004, for the statistics of parameters obtained from ROCs). If a single-point measure must be used, its negative consequences can be minimized by encouraging equal response bias—that is, $H \approx 1 - F$. This requirement can, however, be difficult to follow. For example, in the remember–know recognition memory lit-

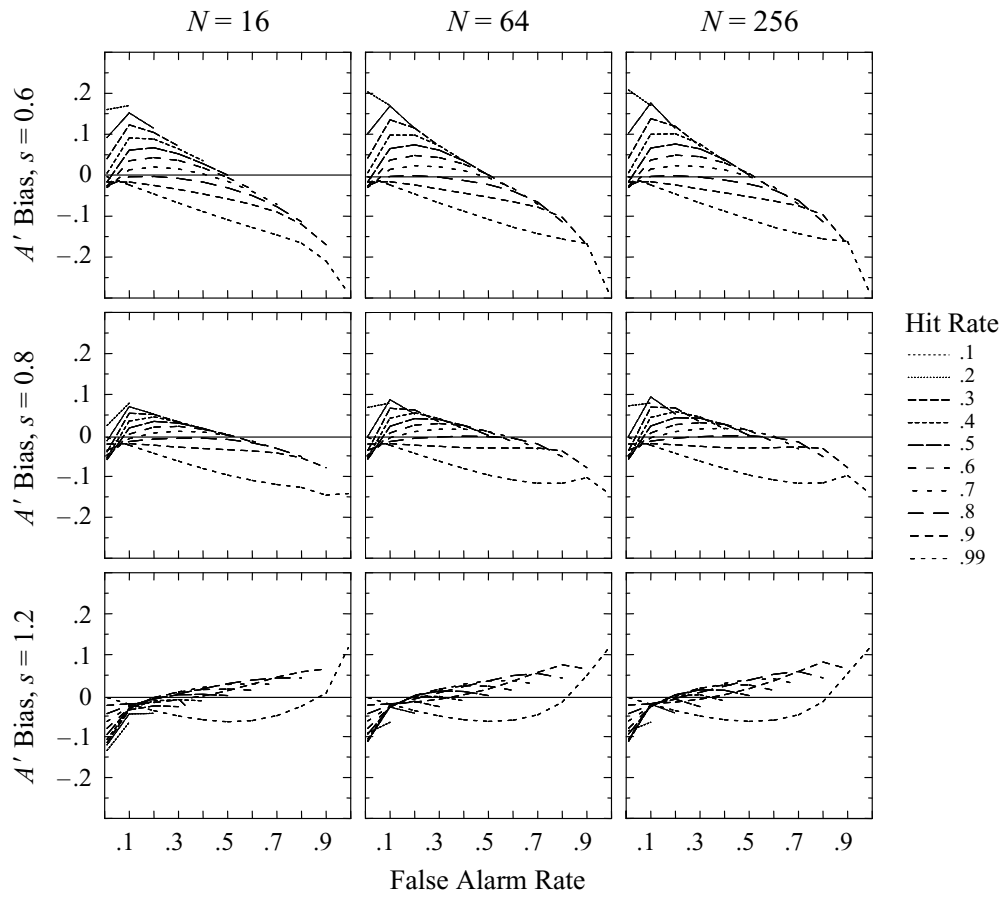

Figure 12. Statistical bias of A' if zROC slope ≠ 1. Rows are slopes of 0.6, 0.8, and 1.2; columns are numbers of trials, N = 16, 64, and 256.

erature, sensitivity is sometimes calculated from "remember hit rates" and "remember false alarm rates." Dunn's (2004) survey of such experiments showed that the latter averaged only about .05, so the corresponding ROC points fall close to the left edge of ROC space. The unfortunate consequences of using single-point measures in this case have been explored by Macmillan et al. (2005).

REFERENCES

Donaldson, W. (1993). Accuracy of d' and A' as estimates of sensitivity. *Bulletin of the Psychonomic Society*, **31**, 271-274.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, **24**, 523-533.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review*, **111**, 524-542.

Gourevitch, V., & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, **32**, 25-33.

Green, D. M. (1964). General prediction relating yes–no and forced-choice results. *Journal of the Acoustical Society of America*, **36**, 1042 (Abstract).

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

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, **27**, 46-51.

Kadlec, H. (1999). Statistical properties of d' and β estimates of signal detection theory. *Psychological Methods*, **4**, 22-43.

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, **3**, 164-170.

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

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, **98**, 185-199.

Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychophysics*, **66**, 406-421.

Macmillan, N. A., Rotello, C. M., & Verde, M. F. (2005). On the importance of models in interpreting remember–know experiments: Comments on Gardiner et al.'s (2002) meta-analysis. *Memory*, **13**, 607-621.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory & Language*, **45**, 160-176.

Miller, J. (1996). The sampling distribution of d'. *Perception & Psychophysics*, **58**, 65-72.

Niewiadomski, M. W., & Hockley, W. E. (2001). Interrupting recognition memory: Tests of familiarity-based accounts of the revelation effect. *Memory & Cognition*, **29**, 1130-1138.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, **10**, 556-569.

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of $d'(e)$. *Psychological Bulletin*, **71**, 161-173.
Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, **1**, 125-126.
Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, **99**, 518-535.
Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin*, **80**, 481-488.
Smith, W. D. (1995). Clarification of sensitivity measure A'. *Journal of Mathematical Psychology*, **39**, 82-89.
Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, **99**, 181-198.
Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 739-746.
Whittlesea, B. W. A. (2002). False memory and the discrepancy-attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General*, **131**, 96-115.
Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, **70**, 1-10.

NOTES

1. The term *bias* is used in two senses in this article: statistical bias (inverse of accuracy) and response bias (tendency by the observer to prefer one of the two responses). Unless the context makes the meaning clear, we avoid referring simply to "bias." Similarly, some authors use the term *accuracy* as a synonym for sensitivity, but in this article it always refers to statistical accuracy.

2. Results above and below the major diagonal mirror one another, so the latter can be easily derived.

3. We adopt the standard convention of representing an estimator of the parameter p by $\hat{p}$.

4. The sampling distribution of A_z is found from Equations 2, 3, and 5. Although the term s does appear in Equation 2, it is set equal to 1 for single (F, H) observations, as noted earlier.

ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, www.psychonomic.org/archive/.

To access these files, search the archive for this article using the journal (*Perception & Psychophysics*), the first author's name (Verde), and the publication year (2006).

File: Verde-P&P-2006.zip

Description: The compressed archive file contains four files:

Sensitivity_Statistics.pdf, containing the complete dataset of estimated d', A', and A_z across a range of parameter values and correction methods.

Sensitivity_Statistics.txt, containing the sensitivity data in .txt form.

Key to File Sensitivity_Statistics.pdf, containing the key for reading the sensitivity data.

Key to File Sensitivity_Statistics.txt, containing the key in .txt form.

Author's e-mail address: michael.verde@plymouth.ac.uk.

(Manuscript received April 14, 2005;
revision accepted for publication July 20, 2005.)