Behavior Research Methods
2005, 37 (3), 436-449

The optimal correction for estimating extreme discriminability

GLENN S. BROWN and K. GEOFFREY WHITE
University of Otago, Dunedin, New Zealand

Discriminability measures such as d' and log d become infinite when performance is extremely accurate and no errors are recorded. Different arbitrary corrections can be applied to produce finite values, but how well do these values estimate true performance? To answer this question, we directly calculated the effects of a range of different corrections on the sampling distributions of $\hat{d}'$ and log $\hat{d}$. Many arbitrary corrections produced better estimates of discriminability than did the intuitively plausible technique of rerunning problem conditions. We concluded that when it is not possible to run more trials and when other techniques are not appropriate, the best correction overall is to add a correction constant between 0.25 and 0.5 to all response counts, regardless of their value.

Discrimination tasks are often used to explore behavioral and psychological phenomena. In these tasks, measures of performance are typically based on transforming either the proportion of correct responses or the ratio of correct to incorrect responses. A common problem with such measures is that they tend not to yield accurate estimates of high discriminability. When performance is extremely accurate, there are instances of no errors. This results in an indeterminate measurement of performance. In this article, we briefly review the major techniques that others have used to deal with such problem cases. Then we evaluate the mathematical procedures for dealing with extreme estimates of discriminability. Our conclusion is that when it is not possible or practical to use nonmathematical means of avoiding indeterminate measurements, it is best to add a constant from 0.25 to 0.5 to all response counts.

The Yes–No Task

The present article applies principally to data obtained from yes–no tasks, described comprehensively by Macmillan and Creelman (1991), and conceptually similar tasks, such as symbolic matching-to-sample (see, e.g., Alsop, Rowley, & Fon, 1995) and delayed matching-to-sample (see, e.g., White, 1985) tasks. These tasks involve a sequence of many discrete trials. In each trial, subjects experience one of two stimuli (S_1 or S_2). As an example, for word recognition, S_1 and S_2 might be distractor and target words.

For the present purposes, the type of stimulus is not important. Instead, the important point is that each stimulus has a correct response (B, for *behavior*) associated with it: B_1 is correct when S_1 is presented, and B_2 is correct when S_2 is presented. In each trial, subjects choose between B_1 and B_2 once S_1 or S_2 has been presented. In a word recognition test, for example, B_1 and B_2 are *no* and *yes* responses. This procedure gives rise to four types of response: correct and incorrect responses on either S_1 or S_2 trials. In the terminology of signal detection theory (Green & Swets, 1966), these correspond, respectively, to the four cells of a signal detection matrix: correct rejections, false alarms, hits, and misses (nominating S_2 as the signal stimulus).

Performance Measures

To measure performance in yes–no tasks, researchers often count the frequency of each of the four response types. From here, one of the simplest measures is percent correct (the proportion of all correct responses—i.e., of hits plus correct rejections). Percent correct has the advantage that it is never infinite, but it is disadvantaged by ceiling effects. Specifically, when performance is very accurate, percent correct asymptotically increases to its ceiling of 100%. Because of this fact, percent correct does not vary on an equal-interval scale; as the ratio of correct to incorrect responses gets larger and larger, the change that it produces in percent correct gets smaller and smaller, making interactions between conditions difficult to interpret (Loftus, 1978; Wixted, 1990). Another disadvantage of percent correct is that it can be "contaminated" by response bias. *Response bias* is a general preference for one choice (B_1 or B_2) over the other, regardless of which stimulus was presented. When there is response bias, percent correct drops even though the subject is no less able to discriminate the sample stimuli.

The present research was presented to the Society for Quantitative Analyses of Behavior, New Orleans, May 2001, and was supported by a Claude McCarthy Fellowship to G.S.B. This research was also supported in part by a University of Otago Postgraduate Scholarship to G.S.B. We thank Jonathan Vaughan, Tony Nevin, and an anonymous reviewer for their helpful comments. Reprints and copies of Visual Basic programs used to conduct the analyses may be requested from K. G. White, Department of Psychology, University of Otago, Dunedin, New Zealand (e-mail: kgwhite@otago.ac.nz).

Note—This article was accepted by the previous editor, Jonathan Vaughan.

Other common measures of performance are the discriminability measures d', log d, and ln α (see Macmillan & Creelman, 1991, for a comprehensive account of d' and ln α). Their respective equations are

$$d' = Z\left(\frac{\text{Hits}}{\text{Hits} + \text{Misses}}\right) - Z\left(\frac{\text{False Alarms}}{\text{False Alarms} + \text{Correct Rejections}}\right), \quad (1)$$

$$\log d = \frac{1}{2} \cdot \log_{10}\left(\frac{\text{Hits}}{\text{Misses}} \cdot \frac{\text{Correct Rejections}}{\text{False Alarms}}\right), \quad (2)$$

and

$$\ln \alpha = \frac{1}{2} \cdot \log_{e}\left(\frac{\text{Hits}}{\text{Misses}} \cdot \frac{\text{Correct Rejections}}{\text{False Alarms}}\right). \quad (3)$$

d' was derived from signal detection theory (Green & Swets, 1966), log d from behavioral detection theory (Davison & Tustin, 1978), and ln α from Luce's (1963) choice theory (by McNicol, 1972). As Equations 1, 2, and 3 reveal, these measures are based on transformations of the same four values: frequencies of hits, misses, correct rejections, and false alarms. In theory, they cannot be contaminated by response bias (Davison & Tustin, 1978; Macmillan & Creelman, 1991), and since their scales are not bounded, they do not suffer from ceiling effects. They also share the advantage of varying on an equal-interval scale—that is, when scores increase or decrease, the change is directly proportional to a change in the assumed underlying psychological process (Wixted, 1990), making it is easier to interpret differences between conditions, especially with interactions (Loftus, 1978). The common problem of these measures, however, is that if any of the four values of hits, misses, correct rejections, or false alarms are zero, the discriminability estimate is undefined. This problem is the focus of the present article.

Before progressing, it is important to point out other similarities between the three measures. First, d' and log d have a roughly linear relationship (Figure 1). Second, the formulae for log d (Equation 2) and ln α (Equation 3) are identical in all but the base values of their logarithms. Specifically, log d uses a base-10 logarithm, whereas ln α uses a base-e logarithm, so ln α is always 2.303 times greater than log d. For this reason, we shall ignore ln α from here on. It should be kept in mind, though, that our conclusions about log d apply equally to ln α.

The Problem of Infinite Estimates of Discriminability

In theory, d' and log d may take on any value without being restricted by a ceiling. In practice, however, experimental estimates of these parameters may not do so. Because experiments always arrange a finite number of trials, estimated values of $\hat{d}'$ or log $\hat{d}$ may only take on a finite number of discretely distributed values (Miller, 1996).[1] This also means that there is a maximum finite value that the estimates can assume, the *maximum obtainable value* (MOV), even if the theoretical d' or log d parameter that it estimates is much greater. Unless data are adjusted, the MOV is determined entirely by the number of trials. Figure 2 shows that when the number of trials increases, the MOVs of $\hat{d}'$ and log $\hat{d}$ rise accordingly. Estimates more extreme than the MOV are infinite.

Infinite estimates of discriminability are awkward to deal with. For one thing, they cannot be plotted meaningfully on a graph. Also, if an infinite value is used when

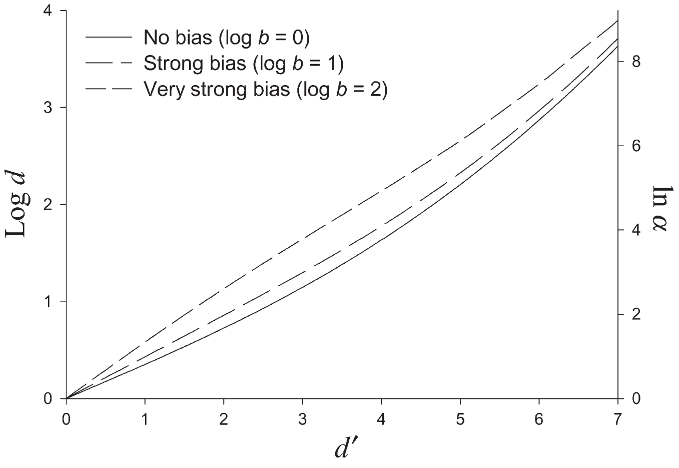

Figure 1. Log d (or, equally, ln α) as a function of d'. The horizontal axis is a logarithmic scale. The function is plotted for three different levels of bias. Bias is measured by log b, the logarithm of the geometric mean of the ratio of B_1 to B_2 responses in S_1 and S_2 trials (Davison & Tustin, 1978)—see Equation 4 in text. No bias (log b = 0) is shown by the solid line. Strong (log b = 1) and very strong (log b = 2) biases are shown by the long- and medium-dashed lines, respectively. Negative biases produce exactly the same functions as positive biases of the same magnitudes.

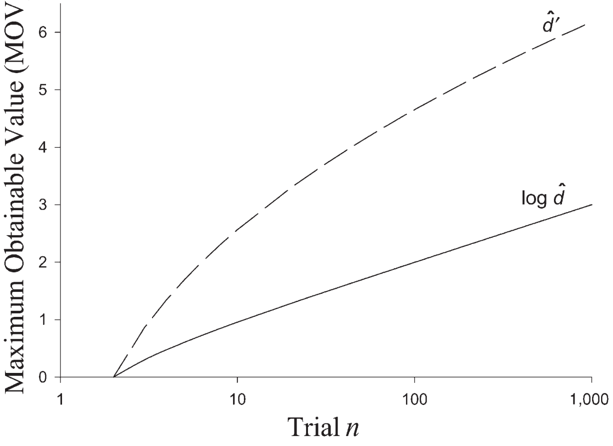

Figure 2. The maximum obtainable value (MOV) of log $\hat{d}$ (solid line) and $\hat{d}'$ (dashed line) as approximately linear functions of the logarithm of trial number. The MOV is the highest finite value possible.

calculating a mean, that mean will also be infinite, and this is true regardless of the number and size of the other data. Similarly, if mathematical fits are being applied to several related data points and one of those values is infinite, the fit will not accurately represent the number and magnitude of the other data points.

Methods of Dealing With Problem Cases

Run more trials. Sampling error dictates that, from time to time, infinite $\hat{d}'$ and log $\hat{d}$ estimates will occur. One of the most obvious solutions to this problem is to run more trials, as Figure 2 shows clearly. Because the MOV increases when more trials are added, infinite estimates must accordingly become less likely. This solution is almost always the best. It is only limited if there are practical reasons that no more trials can be run (for example, because of time constraints in running the experiment). Happily, only a few trials need to be added in order to increase a small MOV noticeably. On the flip side, though, a massive number of trials must be added to noticeably increase a large one. Figure 2 demonstrates this fact also, because the MOV increases almost linearly with the logarithm of trial number, not simply with trial number untransformed.

Rerun problem conditions. Another way to deal with infinite measurements is to rerun problem conditions if a zero occurs. Response counts from the problem condition are then either discarded or added to response counts from the condition's nonproblematic repetition. Either method is intuitively appealing, because response frequencies remain untransformed. In practice, though, problems arise either way. Inserting another condition can disrupt the order of conditions in an experiment, and, at the very least, it generates an unbalanced design in which exposure to the repeated conditions is greater than exposure to other conditions. If instead the problem data are combined with rerun data, however, that condition contains more trials than the other conditions. Miller (1996) showed that this is undesirable, since obtained discriminability estimates depend on the number of trials in these situations. Besides, pooling by addition is inappropriate unless the underlying discriminability and bias parameters remain almost exactly the same. (Bias deflates measurements of percent correct for similar reasons.) If the problem data are discarded, estimates of discriminability are mathematically biased toward lower values, as was demonstrated by Miller and is also shown in the present article.

Discard problem data points. Another way to deal with infinite discriminability estimates is to discard the problem data point. This strategy causes other problems, however. For example, when curves are fitted to discriminability values in memory procedures such as the delayed matching-to-sample task (see, e.g., White, 1985, 2001), the curve fit is more unreliable when a data point is removed. This is especially true when the data point is at the shortest delay, where discriminability tends to be highest. Ignoring the data point also produces problems when averaging performance across subjects or conditions. For example, in a situation of four perfect performances and one poor performance, the "average" measurement would indicate poor performance. Clearly, this does not summarize performance accurately.

Make the task more difficult. One can also avoid infinite discriminability estimates by making sure that the task is difficult. In that case, errors will be common, discriminability estimates will be low, and infinite estimates of discriminability will therefore be unlikely. Unfortunately, such a situation is sometimes not desirable. First, if a task were to be difficult for everybody, variation between subjects would dictate that the task could become virtually impossible for many subjects. Second, there are sometimes good reasons for wanting to investigate a range of discriminability values. For example, the effects of bias tend to be different when discriminability is high versus when it is low (White & Wixted, 1999). Nevertheless, if the experiment and its subjects permit, making the task more difficult is one of the preferred solutions.

Pool S_1 and S_2 data. McNicol (1972) suggested that problem cases can be dealt with by pooling data from S_1 and S_2 trials, which is possible if 0 errors occur for one type of trial but not for the other. By summing the response frequencies from S_1 and S_2 trials, the resulting measurement of discriminability is finite and is based purely on empirical data. This solution does pose two problems, however. First, it allows a greater range of discriminability values than do unpooled data; to treat all conditions fairly, response counts must be pooled in every condition and for every subject. Second, if response bias exists, this bias contaminates and thus lowers estimates of discriminability (Brown, 2003). When 0 errors occur on only one type of trial, performance on S_1 trials differs from that on S_2 trials, and response bias exists (at least empirically). If this asymmetry is due to sampling error alone, it is quite valid to use the pooled data. If, on the other hand, the asymmetry is due to true response bias, the pooled data can noticeably underestimate the value of d' or log d, particularly when there is a large difference between S_1 and S_2 trials. Since the source of asymmetry is generally unknown, it is perhaps prudent to assume that any pooled estimates are contaminated by bias and hence risk underestimating the true values of d' and log d. Pooling, therefore, is only acceptable if it is performed on every subject in every condition and if there is little difference between S_1 and S_2 trials.

Pool data across subjects. To avoid response counts of 0, it is also possible to pool data across subjects if subjects participate in the same conditions (see, e.g., White, 1985). Arguments similar to those for pooling across S_1 and S_2 trials apply: By pooling, the likelihood of a response count of 0 decreases, but if different subjects exhibit different levels of response bias, the pooled estimates underestimate both discriminability and response bias. Furthermore, it is difficult to submit pooled estimates to later analyses, because between-subjects variability cannot be calculated.

Mathematical Corrections of Extreme Measurements of Performance

When none of the above methods of dealing with infinite discriminability values are appropriate, different mathematical corrections can be used. In describing these corrections, we assume that adjustments are made to frequencies (not proportions) in cells of a signal detection matrix. We also assume that Equations 1, 2, and 3—which only use cell frequencies and not marginal totals—are used to calculate discriminability.

One of the simplest corrections is to replace any values of 0 in the signal detection matrix with the constant 1.0 (see, e.g., Jones & White, 1992; Watson & Blampied, 1988). For example, if there are 100 hits and 0 misses, using this correction results in 100 hits and 1 miss. A very similar rule is the ±0.5 rule proposed by Murdock and Ogilvie (1968), which replaces counts of 0 with 0.5 rather than 1.0. At the same time, however, it reduces response counts equal to N by 0.5 (N refers to the number of trials of a particular type). Thus, it would convert 100 hits and 0 misses to 99.5 hits and 0.5 misses. This is mathematically equivalent to Macmillan and Kaplan's (1985) $1/(2N)$ rule for proportions, which adds $1/(2N)$ to proportions of 0 and subtracts $1/(2N)$ from proportions of 1. If the proportion of responses is adjusted by $1/(2N)$, the corresponding frequency of responses is adjusted by $N \times 1/(2N) = 0.5$—that is, the ±0.5 rule. For the remainder of this article, we treat these two together as the ±0.5 rule. For log-linear models, Goodman (1970) also used a constant of 0.5, but he added it to all values in the matrix, regardless of their content and regardless of whether any cells contained 0. This correction converts 100 hits and 0 misses to 100.5 hits and 0.5 misses.

We categorize the above correction methods into two styles. The first adds a constant to all cells, regardless of content, as in Goodman's (1970) technique. The second, exemplified by all the other rules, adds a constant only to cells containing 0. Some rules in this category also reduce response counts equal to N in order to leave the marginal totals unchanged. In practice, however, this has virtually no effect on $\hat{d}'$ or log $\hat{d}$ unless there are very few trials indeed.[2] Thus, the ±0.5 rule is almost identical to simply replacing counts of 0 with 0.5. Beyond the two categories, the only other difference between correction methods is the constant used in the correction. All but one of the above techniques uses a constant of 0.5. This choice of constant is largely arbitrary, although Kadlec (1999) argues that because response counts are whole numbers, a count of 0 represents any real value up to 0.5. To date, however, there has been no systematic examination of whether other constants produce better estimates of performance.

Previous Comparisons of Correction Procedures

Using a series of Monte Carlo simulations, Hautus (1995) compared the effects of the ±0.5 and the Goodman (1970) rules on estimates of d' (see also Hautus, 1997; Hautus & Lee, 1998). He found that both techniques estimated d' poorly when there were very few trials. He also demonstrated that the Goodman rule usually produced better estimates of d' than the ±0.5 rule. The Goodman rule was also more conservative, in that it consistently underestimated d', whereas the ±0.5 rule often overestimated it.

Kadlec (1999) also used Monte Carlo simulations to examine four different correction procedures: the ±0.5 rule, the Goodman (1970) rule, a Murdock and Ogilvie (1968) rule in which 0.0001 was the correction constant (after Miller, 1996, who used this rule to show the effects of arbitrary constants), and discarding problem conditions (which is analogous to rerunning and replacing problem conditions). Kadlec found that the ±0.0001 correction often overestimated d' massively. Discarding problem conditions, on the other hand, tended to underestimate d' the most. The ±0.5 rule and the Goodman rule performed similarly and estimated d' the best.

Miller (1996) used a different technique to demonstrate how the different corrections affect estimates of d'. Rather than conducting simulations, he directly computed sampling distributions of $\hat{d}'$. This is a simple but computationally intensive technique based on the idea that the numbers of correct and incorrect responses for S_1 and S_2 trials follow a binomial distribution. If the true underlying proportions of correct S_1 and S_2 trials are known, as well as the number of trials of each type, then it is possible to calculate the likelihood of obtaining a given measurement of $\hat{d}'$. As a consequence, the full sampling distribution of $\hat{d}'$ and its mean and variance can be determined.

Miller (1996) fully explained how to directly compute the sampling distribution of $\hat{d}'$. In demonstrating this technique, he also investigated three different correction procedures: the ±0.5 rule, the ±0.0001 rule, and discarding problem conditions. Like Hautus (1995) and Kadlec (1999), Miller found that all techniques estimated d' poorly when there were very few trials. Miller also showed that the ± 0.5 rule often overestimated d', just as Hautus and Kadlec had. Furthermore, like Kadlec, Miller showed that the ± 0.0001 rule tended to markedly overestimate d'. Thus, both Miller's computational technique and the Monte Carlo simulations of Kadlec and Hautus reached similar conclusions.

An important finding from these studies was that rerunning problem conditions produced worse estimates of d' than did the ± 0.5 rule (Kadlec, 1999; Miller, 1996). An arbitrary mathematical correction therefore outperformed an intuitively plausible correction procedure that did not transform response counts. This suggests that an in-depth investigation of different mathematical correction procedures could be worthwhile. Even taken together, though, Hautus's (1995), Kadlec's, and Miller's analyses do not represent a comprehensive comparison of the different correction procedures. We attempt to perform such a comparison here.

COMPARISONS OF CORRECTION PROCEDURES

Method and Rationale

Our comparisons of the different correction procedures were based on Miller's (1996) technique for directly computing the sampling distribution of $\hat{d}'$. Using the same logic, we were able to modify Miller's equations to also compute the sampling distribution of log $\hat{d}$. These equations are shown in the Appendix. It is useful to examine estimations of log d because, although it is not as popular as d', log d is much easier to calculate, and its underlying theory (Davison & Tustin, 1978) treats bias more realistically. Log d divides bias into different components: inherent bias (a general preference for one choice over another) and bias generated by response payoffs (McCarthy & Davison, 1981). Another reason to examine log d is that, as shown earlier, its calculation is almost identical to that of ln α.

To calculate a sampling distribution of either $\hat{d}'$ or log $\hat{d}$, we began by assuming an underlying probability of hits (B_2 on S_2 trials) and false alarms (B_2 on S_1 trials). These probabilities are jointly determined by underlying discriminability and bias parameters. High discriminability produces more hits but fewer false alarms, and bias toward B_2 (for instance) inflates both hits and false alarms. Although the hit and false alarm probabilities are stable (in theory), chance alone dictates that the obtained number of hits and false alarms will vary. In fact, because the choice between B_1 and B_2 is binary, the obtained values vary according to a binomial distribution. It is thus possible to calculate the likelihood of obtaining a particular number of hits or false alarms. To find the probability of obtaining a specific combination of these values, their independent probabilities are multiplied together, and this can be done for each possible combination of hits and false alarms. Because each combination has an associated $\hat{d}'$ or log $\hat{d}$ value, the resulting table of combinations and probabilities defines the sampling distribution of $\hat{d}'$ or log $\hat{d}$.

To determine how well a given correction procedure works, we compared the means of sampling distributions of $\hat{d}'$ and log $\hat{d}$ to their true values. The mean of the sampling distribution (i.e., its expected value) is useful because this measure takes into account not only each possible measured value, but also that value's likelihood. When a specific measurement of extreme discriminability is both probable and substantially different from true discriminability because of the correction procedure used, the mean of the sampling distribution clearly misestimates the true value of d' or log d. This misestimation represents the amount of *statistical bias*—that is, the expected difference between measured and true discriminability.

The sampling distributions of $\hat{d}'$ and log $\hat{d}$ rely on many factors. These include the true underlying d' or log d, the level of response bias, the number of trials, the correction technique used, and the correction constant used. We varied all of these factors systematically to determine how they influence the sampling distributions of $\hat{d}'$ and log $\hat{d}$, and hence the estimates of discriminability.

Effects of Different Correction Procedures on Estimates of Discriminability

Figure 3 contains nine graphs that all compare the mean of the sampling distribution of $\hat{d}'$ to its true underlying value for three different levels of bias. Bias was measured by the response criterion c (Macmillan & Creelman, 1991). We refer to the mean of the sampling distribution as *expected* $\hat{d}'$ and the true underlying value as *true* d'. In each graph in Figure 3, true d' is shown on the horizontal axis. For each true d', the corresponding expected $\hat{d}'$ is plotted on the vertical axis. To make it easier to compare the two, true d' is also plotted on both axes, and thus appears as a straight diagonal line. In each graph, the expected $\hat{d}'$ closely matches true d' when true d' is low. As it nears the MOV (represented by the horizontal line), expected $\hat{d}'$ underestimates the true value more and more. This underestimation is more pronounced, and begins at lower values of true d', when there is more bias.

The nine graphs in Figure 3 are separated into three rows and three columns. The three different rows are for three different numbers of trials. Here, S_1 and S_2 trials each number 32, 128, and 512. The three different columns represent three different conventions for dealing with problem cases (i.e., cases in which one or more cells in the signal detection matrix contain 0). In the first column, the Goodman (1970) rule is applied: 0.5 was added to all cells in the signal detection matrix, regardless of their content. In the second column, 0.5 is added only to cells containing 0, but response counts equal to N are never adjusted. This has effects virtually identical to those of the ± 0.5 rule, but if anything this correction has a tiny advantage in that it very slightly increases the MOV. It is also simpler. We will therefore dispense with analyses of the pure ± 0.5 rule for reasons of economy, but identi-

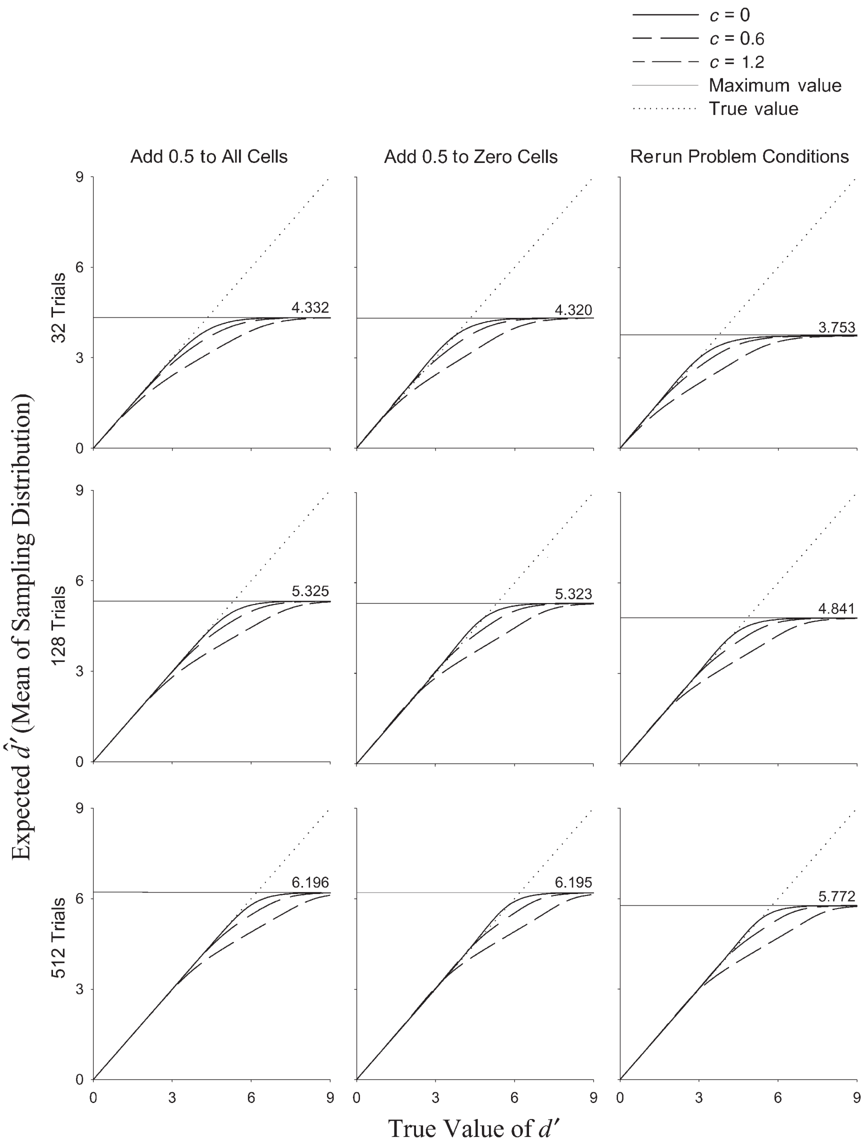

Figure 3. Comparisons of expected $\hat{d}'$ (the mean of the sampling distribution of $\hat{d}'$) with the true underlying value of d'. Functions are shown for 32 trials (top row), 128 trials (middle row), and 512 trials (bottom row) of each type. Functions are also differentiated according to whether the correction procedure added 0.5 to all cells (left column), added 0.5 only to cells containing 0 (middle column), or excluded problem conditions (right column). Functions are plotted for three different levels of bias, as measured by response criterion c: no bias ($c = 0$, solid lines), moderate bias ($c = 0.6$, long-dashed lines), and strong bias ($c = 1.2$, medium-dashed lines). The horizontal line on each graph represents the maximum value of $\hat{d}'$. The dotted diagonal line represents equality between expected $\hat{d}'$ and true d'.

cal conclusions will apply. In the third column, problem cases are excluded. This is analogous to discarding data from problem conditions and rerunning those conditions (Miller, 1996).

For the different conventions, response bias produces about the same underestimation of d', regardless of the correction procedure used. Also, the MOV is almost exactly the same whether the Goodman method is used or 0.5 is added only to cells that contain 0. This is not surprising, since the lowest value in the signal detection matrix is 0.5 for both corrections. There is a tiny difference between the two simply because of the difference in the number of cells in the signal detection matrix to which the two methods add 0.5. In comparison with these two methods, rerunning problem conditions always yields a lower MOV, simply because the lowest possible value in the signal detection matrix, and hence the lowest possible denominator in a ratio, is 1.0 rather than 0.5.

Figure 3 reveals an important difference between the Goodman convention and the other two methods. With the Goodman convention, expected $\hat{d}'$ exclusively underestimates true d'. Reassuringly, this is consistent with Hautus's (1995) results based on Monte Carlo simulations. In contrast, the other two conventions tend to slightly overestimate true *d′* in many instances (at least when bias is negligible). The overestimation is most pronounced when the values of true *d′* begin to approach the MOV. The conclusion is that the correction rule is less likely to overestimate true *d′* when it adds a constant to all cells than when it adds a constant only to cells containing 0 or when it requires problem conditions to be rerun. Another key observation in Figure 3 is that the relationship between the correction procedures remains the same regardless of the number of trials.

Figure 4 plots information similar to that found in Figure 3. The main difference is that it represents log *d* rather than *d′*. Also, it measures response bias in terms of log *b* (Davison & Tustin, 1978; see Equation 4 below) rather than the response criterion, *c*. The level of bias is roughly equivalent, however. Log *b* is symmetrical with log *d* in

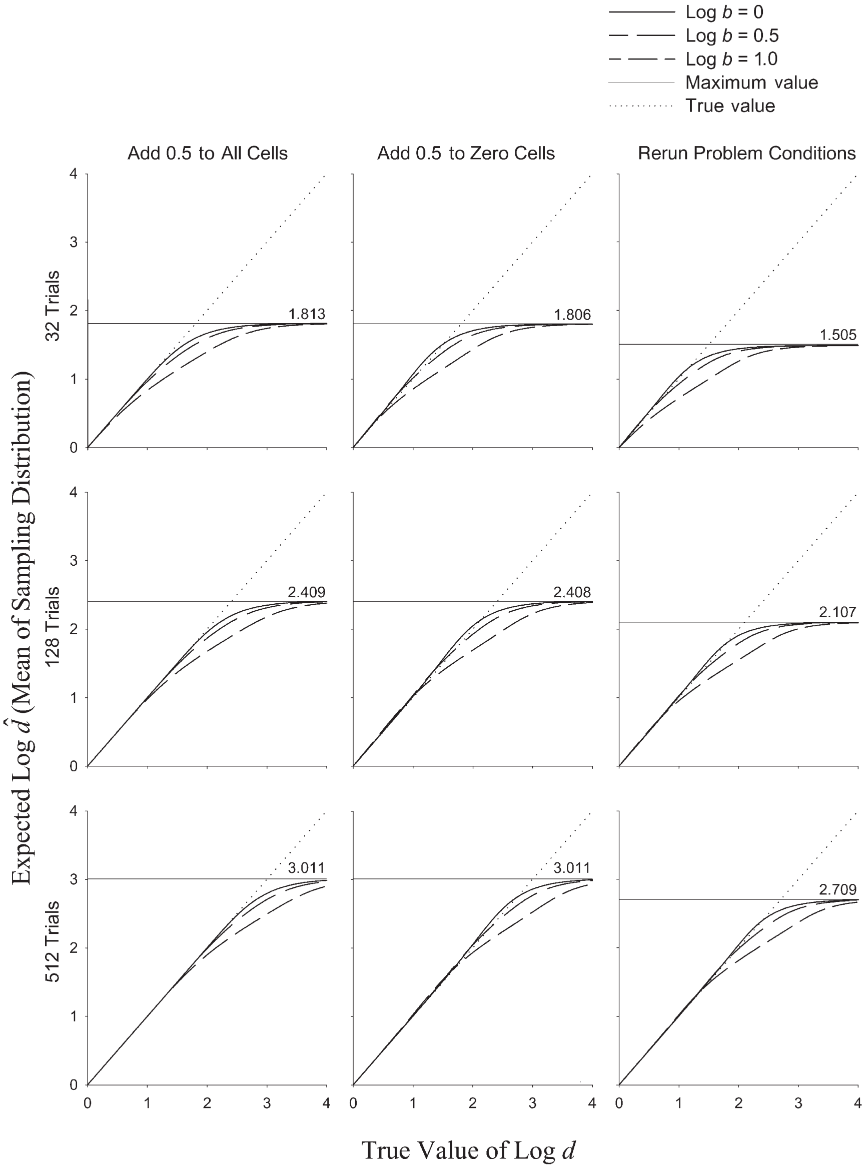

Figure 4. Comparisons of expected log $\hat{d}$ (the mean of the sampling distribution of log $\hat{d}$) with the true underlying value of log *d*. Functions are shown for 32 trials (top row), 128 trials (middle row), and 512 trials (bottom row) of each type. Functions are also differentiated according to whether the correction procedure added 0.5 to all cells (left column), added 0.5 only to cells containing 0 (middle column), or excluded problem conditions (right column). Functions are plotted for three different levels of bias, as measured by log *b*: no bias (log *b* = 0, solid lines), moderate bias (log *b* = 0.5, long-dashed lines), and strong bias (log *b* = 1.0, medium-dashed lines). The horizontal line on each graph represents the maximum value of log $\hat{d}$. The dotted diagonal line represents equality between expected log $\hat{d}$ and true log *d*.

Equation 2 and is identical to the measure of bias derived from Luce's (1963) choice theory, except that it uses a base-10 logarithm rather than the natural logarithm.

$$\log b = \frac{1}{2}\cdot\log_{10}\left(\frac{\text{Hits}}{\text{Misses}}\cdot\frac{\text{False Alarms}}{\text{Correct Rejections}}\right). \quad (4)$$

The relationships shown for log *d* in Figure 4 are almost exactly the same as those shown for *d′* in Figure 3. This is not surprising, given the roughly linear relationship between *d′* and log *d* shown in Figure 1. The main point of Figure 4 is to show that when investigating the different conventions, it matters little whether *d′* or log *d* is used. Because of this, we simplify the remainder of the present article by concentrating mostly on only one measure, *d′*. We further simplify matters by presenting results only for 128 trials apiece for S_1 and S_2. In Figures 3 and 4,

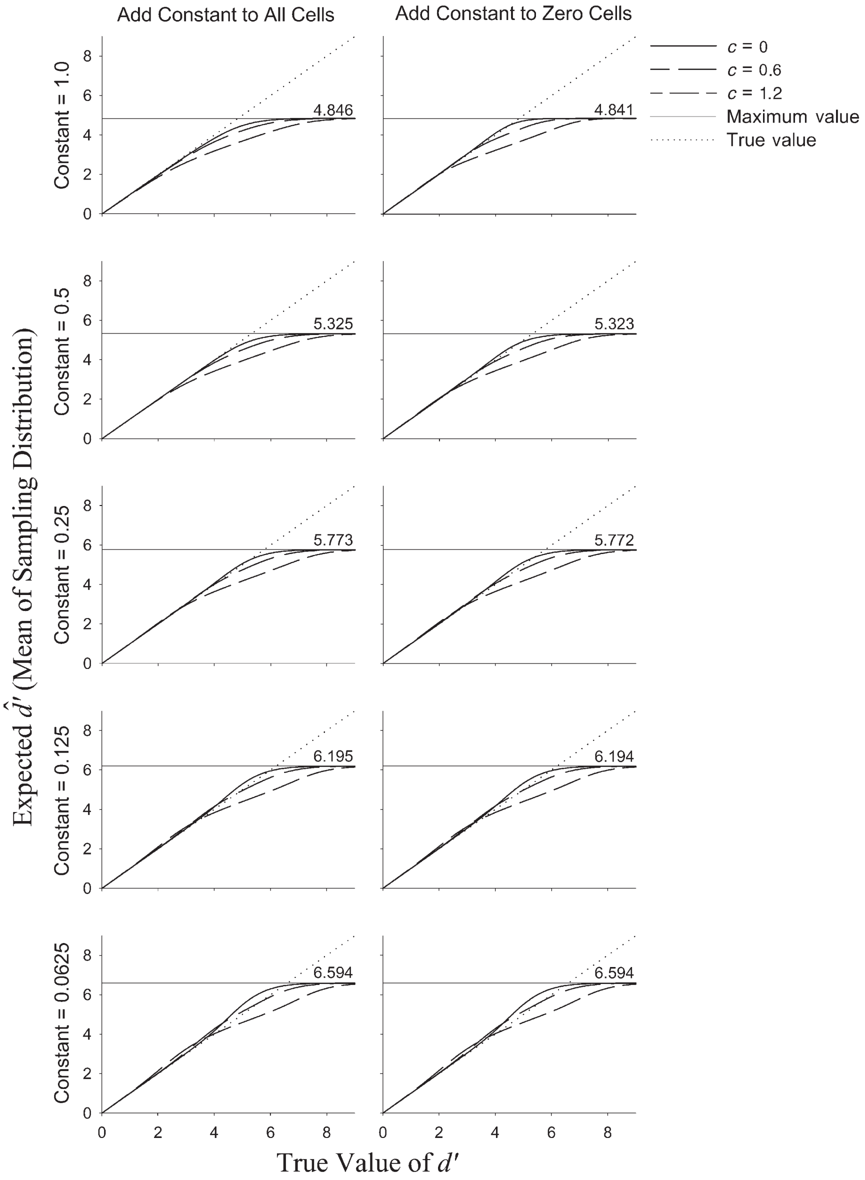

Figure 5. Comparisons of mean expected $\hat{d}'$ with true *d′* for different correction procedures, correction constants, and levels of bias, for 128 trials of each type. Functions are shown for adding correction constants to all cells (left column) or only to cells containing 0 (right column). Functions are also differentiated according to correction constant: 1.0 (top row), 0.5 (second row), 0.25 (middle row), 0.125 (next-to-bottom row), and 0.0625 (bottom row). Functions are plotted for three different levels of bias, as measured by response criterion *c*: no bias ($c = 0$, solid lines), moderate bias ($c = 0.6$, long-dashed lines), and strong bias ($c = 1.2$, medium-dashed lines). The horizontal line on each graph represents the maximum value of $\hat{d}'$. The dotted diagonal line represents equality between expected $\hat{d}'$ and true *d′*.

comparisons between the conventions produced the same conclusions, regardless of the number of trials (see also Figure 7).

Effects of Different Constants on Estimates of Discriminability

Figures 3 and 4 show that the MOV is determined by the number of trials and by the size of the smallest value in the signal detection matrix. For most corrections, the smallest value is the constant added to cells containing 0, which is 0.5 in the case of the ±0.5 and Goodman rules. The choice of 0.5 as a constant is largely arbitrary, although Kadlec (1999) did provide some intuitive justification. Nonetheless, the effect of using different constants has not been systematically studied. We have done just that, however, by testing the effects of different correction constants on the sampling distribution of $\hat{d}'$. Figure 5 contains 10 graphs of the same style as those contained in Figure 3. They compare true d' to expected $\hat{d}'$ for 128 trials apiece for S_1 and S_2. In the left column, correction constants were added, Goodman style, to all cells in the signal detection matrix, regardless of content. In the right column, correction constants were applied only to cells containing 0. The five rows of Figure 5 systematically vary the constant from a high value of 1.0 to a low value of 1/16 (0.0625).

Figure 5 shows that smaller correction constants produce higher MOVs of $\hat{d}'$. Taken alone, this gives smaller correction constants an advantage: They are less likely to underestimate d' when true d' is very high. Not surprisingly, however, they also have a disadvantage: At less extreme values of true d', and when bias is small, they produce greater overestimation. The overestimation appears to be slightly more pronounced when the constant is added only to cells containing 0 rather than to all cells.

THE OPTIMAL CORRECTION PROCEDURE

An optimal correction procedure should balance two conflicting goals. The first is to possess a high MOV, since doing so permits a broader range of scores. A high MOV also means that underestimates of d' become less likely and that an equal-interval scale of discriminability is preserved for a wider range of values. The second goal is to ensure that d' values are not greatly overestimated. These two goals necessarily oppose each other. Given that perfect scores can be difficult to interpret and that lower scores are generally more likely, it is perhaps prudent to focus first on the latter, more conservative goal.

To decide upon the best correction constant, three questions must be answered. First, according to the criteria above, which is the best style of correction? Our analyses show that this is really a choice between a Goodman-style method of adding a constant to all cells or, alternatively, adding a constant only to cells that contain 0 (much like the ±0.5 technique). Second, does the number of trials influence which is the best correction constant to use? Finally, if the number of trials does not matter, which is the best choice of correction constant?

Method for Finding the Maximum Mean Overestimation

Since it is important to avoid excessive overestimation, an important measure of a correction method's success is the maximum mean overestimation that it produces (that is, the largest amount by which d' is overestimated by the mean of its sampling distribution, expected $\hat{d}'$). We found the maximum mean overestimation for each constant and correction procedure. To do this, we used a series of recursive numerical searches. Since the largest overestimations occur when there is no bias, bias was held at zero. For each search, true d' was increased incrementally and compared with expected $\hat{d}'$ each time. This eventually revealed the two values of true d' that produced the two largest mean overestimations. A search using smaller increments of true d' was then conducted within those two d' values, a search by even smaller increments within the resulting

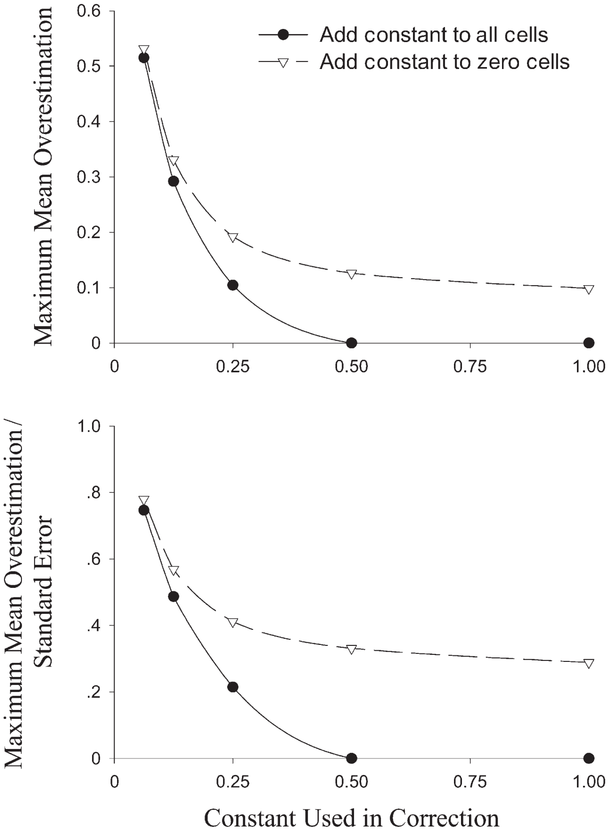

Figure 6. For 128 trials of each type, the maximum amount by which expected $\hat{d}'$ (the mean of the sampling distribution of $\hat{d}'$) overestimates true d' for different correction constants. The overestimation is expressed absolutely (top panel) or as a proportion of the standard error (bottom panel). Both graphs compare adding the correction constant to all cells (solid lines, filled circles) with adding the constant only to cells containing 0 (dashed lines, open triangles).

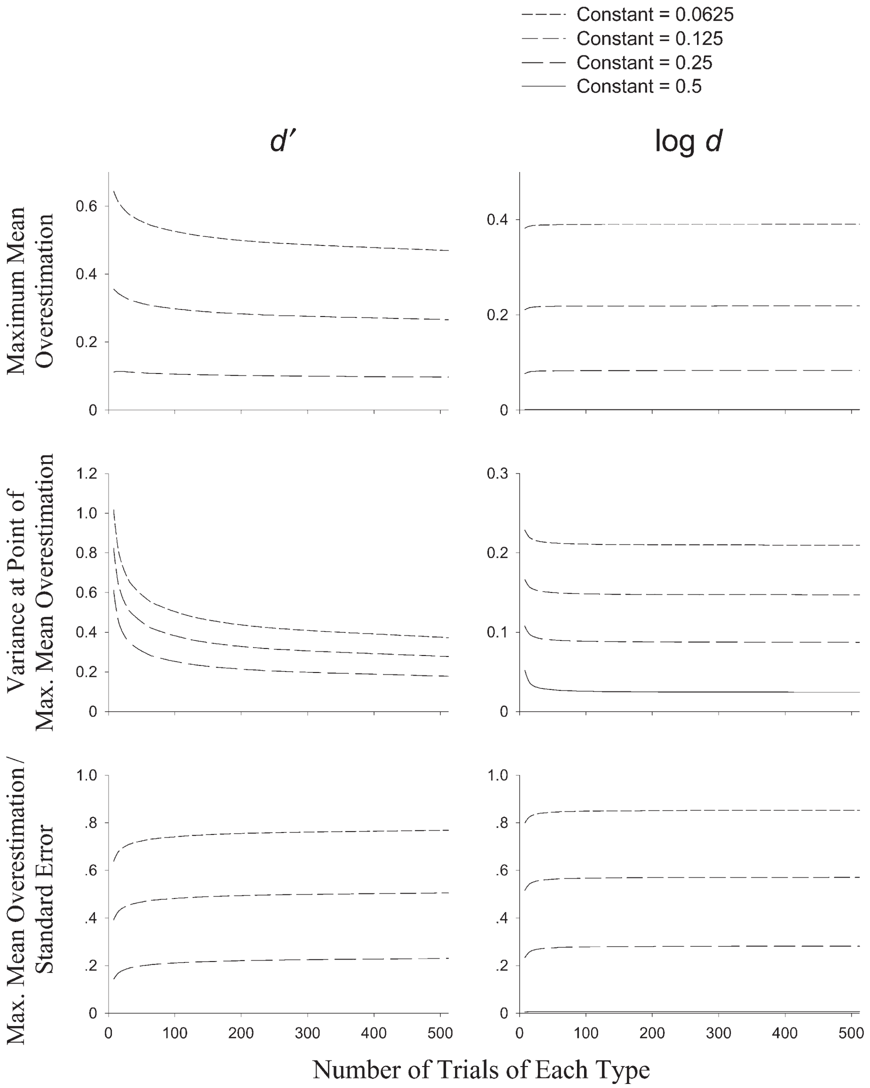

Figure 7. The effect of the number of trials of each type on overestimations of *d′* (left column) and log *d* (right column). The top panels show the maximum amount by which expected $\hat{d}'$ or log $\hat{d}$ overestimates true *d′* or log *d* for different trial *n*s. The middle panels show how the variance at the point of maximum overestimation varies with trial number. The bottom panels illustrate the effect of trial number on the maximum overestimation as a proportion of the standard error. Functions are shown for different correction constants: 0.0625 (short-dashed line), 0.125 (medium-dashed line), 0.25 (long-dashed line), and 0.5 (solid line). These correction constants were added to all cells, regardless of their value. When the correction constant was 0.5, there was no mean overestimation of *d′*, so the corresponding functions are not shown. With the same correction constant, the maximum mean overestimation of log *d* was almost 0, and is thus almost indistinguishable from the abscissa.

values, and so on. The recursion continued until the two values of true *d′* differed from each other by less than 0.0001. (It should be noted that the search was broadened if either the highest or the lowest true *d′* within a range produced the biggest overestimation.) The resulting value represented the true *d′* that was most overestimated by the mean of its sampling distribution.

The Best Style of Correction

We used the data in Figure 6 to decide which style of correction is better: adding a constant to all cells or adding a constant to cells containing 0. The top panel of Figure 6 shows how the different correction constants and correction procedures affect the maximum mean overestimation of *d′*. These calculations are based on 128 trials apiece for S_1 and S_2, but a similar pattern of results applies irrespective of the number of trials. The bottom panel of Figure 6 shows these same values expressed as a proportion of the standard error of the sampling distribution of $\hat{d}'$ at the point of maximum overestimation. These proportions are important because they show the size of the overestimation relative to the spread of the sampling distribution. Large overestimations are more acceptable when estimates vary wildly anyway (i.e., if the sampling distribution is well spread and the standard error is large). This is because, in context, the overestimation is then very difficult to detect and is therefore of little practical significance. For similar reasons, a small mean overestimation may be unacceptable if the standard error is tiny.

Figure 6 shows that when the smallest correction constants are used, the highest maximum mean overestimation results, and is similar for both styles of correction. When larger correction constants are used, less overestimation results, and a difference between the correction styles becomes clear. Specifically, adding a constant to all cells produces less overestimation than does adding a constant only to cells that contain 0. This is because a Goodman-style procedure moderates not only the most extreme measurements of $\hat{d}'$, but to a lesser extent measurements that are not quite as extreme. To avoid overestimation while maintaining a high MOV, then, the Goodman-style procedure of adding a constant to all cells is superior to adding a constant only to cells that contain 0.

Kadlec (1999) did not agree that the Goodman correction was superior, but concluded that the ± 0.5 rule outperformed the Goodman correction. She did not, however, analyze the maximum mean overestimation, as we did. Furthermore, her preference arose only because the ± 0.5 rule was slightly better for estimating the likelihood ratio β (see Macmillan & Creelman, 1991). Here, our focus is on how best to estimate d', and for this purpose, a Goodman-style correction is superior. Two questions remain, however: What is the best correction constant to use? And does it depend on the number of trials?

Does Trial *n* Determine the Optimal Correction Constant?

To help evaluate whether or not the number of trials determines the best correction constant to use, Figure 7 shows how the maximum mean overestimation, its variance, and its size with respect to the standard error change when the number of trials is increased. These values are plotted separately for four different correction constants: 0.5, 0.25, 0.125, and 0.0625. They are also plotted separately for d' and log d. (We show both here because their patterns differ slightly.) With log d, for all but the smallest number of trials (less than about 20), there is almost no influence of trial number on any of the measurements, no matter which constant is used. This fact justifies using the same constant regardless of the number of trials. (It does not, however, suggest that running few trials is as good as running many: When few trials are used, the MOV is still much lower.)

When d' is used, the graphs in Figure 7 are a little harder to interpret. As long as the trial n is greater than 100, maximum mean overestimation—whether expressed absolutely or as a proportion of the standard error—is about the same regardless of the number of trials. As trial n drops below 100, however, the maximum mean overestimation increases substantially for correction constants less than 0.25. For all correction constants, variance increases considerably with the same drop in trial n. For this reason, like Kadlec (1999), we caution against using d' with fewer than 100 trials of each type, especially if a correction constant of less than 0.25 is used. When trial N is necessarily small, log d may be a more reliable measure.

The Optimal Correction Constant

To decide on the optimal correction constant is largely to decide where to strike the balance between underestimation of the highest discriminability values and overestimation of less extreme values. With little doubt, the upper limit for the correction constant should be 0.5. This produces virtually no mean overestimation and is thus very conservative. With smaller correction constants, however, there is an increased likelihood of overestimation, and hence an increased likelihood of Type I errors. To avoid assigning an overly arbitrary lower limit for the correction constant, it is useful to look at what response counts logically represent. Response counts of 1 represent all real values between 0.5 and 1.5 as their halfway point, and response counts of 0—perhaps counterintuitively—represent all real values between 0 and 0.5 (Kadlec, 1999). The latter range might be better represented by its own halfway point, 0.25. It is difficult to find a logical reason for using a value lower than 0.25.

Visually, Figure 5 confirms that 0.25 could be a suitable correction constant. It maintains an almost linear relationship between expected $\hat{d}'$ and true d' for a wide range of values. Furthermore, its MOV is higher than that produced by a correction constant of 0.5. At the same time, however, Figure 6 shows that it still produces less overestimation than the widely used ± 0.5 rule. The same is not true when the correction constant is much lower than 0.25, however. Further support for a correction constant of 0.25 is demonstrated in Figure 7, where the maximum mean overestimation produced by a correction constant of 0.25 does not exceed 25% of the standard error. Any overestimation would be very difficult to detect statistically, and thus would be of little practical significance. We therefore conclude that any correction constant between 0.25 and 0.5 is acceptable.

SUMMARY AND RECOMMENDATIONS

In yes–no tasks, discriminability can be measured using d', log d, or ln α. When performance is very good or when there are very few trials, infinite estimates of discriminability are likely. Such measurements are problematic because they cannot easily be plotted on a graph, used in calculations of means, or employed in curve-fitting procedures. Moreover, it is theoretically questionable whether discriminability can truly be infinite.

Infinite estimates are best avoided by running more trials. If it is not viable to do so, researchers could instead make the task more difficult, rerun problem conditions, discard problem data points, pool S_1 and S_2 data, or pool data across subjects. Unfortunately, these options are sometimes inappropriate and often problematic. In such cases, it is best to apply a mathematical correction.

We compared the different types of mathematical correction procedures. Our analyses indicated that it is best to add a constant to all cells in the signal detection matrix, regardless of their content. The constant used should be the same, regardless of the number of trials. However, using d' with fewer than about 100 trials of each type (see Kadlec, 1999) is not recommended. If low trial N is unavoidable, log d or ln α may be more reliable measures.

We concluded that it is best to use a correction constant between 0.25 and 0.5. This will result in less overestimation than is produced by the widely used ±0.5 rule [also known as the $1/(2N)$ rule for proportions]. At the same time, unavoidable underestimates of discriminability will be as small as possible. Any correction constant between 0.25 and the more commonly used 0.5 produces an acceptable level of overestimation (Figure 7). A correction constant of 0.5 remains satisfactory, as has been confirmed by prior analyses (Hautus, 1995; Kadlec, 1999), but 0.25 permits a wider range of discriminability estimates.

Typically, experiments have a small number of trials and have measured d' values in the range 0–3.0, such as in studies of recognition memory. In these circumstances, our suggested correction method produces the best estimates of discriminability of all of the techniques examined and is also the least sensitive to response bias. It should be noted, however, that when bias is particularly strong ($c > \pm 0.5$, approximately), all correction methods tend to underestimate discriminability.

As a final note, caution should be exercised when any mathematical correction is used. Although such corrections allow estimates to be finite and as accurate as possible, they still only represent a "best guess." If the conclusions of an experiment or analysis depend heavily on the mathematical correction used, then using other methods to avoid undefined measurements is essential, especially when response bias is strong.

REFERENCES

Alsop, B., Rowley, R., & Fon, C. (1995). Human symbolic matching-to-sample performance: Effects of reinforcer and sample-stimulus probabilities. *Journal of the Experimental Analysis of Behavior*, **63**, 53-70.

Brown, G. S. (2003). *Memory and reinforcement.* Unpublished doctoral thesis, University of Otago, Dunedin, NZ.

Davison, M. C., & Tustin, R. D. (1978). The relation between the generalized matching law and signal-detection theory. *Journal of the Experimental Analysis of Behavior*, **29**, 331-336.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, **65**, 226-256.

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.

Hautus, M. J. (1997). Calculating estimates of sensitivity from group data: Pooled versus averaged estimators. *Behavior Research Methods, Instruments, & Computers*, **29**, 556-562.

Hautus, M. J., & Lee, A. J. (1998). The dispersions of estimates of sensitivity obtained from four psychophysical procedures: Implications for experimental design. *Perception & Psychophysics*, **60**, 638-649.

Jones, B. M., & White, K. G. (1992). Sample-stimulus discriminability and sensitivity to reinforcement in delayed matching to sample. *Journal of the Experimental Analysis of Behavior*, **58**, 159-172.

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

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, **6**, 312-319.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103-189). New York: Wiley.

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

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.

McCarthy, D., & Davison, M. [C.] (1981). Towards a behavioral theory of bias in signal detection. *Perception & Psychophysics*, **29**, 371-382.

McNicol, D. (1972). *A primer of signal detection theory.* London: Allen & Unwin.

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

Murdock, B. B., Jr., & Ogilvie, J. C. (1968). Binomial variability in short-term memory. *Psychological Bulletin*, **70**, 256-260.

Watson, J. E., & Blampied, N. M. (1988). Quantification of the effects of chlorpromazine on performance under delayed matching to sample in pigeons. *Journal of the Experimental Analysis of Behavior*, **51**, 317-328.

White, K. G. (1985). Characteristics of forgetting functions in delayed matching to sample. *Journal of the Experimental Analysis of Behavior*, **44**, 15-34.

White, K. G. (2001). Forgetting functions. *Animal Learning & Behavior*, **29**, 193-207.

White, K. G., & Wixted, J. T. (1999). Psychophysics of remembering. *Journal of the Experimental Analysis of Behavior*, **71**, 91-113.

Wixted, J. T. (1990). Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 927-935.

NOTES

[1] In this article, symbols with "hats" (e.g., $\hat{d}'$) are data-based estimates of true, underlying parameter values, which are shown without hats (e.g., d'). These "true" underlying parameters are theoretical constructs.

[2] Even with a mere 10 trials of each type, subtracting 0.5 from values equal to N has an effect of less than 2% (0.047 for $\hat{d}'$ and 0.022 for log $\hat{d}$). With 100 trials of each type—the minimum recommended by Kadlec (1999)—the difference is less than 0.01% (0.0034 for $\hat{d}'$ and 0.0022 for log $\hat{d}$). Thus, the $1/(2N)$ and ±0.5 rules are virtually identical to adding 0.5 only to cells that contain 0.

(Continued on next page)

APPENDIX
Computations for the Sampling Distribution of Log $\hat{d}$

Miller (1996) described how to generate the sampling distribution of $\hat{d}'$ and calculate its mean and variance. We adapted Miller's equations to enable us to perform analogous calculations for the sampling distribution of log $\hat{d}$. We have generally retained Miller's notation and will explain our equations in the same order he used. The important assumption is that performance on S_1 trials is independent of performance on S_2 trials. Given this assumption, calculations on S_1 and S_2 trials can be performed separately. These independent calculations can then be combined in order to examine overall performance.

B_1 and B_2 are alternative responses on each trial. After Miller (1996), we used the binomial distribution to calculate how likely it is that $N_{B_1|S_1}$, a particular number of B_1 responses on S_1 trials (correct rejections), will be obtained:

$$\Pr\left(N_{B_1|S_1} = k\right) = \binom{N_{S_1}}{k} p_{B_1|S_1}^{k}\left(1 - p_{B_1|S_1}\right)^{N_{S_1} - k}, \quad k = 0, 1, \ldots, N_{S_1}, \tag{A1}$$

where $p_{B_1|S_1}$ is the true, underlying probability of obtaining B_1 on an S_1 trial and N_{S_1} is the number of S_1 trials. Put in words, Equation A1 states that the number of obtained B_1 responses on S_1 trials follows a binomial distribution of probability $p_{B_1|S_1}$ and N_{S_1} observations. The likelihood of obtaining a particular number of B_2 responses on S_2 trials (hits) is expressed by a similar equation:

$$\Pr\left(N_{B_2|S_2} = k\right) = \binom{N_{S_2}}{k} p_{B_2|S_2}^{k}\left(1 - p_{B_2|S_2}\right)^{N_{S_2} - k}, \quad k = 0, 1, \ldots, N_{S_2}. \tag{A2}$$

On S_1 trials, the log of correct to incorrect responses can be calculated directly from $N_{B_1|S_2}$ and N_{S_1}, and on S_2 trials it can be calculated from $N_{B_2|S_2}$ and N_{S_2}. Because of this, Equations A1 and A2 can easily be adapted to compute the probability of obtaining a particular log ratio of correct to incorrect responses on S_1 or S_2 trials. The resulting equations are

$$\Pr\left[\log\left(\frac{\hat{p}_{B_1|S_1}}{1 - \hat{p}_{B_1|S_1}}\right) = \log\left(\frac{k}{N_{S_1} - k}\right)\right] = \binom{N_{S_1}}{k} p_{B_1|S_1}^{k}\left(1 - p_{B_1|S_1}\right)^{N_{S_1} - k}, \quad k = 0, 1, \ldots, N_{S_1} \tag{A3}$$

and

$$\Pr\left[\log\left(\frac{\hat{p}_{B_2|S_2}}{1 - \hat{p}_{B_2|S_2}}\right) = \log\left(\frac{k}{N_{S_2} - k}\right)\right] = \binom{N_{S_2}}{k} p_{B_2|S_2}^{k}\left(1 - p_{B_2|S_2}\right)^{N_{S_2} - k}, \quad k = 0, 1, \ldots, N_{S_2}, \tag{A4}$$

where $p_{B_1|S_1}$ is the obtained probability of B_1 on S_1 trials and $p_{B_2|S_2}$ is the obtained probability of B_2 on S_2 trials. Note that for ln α rather than for log d, base-e rather than base-10 logs would be used in all equations here.

For the purposes of generating a sampling distribution, the true probability of $p_{B_1|S_1}$ and $p_{B_2|S_2}$ can be calculated directly from the underlying values of log d and log b. Assuming that log d represents a bias toward correct responses and log b represents a bias toward B_1 responses, the appropriate equations are

$$p_{B_1|S_1} = \frac{10^{\log d} \cdot 10^{\log b}}{1 + \left(10^{\log d} \cdot 10^{\log b}\right)} \tag{A5}$$

and

$$p_{B_2|S_2} = \frac{10^{\log d}}{10^{\log b} + 10^{\log d}}. \tag{A6}$$

Equations A5 and A6 are derived from Equations 2 and 4 in the text. Note that if ln α is being used, e rather than 10 should be raised to the power of the appropriate bias and discriminability terms.

In the present article, the calculation of the mean of the sampling distribution of log $\hat{d}$ is very important. To compute it, the first step is to calculate the expected value of the log of correct to incorrect responses for S_1 and S_2 trials independently. For each trial type, this is accomplished by multiplying each possible value of the log of correct to incorrect responses by its probability. The appropriate equations are

$$\mathrm{E}\left[\log\left(\frac{\hat{p}_{B_1|S_1}}{\hat{p}_{B_2|S_1}}\right)\right] = \sum_{k=0}^{N_{S_1}} \log\left(\frac{k}{N_{S_1} - k}\right)\binom{N_{S_1}}{k} p_{B_1|S_1}^{k}\left(1 - p_{B_1|S_1}\right)^{N_{S_1} - k}, \quad k = 0, 1, \ldots, N_{S_1} \tag{A7}$$

and

$$\mathrm{E}\left[\log\left(\frac{\hat{p}_{B_2|S_2}}{\hat{p}_{B_1|S_2}}\right)\right] = \sum_{k=0}^{N_{S_2}} \log\left(\frac{k}{N_{S_2} - k}\right)\binom{N_{S_2}}{k} p_{B_2|S_2}^{k}\left(1 - p_{B_2|S_2}\right)^{N_{S_2} - k}, \quad k = 0, 1, \ldots, N_{S_2}. \tag{A8}$$

When calculating these expected values, a suitable correction procedure is applied. If this is not done, the result is undefined. Once the independent expected values have been calculated, they can be combined to give us

APPENDIX (Continued)

E[log $\hat{d}$], the expected value of log $\hat{d}$. It is represented by the mean of the expected log of correct to incorrect responses for S_1 and S_2 trials.

$$\mathrm{E}\left[\log \hat{d}\right] = \frac{\mathrm{E}\left[\log\left(\frac{p_{B_1|S_1}}{p_{B_2|S_1}}\right)\right] + \mathrm{E}\left[\log\left(\frac{p_{B_2|S_2}}{p_{B_1|S_2}}\right)\right]}{2}. \tag{A9}$$

It is also important in the present article to calculate the variance of the sampling distribution of log $\hat{d}$. For this calculation, it is necessary to compute the second raw moment. This requires a simple adjustment of Equations A7 and A8. The resulting equations are

$$\mathrm{E}\left[\log\left(\frac{\hat{p}_{B_1|S_1}}{\hat{p}_{B_2|S_1}}\right)^2\right] = \sum_{k=0}^{N_{S_1}}\left[\log\left(\frac{k}{N_{S_1}-k}\right)\right]^2 \binom{N_{S_1}}{k} p_{B_1|S_1}^{k}\left(1-p_{B_1|S_1}\right)^{N_{S_1}-k}, \quad k = 0, 1, \ldots, N_{S_1} \tag{A10}$$

and

$$\mathrm{E}\left[\log\left(\frac{\hat{p}_{B_2|S_2}}{\hat{p}_{B_1|S_2}}\right)^2\right] = \sum_{k=0}^{N_{S_2}}\left[\log\left(\frac{k}{N_{S_2}-k}\right)\right]^2 \binom{N_{S_2}}{k} p_{B_2|S_2}^{k}\left(1-p_{B_2|S_2}\right)^{N_{S_2}-k}, \quad k = 0, 1, \ldots, N_{S_2}. \tag{A11}$$

Equations A8 and A10, and A9 and A11, can then be combined to compute the variance for each type of trial:

$$\mathrm{Var}\left[\log\left(\frac{\hat{p}_{B_1|S_1}}{\hat{p}_{B_2|S_1}}\right)\right] = \mathrm{E}\left[\log\left(\frac{p_{B_1|S_1}}{p_{B_2|S_1}}\right)^2\right] - \mathrm{E}\left[\log\left(\frac{p_{B_1|S_1}}{p_{B_2|S_1}}\right)\right]^2 \tag{A12}$$

and

$$\mathrm{Var}\left[\log\left(\frac{\hat{p}_{B_2|S_2}}{\hat{p}_{B_1|S_2}}\right)\right] = \mathrm{E}\left[\log\left(\frac{p_{B_2|S_2}}{p_{B_1|S_2}}\right)^2\right] - \mathrm{E}\left[\log\left(\frac{p_{B_2|S_2}}{p_{B_1|S_2}}\right)\right]^2. \tag{A13}$$

To compute Var[log $\hat{d}$], the overall variance of the sampling distribution of log $\hat{d}$, Equations A12 and A13 are simply combined. The resulting equation is

$$\mathrm{Var}\left[\log \hat{d}\right] = \mathrm{Var}\left[\log\left(\frac{p_{B_1|S_1}}{p_{B_2|S_1}}\right)\right] + \mathrm{Var}\left[\log\left(\frac{p_{B_2|S_2}}{p_{B_1|S_2}}\right)\right]. \tag{A14}$$

We used Equations A1 through A14 to calculate the characteristics of the sampling distribution of log $\hat{d}$. To calculate the characteristics of the sampling distribution of $\hat{d}'$, we used the analogous equations described by Miller (1996). Similar principles can be used to calculate characteristics of the sampling distribution of any of the common performance measures used in yes–no tasks. The calculation, however, is somewhat more cumbersome for measures that additively combine response counts over S_1 and S_2 trials, such as percent correct. This is because distributions for S_1 and S_2 trials cannot then be calculated independently.

(Manuscript received November 3, 2003;
revision accepted for publication November 22, 2004.)