Behavior Research Methods, Instruments, & Computers
1995, 27 (1), 46-51

Corrections for extreme proportions and their biasing effects on estimated values of d'

MICHAEL J. HAUTUS
University of Auckland, Auckland, New Zealand

Estimating d' from extreme false-alarm or hit proportions ($p = 0$ or $p = 1$) requires the use of a correction, because the z score of such proportions takes on infinite values. Two commonly used corrections are compared by using Monte-Carlo simulations. The first is the $1/(2N)$ rule for which an extreme proportion is corrected by this factor before d' is calculated. The second is the log-linear rule for which each cell frequency in the contingency table is increased by 0.5 irrespective of the contents of each cell. Results showed that the log-linear rule resulted in less biased estimates of d' that always underestimated population d'. The $1/(2N)$ rule, apart from being more biased, could either over- or underestimate population d'.

The estimation of d' from data obtained using a single-interval detection-theoretic task, such as the yes/no task, is relatively straightforward. If it is assumed that the distributions of the sensory evidence derived from the noise and signal-plus-noise stimuli are both normal and of equal variance, the relation between d', the hit rate (H), and the false-alarm rate (F), is simply

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

The function $Z(x)$ is the *inverse-normal transform* and is equal to the value, z, of the standardized normal variate, Z, so that $p(Z \leq z) = x$. $Z(x)$ is not to be confused with the traditional z score, in which case the two terms on the right-hand side of Equation 1 would be reversed. It may be noted that both of these conventions have been employed in the psychophysical literature (cf. e.g., Egan, 1975, p. 61; Macmillan & Creelman, 1991, p. 9).

The equal-variance assumption has a sound theoretical basis for cases in which the observer must discriminate between two stimuli that are presented separately. This sort of task, termed a *difference discrimination* by Laming (1986), can be contrasted with an *increment detection* task, in which the observer must detect an increment in an otherwise continuous stimulus background.[1] Fortunately, most experiments can be structured to take the form of a difference discrimination so that the computational simplicity of the normal–normal equal-variance model can be employed.

There is, however, a problem with Equation 1. It is difficult to obtain a sensible estimate of d' when either F or H is equal to zero or one. This is because the inverse-normal transform takes on an infinite value for these cases. The traditional treatment for this problem distinguishes between extreme proportions that are due to sampling variability and those that are due to perfect discrimination. Truly perfect discrimination would require an observer to detect the signal correctly on every occasion that it was present ($H = 1$) *and* never to say the signal was present when, in fact, it was not present ($F = 0$). If either $H \neq 1$ or $F \neq 0$, it is unlikely that the observer is able to discriminate between the stimuli perfectly, and it is reasonable to assume that the extreme proportion is due to sampling variability.

Two methods for dealing with extreme proportions due to sampling variability are widely used. A method that has found much use in log-linear analysis requires the addition of 0.5 to each cell in the two-by-two contingency table that defines the performance of the observer (e.g., Fienberg, 1980; Goodman, 1970; Knoke & Burke, 1980). Row and column totals are increased by one. This transformation, which will be termed the *log-linear rule*, makes extreme proportions impossible to obtain, and d' is estimated by entering the corrected values of H and F into Equation 1. Proponents of this method recommend that the contingency table be transformed irrespective of whether extreme proportions are obtained (Snodgrass & Corwin, 1988). The second approach has been referred to as the $1/(2N)$ rule (Macmillan & Kaplan, 1985) and is probably more familiar to psychophysicists. Extreme proportions of zero or one are replaced with values of $1/(2N)$ and $1-1/(2N)$ respectively, where N equals the number of trials upon which the proportion is based. No correction is applied to proportions other than zero or one. The aim of the present paper is to investigate the biasing effects of these two corrections on estimates of d'.

Distinctions Between the Rules

There are two basic distinctions between these two methods of correction. First, the log-linear rule is applied to all entries in the contingency table irrespective of their nature. Therefore, both F and H are *always* cor-

I would like to thank John Irwin for his valuable comments on the draft of this manuscript. Requests for reprints should be sent to M. J. Hautus, Department of Psychology, The University of Auckland, Private Bag 92019, Auckland, New Zealand.

rected proportions. On the other hand, the $1/(2N)$ rule is applied only to cell frequencies of zero or N—that is, frequencies that result in probabilities of zero or one. Second, it is easy to show that the corrected probabilities obtained using the log-linear rule for cell frequencies of zero or N are equal to $1/[2(N+1)]$ and $1-1/[2(N+1)]$, respectively. This suggests that the form of the two rules is very similar; they differ mainly in the range of data to which they are applied.

Monte-Carlo Simulations

A series of Monte-Carlo simulations was conducted to determine the biasing effects of each of the two corrections on estimates of d'. The simulations were of the single-interval, yes/no task under the assumption of the normal–normal, equal-variance model of detection theory. Experiments were based on 20, 50, 100, 200, and 400 trials, and the criteria, f, upon which judgments were based were defined by their false-alarm rates of .05, .25, .50, .75, and .95. The population values of d' ranged from zero to three in steps of 0.5. The two corrections were applied to F and H where appropriate. Distributions of estimated d', each based on 10,000 simulations, were collected for each combination of d', N, and f.

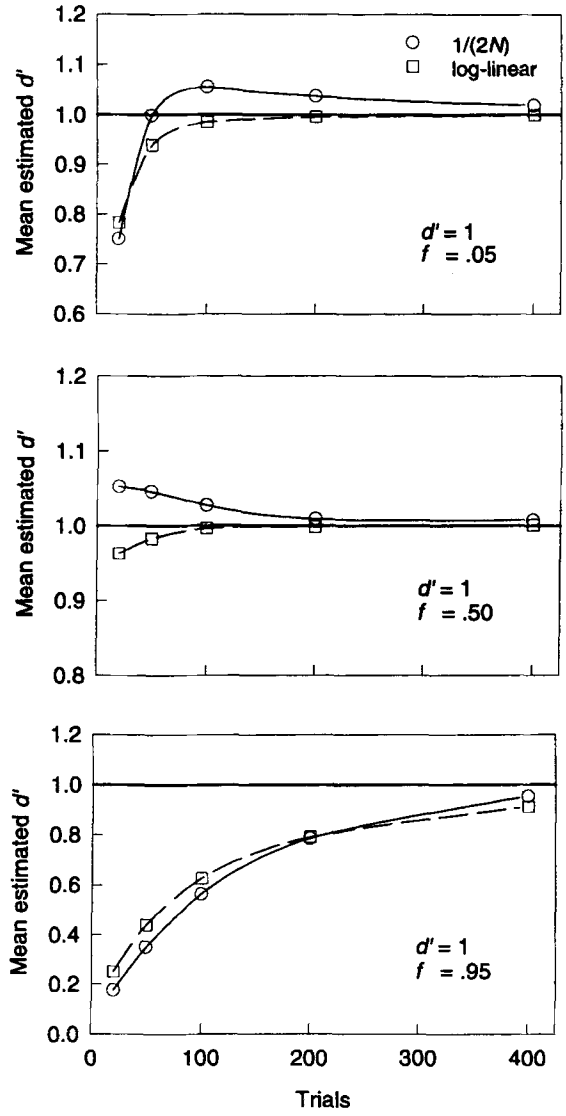

Figure 1. Mean estimated values of d' obtained when population $d' = 1$.

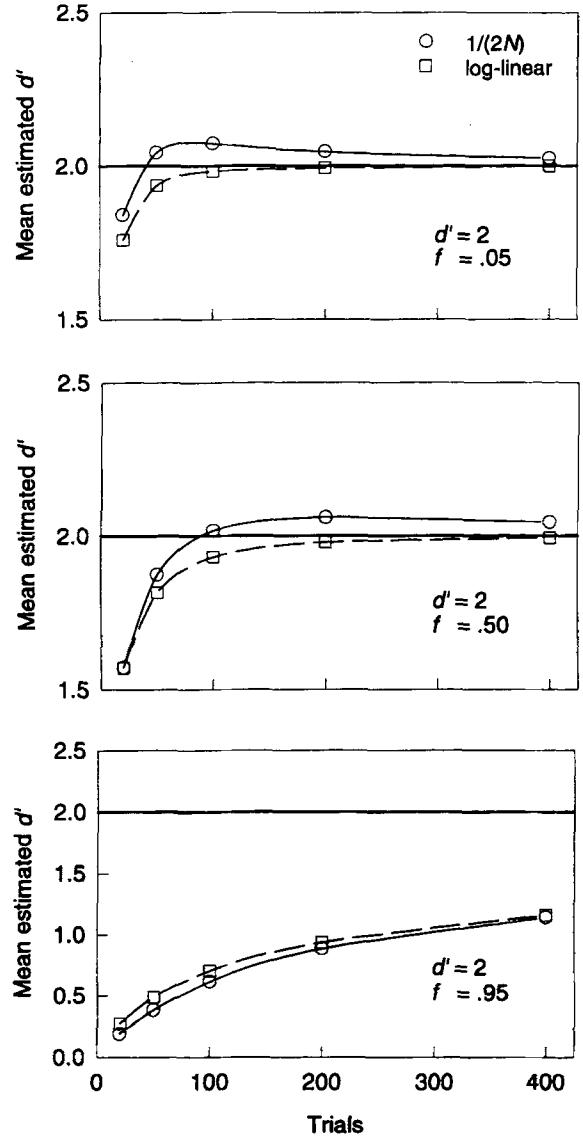

Figure 2. Mean estimated values of d' obtained when population $d' = 2$.

Figures 1–3 illustrate the mean values of estimated d' for each of three criterion values (.05, .50, .95), two corrections, and three values of population d' (1.0, 2.0, and 3.0). The fitted curves are cubic splines through the data points. The first point to note is that estimated d' converged on population d' as the number of trials was increased, but for some conditions a substantial bias remained even if as many as 400 trials were employed. Because the variance of a proportion decreases as $1/N$,

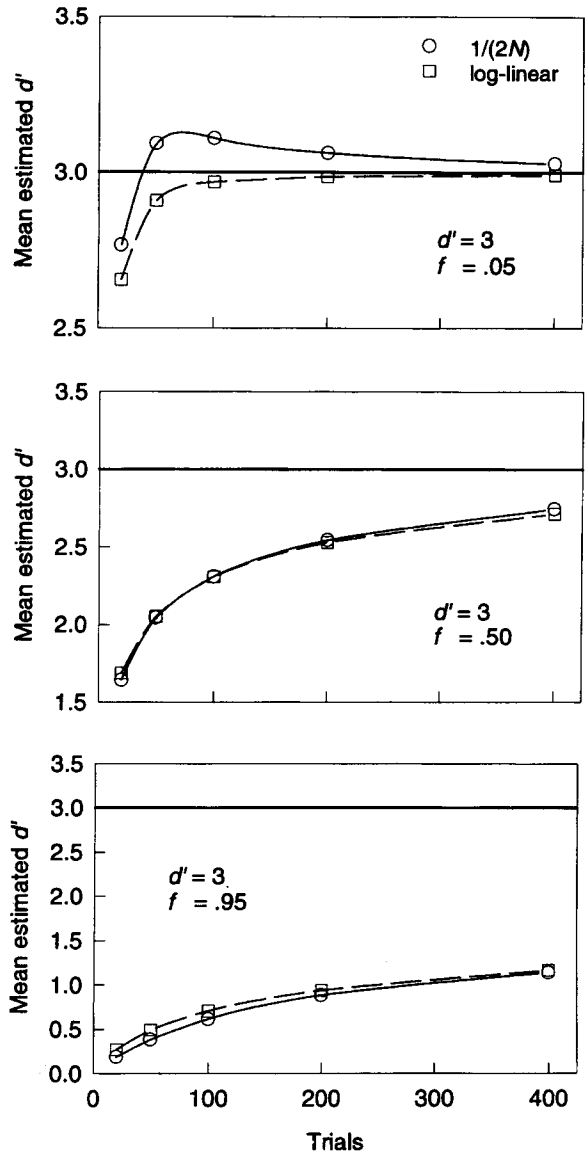

Figure 3. Mean estimated values of d' obtained when population $d' = 3$.

values of H and F based on small N are more likely to require a correction. Estimates of d' based on the $1/(2N)$ rule converged less rapidly than those based on the log-linear rule. Furthermore, estimates based on the log-linear rule always converged from below and therefore always underestimated population d'. This was true also for those cases not illustrated in the figures. A second point is that the criterion adopted by the observer had a major effect on the magnitude of the bias of estimated d'. The bias increased rapidly as the observer became more lax—that is, as F and H increased. For example, when $d' = 1, f = .95$, and $N = 100$, the estimated value of d' was only 0.56 for the $1/(2N)$ rule. It is clear that substantial biases occur irrespective of which correction rule is used.

Gourevitch and Galanter (1967) demonstrated that observed z scores have an approximately normal distribution. This means that estimates of d' are distributed as the difference between two approximately normal distributions (see Equation 1). Such a difference is also approximately normally distributed (see Macmillan and Creelman, 1991, pp. 271–274). To test this hypothesis, a Kolmogorov-Smirnov test of normality was applied to the first 100 values of d' estimated from each simulation. The first 100 values were chosen so that the tests were not too severe, because the distribution of estimated d' is only approximately normal. The majority of distributions were not significantly different from normal ($\alpha = .05$). However, distributions tended to depart from normality as N decreased, as criteria became more extreme, and as population d' increased; these are all situations for which the likelihood of requiring a correction is increased.

To illustrate the distributions of estimated d', histograms based on 50,000 simulations each were generated. Figure 4 shows the distribution of estimated d' when population $d' = 1$, the criterion $f = .95$, and the $1/(2N)$ rule was applied. The number of trials for each simulation was $N = 100$ for the left panels and $N = 400$ for the right panel (these cases can also be found in the bottom panel of Figure 1). The lightly shaded regions correspond to the number of simulations on which $H < 1$—that is, those simulations in which no correction was applied. The heavily shaded regions correspond to simulations in which a correction was applied to H. The solid curves are the best-fitting normal density functions. For the case of $N = 100$, the distribution is dominated by estimates of d' derived from corrected hit rates. These estimates tend to underestimate population d', thereby producing a bias in mean estimated d'. As N increases, the variance of the hit rate gets smaller and fewer corrections are required. Mean estimated d' begins to converge on population d'. This is evident in Figure 4 for the case of $N = 400$.

Figure 5 shows the distribution of estimated d' for the case where population $d' = 3, f = .05$, and $N = 100$ (this case is also illustrated in the top panel of Figure 3). The distributions in the left and right panels were derived by using the $1/(2N)$ and log-linear rules, respectively. Sampled values of $H = 1$ occurred rarely; hence, very few corrections were applied by using the $1/(2N)$ rule. However, when corrections were applied, the resulting estimates of d' were in the upper tail of the distribution, causing the mean of the estimates to overestimate population d' slightly. For the log-linear rule, corrections were applied to all pairs of H and F. This has the effect of skewing the entire distribution slightly to the left. This can be observed visually in Figure 5 or can be determined by comparing the coefficient of skewness for the two distributions. The coefficient for this case was -0.331 for the $1/(2N)$ rule and -0.512 for the log-linear rule. This negative skew weighted the distribution so that population d' was underestimated. The skew reduces as N increases, as does the frequency of extreme propor-

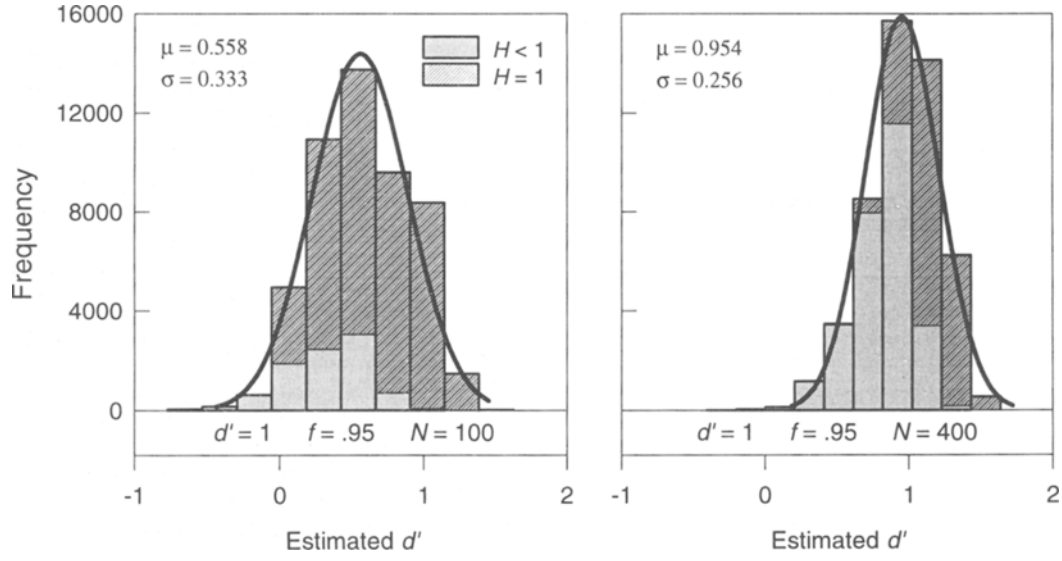

Figure 4. Distributions obtained for 50,000 simulations with population $d' = 1$ and $f = .95$. Each simulation was based on either $N = 100$ (left panel) or $N = 400$ (right panel). The solid curve is the best-fitting normal density function.

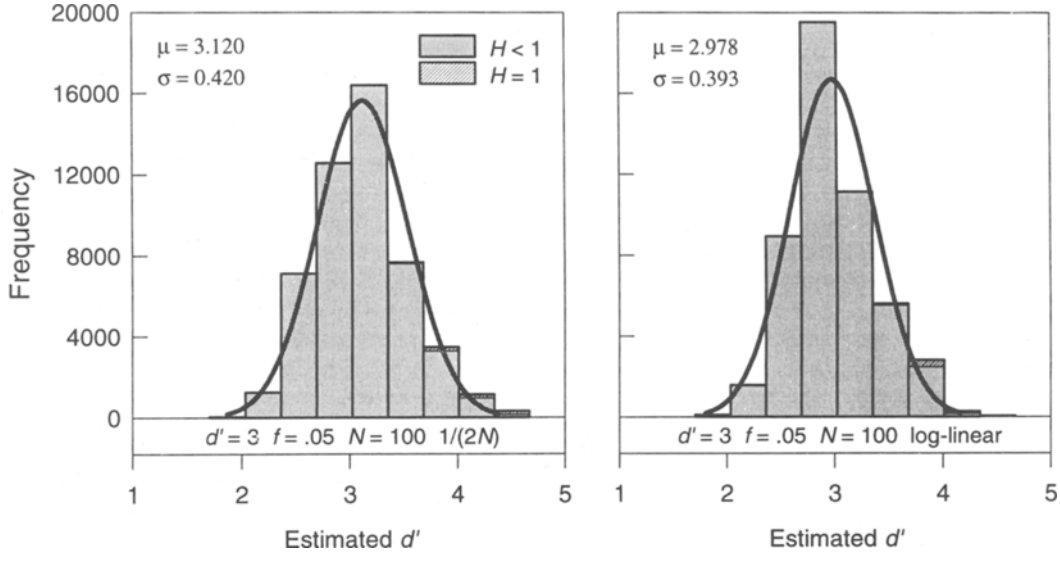

Figure 5. Distributions obtained for 50,000 simulations with population $d' = 3$, $f = .05$, and $N = 100$. The $1/(2N)$ rule (left panel) and the log-linear rule (right panel) were employed on the same simulated sequence.

tions. For this reason, mean estimates of d' based on the log-linear rule tend to converge on population d' from below.

A second set of simulations was conducted to gain information on the effect of observer criterion on the magnitude of the bias of estimates of d'. This time only one criterial value was employed for each value of population d'. The other parameters were kept the same as those in the previous simulations. The criterion was set at $F = 1 - \Phi(d'/2)$, which is the criterion adopted by an unbiased observer. Figure 6 illustrates mean estimated d' for an unbiased observer at population d' values of 1.0, 2.0, and 3.0. The log-linear rule resulted in relatively unbiased d' estimates over the range considered, at least for N greater than 50. The $1/(2N)$ rule yielded d' estimates that converged on population d' much less rapidly. Furthermore, the rate of convergence decreased as d' increased. A comparison of Figures 1–3 with Figure 6 reveals that the magnitude of bias introduced by the corrections is greater for criteria that are more distant from a likelihood ratio of one (that of an unbiased observer). This occurs whether the criterion is strict or lax.

These simulations have focused on the detection-theoretic index of sensitivity, d'; however, the results can be generalized to other indices that can be considered near relatives of d'. Examples of these indices are η (Luce, 1959, 1963), the log-odds ratio, LOR (Goodman, 1970),

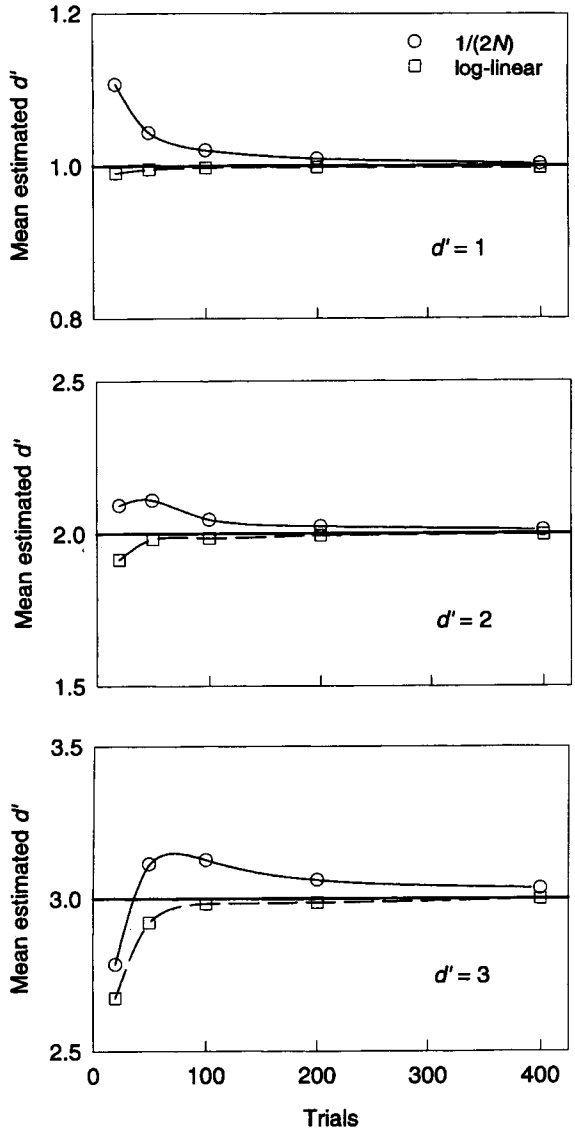

Figure 6. Mean estimated values of d' obtained for an unbiased observer.

and log(d) (Davison & Tustin, 1978). This was demonstrated by a simulation conducted to determine the mean estimate of log(d) for population values of log(d) equal to 0.351, 0.725, and 1.145. These values of log(d) are equivalent to d's of 1, 2, and 3. The results of the simulation were similar to those reported for d'.

Reduction of Bias Magnitude

One approach to eliminating the biasing effects of corrections for extreme proportions is to employ a methodology that can reduce such values. The rating method of detection theory is an ideal candidate in this respect, although it requires a small amount of additional effort on the part of both the experimenter and the subject. For this method, the subject provides a series of ratings rather than a simple binary response (see Green & Swets, 1966). During initial training, the subjects are guided to adjust their criteria so that all rating categories are used (although not used equally), thereby reducing the number of categories to which no responses are made. This process is possible for all but highly (almost perfectly) discriminable stimuli. Hence, a number of points are obtained in ROC (receiver operating characteristic) space. It is possible that extreme categories may still contain zeros, but most unlikely that most categories will have them. Once the data are collected, a computer program can be used to estimate d'. An example of a suitable method was published by Dorfman and Alf (1969).

The problem of what to do with any rating category for which the cumulative cell frequency is either zero or N still remains, however. One option is to use either the $1/(2N)$ rule or the log-linear rule to determine a corrected proportion for such cells. Dorfman and Alf's (1969) program, for example, used the $1/(2N)$ rule. However, the use of a correction introduces bias into the estimated value of d'. Furthermore, it is known that small cell frequencies tend to result in goodness-of-fit statistics (such as χ^2 or G^2) that reject the fitted model more often than they should (Metz, 1989; Wickens, 1992). This suggests that corrections should also be avoided, when possible, in any situation in which the goodness-of-fit of a mathematical model is of interest. An approach that reduces the use of corrections, and hence the admission of bias, is to collapse two or more adjacent cells so that extreme proportions do not occur (McNicol, 1972). Thus, estimated d' is based only on empirically determined proportions, not estimated proportions. In practice, there are few cases that cannot be dealt with in this manner.

Conclusions

Corrections are required in order to calculate d' from proportions of zero or one. The log-linear rule usually results in less biased estimates of d' than does the $1/(2N)$ rule. Furthermore, the log-linear rule leads to estimates of d' that converge asymptotically on population d' always from below. The $1/(2N)$ rule does not provide such consistent performance. Finally, the log-linear rule treats all data equally. No distinction is made between acceptable data ($0 < p < 1$) and unacceptable data ($p = 0$ or $p = 1$). On the other hand, the $1/(2N)$ rule targets problematic data only.

A methodology that may reduce the need to use these corrections is the rating method of detection theory. This method yields a number of points in ROC space and, with correctly trained subjects, it is rare to obtain all points with either F or H equaling zero or one. In most cases, an estimate of d' can be obtained without using corrections by collapsing categories.

REFERENCES

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.

Dorfman, D. D., & Alf, E., Jr. (1969). Maximum likelihood estima-

tion of parameters of signal detection theory and determination of confidence intervals—rating-method data. *Journal of Mathematical Psychology*, **6**, 487-496.

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

FIENBERG, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: MIT Press.

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

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

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

KNOKE, D., & BURKE, P. J. (1980). *Log-linear models* (Sage University Paper Series on Quantitative Application in the Social Sciences, 07-020). Beverly Hills: Sage Publications.

LAMING, D. (1986). *Sensory analysis*. New York: Academic Press.

LUCE, R. D. (1959). *Individual choice behavior*. New York: Wiley.

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

MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. New York: 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.

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

METZ, C. E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, **24**, 234-245.

SNODGRASS, J. G., & CORWIN, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.

WICKENS, T. D. (1992). Maximum-likelihood estimation of a multivariate Gaussian rating model with excluded data. *Journal of Mathematical Psychology*, **36**, 213-234.

NOTE

1. Laming (1986) has demonstrated that, for increment detections, the random variables that represent the sensory evidence are noncentral chi-square.

(Manuscript received August 4, 1993;
revision accepted for publication January 21, 1994.)