Journal of Experimental Psychology:
Learning, Memory, and Cognition
2009, Vol. 35, No. 2, 509–527

0278-7393/09/$12.00 DOI: 10.1037/a0014876

Sources of Bias in the Goodman–Kruskal Gamma Coefficient Measure of Association: Implications for Studies of Metacognitive Processes

Michael E. J. Masson
University of Victoria

Caren M. Rotello
University of Massachusetts Amherst

In many cognitive, metacognitive, and perceptual tasks, measurement of performance or prediction accuracy may be influenced by response bias. Signal detection theory provides a means of assessing discrimination accuracy independent of such bias, but its application crucially depends on distributional assumptions. The Goodman–Kruskal gamma coefficient, *G,* has been proposed as an alternative means of measuring accuracy that is free of distributional assumptions. This measure is widely used with tasks that assess metamemory or metacognition performance. The authors demonstrate that the empirically determined value of *G* systematically deviates from its actual value under realistic conditions. A distribution-specific variant of *G*, called G_c, is introduced to show why this bias arises. The findings imply that caution is needed when using *G* as a measure of accuracy, and alternative measures are recommended.

Keywords: discrimination accuracy, gamma coefficient, metamemory measurement, signal detection theory

> Our belief is that each scientific area that has use for measures of association should, after appropriate argument and trial, settle down on those measures most useful for its needs. (Goodman & Kruskal, 1954, p. 763)

In a wide range of cognitive and perceptual tasks, such as recognition memory, metacognitive or metamemory judgments, and perceptual discrimination, a researcher's objective is to assess accuracy of discrimination in a manner that is independent of response bias. For example, in a judgment of learning (JOL) task, subjects are asked to predict whether they will be able to remember each studied item on a subsequent memory test. Subjects vary in how willing they are to make a positive prediction (response bias), so some may nominate many words as likely to be remembered, whereas others may make that claim for very few. The question of interest, however, is how accurately subjects can predict future performance on items, regardless of their varying tendencies to make positive responses.

In this article, we begin with a brief characterization of the distinction between discrimination accuracy and response bias using signal detection theory. We then review the theoretical underpinnings of a widely used measure of metamemory and metacognition performance, the Goodman–Kruskal gamma coefficient, *G* (Goodman & Kruskal, 1954). We demonstrate that empirically determined values of *G* systematically deviate from the true value that *G* is intended to estimate and that *G* varies with response bias even in the absence of any underlying change in true discrimination accuracy. We show why *G* behaves this way and recommend an alternative measure of accuracy, grounded in signal detection theory, which can be applied without requiring any change in the procedures typically used to collect metamemory and metacognition data. Finally, we demonstrate through examples—one based on hypothetical data and others based on actual data from the memory and metamemory literature—that misleading conclusions about performance accuracy and associated cognitive mechanisms may be reached when *G* is used and that these problems can be avoided by the use of signal detection analysis.

Measuring Discrimination Accuracy

A classic measurement tool for independently assessing discrimination accuracy and response bias is based on signal detection theory. In the most straightforward application of this technique, it is assumed that items or stimuli vary on a single, continuous dimension (e.g., evidence strength) and that this variation is normally distributed. Moreover, when comparing two classes of items (e.g., studied and nonstudied words), it is further assumed that these distributions have equal variance. Classification decisions are generated by establishing a response criterion such that any item whose value on the dimension of interest exceeds the criterion is classified into one category (e.g., a studied word) and items falling below the criterion are assigned to another category (e.g., a nonstudied word). The extent to which the two strength distributions overlap determines the degree of difficulty in making accurate responses: Substantial overlap will cause many items to be misclassified, regardless of where the criterion is placed; minimal overlap will permit (but not ensure) highly accu-

Michael E. J. Masson, Department of Psychology, University of Victoria, Victoria, British Columbia, Canada; Caren M. Rotello, Department of Psychology, University of Massachusetts Amherst.

Michael E. J. Masson was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (7910-03), and Caren M. Rotello was supported by a grant from the National Institutes of Health (MH60274). We thank Jason Arndt, Aaron Benjamin, and Lael Schooler for their comments on this work.

Correspondence concerning this article should be addressed to Michael E. J. Masson, Department of Psychology, University of Victoria, P.O. Box 3050 STN CSC, Victoria, British Columbia V8W 3P5, Canada. E-mail: mmasson@uvic.ca

rate classifications. Accuracy of responding can be quantified as the distance (d') between the means of the two normal, equally variable, distributions; it is measured in units of the common standard deviation. Formally, accuracy is defined as:

$$d' = z(H)–z(F), \quad (1)$$

where z gives the z score associated with a response proportion, H is the proportion of hits (e.g., classifying a studied word as studied), and F is the proportion of false alarms (e.g., classifying a nonstudied item as studied). Note that the proportions are treated as lower tail regions (area below the z score), so, for example, $z(H)$ becomes positive and grows larger as H increases. The assumption of equal-variance distributions can be relaxed by adopting a more general accuracy measure:

$$d_a = \sqrt{\frac{2}{1 + s^2}}\,(z(H) - sz(F)) \quad (2)$$

where s is the standard deviation of the noise distribution (e.g., nonstudied items), and the standard deviation of the signal distribution (e.g., studied items) is fixed at 1 without loss of generality. (Methods for estimating s are described in Appendix A.) Thus, d_a measures the distance between the means of the two distributions in units that are a compromise between the two standard deviations; in the case of equal variance, d_a equals d'. Notably, the d_a measure will yield a consistent value across variations in response criterion for both equal- and unequal-variance Gaussian evidence distributions; d' is independent of response bias only in the equal-variance case (Macmillan & Creelman, 2005).

Because application of signal detection analysis requires a set of assumptions pertaining to the shape and relative variance of distributions of relevant psychological variables (e.g., memory strength, feeling of knowing [FOK]), it would be desirable to have available an alternative, nonparametric measure that does not rely on such assumptions. One such measure of discrimination accuracy was advocated by Nelson (1984, 1986a, 1986b), who proposed that the Goodman and Kruskal (1954) gamma coefficient (G) could be used as a measure of discrimination accuracy without making the distributional assumptions associated with signal detection theory. Building on the Goodman and Kruskal probabilistic interpretation of G, Nelson characterized discrimination accuracy in terms of a conditional probability. The critical probability, V, specifies how likely it is that an observer will claim that Stimulus J is greater than Stimulus K on an attribute of interest (e.g., memory strength), given that Stimulus J actually is greater than Stimulus K:

$$V = p(J_c > K_c \mid J_a > K_a), \quad (3)$$

where Subscripts c and a refer to the observer's claim and the actual state of affairs, respectively.

Nelson (1984) showed that V is linearly related to the Goodman–Kruskal G statistic in the following way:

$$G = 2V–1. \quad (4)$$

He also proposed that this probabilistic interpretation "makes no assumption about underlying distributions of the attribute, not even that there are underlying distributions versus discrete states" (Nelson, 1986b, p. 130). Moreover, Smith (1995) advocated Nelson's proposal as an alternative that does not depend on "assumptions about underlying ROC [receiver operating characteristic] curves or distributions" (p. 88). It appears that G offers an appealing alternative to d' or d_a as a measure of discrimination accuracy, given its distribution-free probability interpretation.

Although G has been used only occasionally in studies of memory (e.g., Criss & Shiffrin, 2004; Nobel & Shiffrin, 2001) or perception (e.g., Masson & Hicks, 1999), it has been used frequently in studies of metacognitive processes such as JOLs (e.g., Kao, Davis, & Gabrieli, 2005; Souchay, Isingrini, Clarys, Taconnat, & Eustache, 2004). We conducted a survey of articles published between January 2000 and July 2008 in four of the leading journals in the field of memory and cognition (*Journal of Experimental Psychology: General*; *Journal of Experimental Psychology: Learning, Memory, and Cognition*; *Journal of Memory and Language*; and *Memory & Cognition*) to estimate the relative frequency with which G is used to assess metacognition accuracy. Using the keywords *metacog*(*nition*), *metamem*(*ory*), *metacomp*(*rehension*), *feeling of knowing*, *FOK*, *judgments of learning*, and *JOL*, we identified 64 articles that reported empirical work on metacognition. Of these, 31 (nearly half) used G as a primary measure of metacognitive accuracy.

Although computation of G requires no distributional assumptions, we present simulation data showing that estimates of discrimination accuracy based on G are systematically affected by response bias and that corrections for that dependency necessarily entail distributional assumptions. Before presenting the simulations, however, we describe the ways in which V and G may be calculated, and point out some issues that arise in doing so. In particular, we consider treatment of pairs of observations that constitute ties (i.e., $J_c = K_c$ or $J_a = K_a$).

Calculating the Goodman–Kruskal G Coefficient

2×2 Designs

In research designs involving a 2 (stimulus class) $\times$ 2 (response category) design, V can be computed by enumerating the stimulus pairs that are concordant with the conditional probability shown in Equation 3 and those that are discordant (Nelson, 1984). Table 1 shows the data from a hypothetical recognition experiment in which a subject classified 10 studied and 10 nonstudied items as old or new. A concordant pair of items consists of one studied item and one nonstudied item ($J_a > K_a$), where the subject classifies the

Table 1
Data From a Subject in a Hypothetical Recognition Memory Experiment

	Stimulus	
Response	Studied	Nonstudied
Old	7 (a) Hit	4 (b) False alarm
New	3 (c) Miss	6 (d) Correct rejection

Note. The letter in each cell is used to establish the correspondence between cells of the table and the formulas for computing V (Equation 5) and G (Equation 6).

former as old and the latter as new ($J_c > K_c$). Thus, concordant pairs can be constructed by pairing a trial on which a studied item was classified as old (a hit) with a trial on which a nonstudied item was classified as new (a correct rejection). From Table 1, there are 7 hits and 6 correct rejections, so there are $7 \times 6 = 42$ concordant pairings of trials or items in this data set. A discordant pair consists of a nonstudied stimulus that is incorrectly classified as old (false alarm) and a studied stimulus mistakenly classified as new (miss). There are 4 and 3 trials of these types, respectively, producing $4 \times 3 = 12$ discordant pairs. Using the letter labels shown in the cells of Table 1, Nelson defined the computation of V for a 2×2 classification task as the proportion of concordant pairs relative to the total number of concordant and discordant pairs:

$$V = \frac{ad}{ad + bc}. \quad (5)$$

For this example, then, $V = 42/(42 + 12) = .78$. V approaches 1.0 as the hit rate approaches 1.0 and the false-alarm rate approaches .0 (perfect discrimination); as the hit and false-alarm rates become more similar to each other, V approaches .5 (guessing). Equation 4 shows that G approaches 1 as V approaches 1 and that G approaches 0 as V approaches .5. The Goodman–Kruskal G coefficient for the data in Table 1 is $G = (2 \times .78) - 1 = .56$. Combining Equations 4 and 5 and simplifying, G can be computed as

$$G = \frac{ad - bc}{ad + bc}. \quad (6)$$

Nelson (1986a) showed that Equation 6 could be rewritten to define G as a function of the hit rate (H) and the false-alarm rate (F):

$$G = \frac{H - F}{H + F - (2HF)}. \quad (7)$$

$2 \times m$ Designs

For the 2×2 design, G is equivalent to Yule's (1912) Q, although the G coefficient can also be computed for the $2 \times m$ case, in which there are m response options that reflect a subject's confidence rating for each classification response. For example, in a recognition test, a 6-point rating scale could be used, where 1 indicates a high-confidence new response and 6 indicates a high-confidence old response. Indeed, experiments on JOLs and metamemory typically use a confidence scale of this form, and studies of this type very commonly use G to measure accuracy of learning and memory predictions (e.g., Bornstein & Zickafoose, 1999; Dunlosky & Nelson, 1994; Koriat, Ma'ayan, & Nussinson, 2006; Nelson & Dunlosky, 1991).

Computation of G based on response ratings requires a modification of the procedure for counting the number of concordant and discordant pairs—the values that are used in Equation 6 to generate G. As explained by Gonzalez and Nelson (1996), concordant pairs are defined as non-tied pairs (e.g., one studied, one unstudied), in which the ordering on the response rating is consistent with the stimulus categories. One subset of concordant pairs would be any studied item given a "sure old" rating that is paired with any nonstudied item given a rating lower than "sure old". Another subset would be formed from pairing any studied item given the second highest confidence rating paired with any nonstudied item given a confidence rating below the second highest rating. One subset of discordant pairs would consist of nonstudied items given the "sure old" rating paired with studied items given any rating lower than "sure old". This definition of concordant and discordant pairs is perfectly consistent with Equation 3. Nelson (1986a) provided code for a simple computer program to compute G from ratings data.

The Problem of Tied Observations

It is important to note that Equation 6 (and its formal equivalent, Equation 7) does not take account of the entire set of paired observations available in a typical discrimination paradigm, in the sense that the computation excludes any pair of observations formed from stimuli belonging to the same class (e.g., both are studied words in a memory task, $J_a = K_a$) or that involve the same response by the subject (e.g., both are classified as old, $J_c = K_c$). The difficulty in evaluating the status of pairs from the same stimulus class is that there is no basis for determining which (if either) has the higher value on the stimulus dimension (e.g., memory strength). Therefore, the validity of the subject's $J_c > K_c$ classification of the stimuli cannot be ascertained. Similarly, by placing two items in the same response category, the subject signals his or her inability to distinguish them on the dimension of interest.

The approach taken by Goodman and Kruskal (1954) in defining G was to exclude these tied observations from the computation; Nelson (1984, 1986a) followed suit. Spellman, Bloomfield, and Bjork (2008) pointed out that the exclusion of tied pairs reduces the number of observations on which G is based and thereby reduces the stability of this measure. Others have argued for correcting the computation of G for the presence of ties (Kim, 1971; Somers, 1962; Wilson, 1974). The proposed corrections all involve including a subset of tied pairs in the denominator of the G formula, thereby reducing the magnitude of G (see Freeman, 1986, for a review).

Gonzalez and Nelson (1996) argued that the decision of whether to include tied pairs in the computation of G should be based on whether ties are ambiguous. Ambiguity arises when the procedure potentially forces subjects to categorize stimuli that differ on the dimension of interest into a single category. For example, in an old/new recognition memory test, studied items receiving positive responses may vary on an underlying dimension such as memory strength or familiarity, but the subject has no opportunity to express his or her sensitivity to this variation. Using confidence ratings allows the subject to make finer grained distinctions between items classified as old (or new), but even then it is likely that items varying in familiarity will nonetheless be assigned the same confidence rating. This situation is inevitable any time there are more items than response classes. In these cases, ties are "forced by the procedure" (Gonzalez & Nelson, 1996, p. 162), creating an ambiguity in their interpretation: Is the tie intended by the subject or caused by an insufficiently fine grain size in response classes? In such cases (i.e., essentially all real situations), Gonzalez and Nelson recommended ignoring ties and using the standard computation of G, as shown in Equation 6.

Comparison of Empirically Estimated *G* to Actual γ

To illustrate the implications of the information loss due to ignoring tied observations, we generated sets of hypothetical data that simulated a 2 × 2 classification task, such as an old/new recognition memory task, or a 2 × 6 task, such as a recognition task with confidence ratings. In the first simulation, we sampled observations from Gaussian distributions with either equal or unequal variances. The second simulation was based on rectangular distributions. We recognize that there is nothing in the probabilistic interpretation of *G* that requires data generation to be conducted in this way. This approach was adopted because it provided a convenient and principled means of generating observed data and at the same time allowed us to compute an actual population value of discrimination accuracy, γ, against which we could compare values of *G* generated by applying Equation 6 to simulated data.

An earlier study by Benjamin and Diaz (2008) reported simulations of *G* under a range of actual γ values. They found that the function relating *G* to the true γ value was nonlinear such that as γ increased, the value of *G* followed a negatively accelerated function that reached asymptote near 1.0. The shape of this function is a necessary consequence of *G* having upper and lower bounds of 1.0 and −1.0, and it implies that *G* is not capable of correctly representing the magnitude of accuracy differences when accuracy nears asymptote. To address this problem, Benjamin and Diaz introduced a modified measure of accuracy, which they called G^*. This measure is computed as the logit of *V*, where *V* is defined as in Equation 3 above. Thus, G^* is related to *G* in the following way:

$$G^* = log\left(\frac{G + 1}{1 - G}\right). \quad (8)$$

Modifying the accuracy measure in this way defines its upper and lower bounds as ±∞, and Benjamin and Diaz found that G^* varied nearly linearly with actual γ. The promising results reported with this modified measure of *G* led us to include it in the simulations we report here.

When evaluating the performance of *G*, which is a sample-based estimate of the actual, population value of γ, it is important to be clear about exactly how that population value is derived. In the classic statistics literature on the evaluation of *G* as an estimator of γ, the definition of the population value arguably is itself inaccurate because tied observations are ignored. In this literature, population γ is defined by the equivalent of Equation 6 applied to a design matrix, such as Table 1, that represents a population (e.g., see Gans & Robertson, 1981a, pp. 942–943). Simulated samples are then created by randomly selecting cases on the basis of the probabilities associated with each cell of the design matrix. For instance, in Table 1 the probability of selecting a case that produces a hit (responding "old" to a studied item) is 7/20 = .35. Our approach, in contrast, is to define γ using specific assumptions about the continuous distributions that characterize signal and noise targets, which allows us to avoid the occurrence of ties when determining the value of γ for a particular pair of signal and noise distributions. The consequence of ignoring ties thus applies only when computing *G* and not when defining the population γ against which it is assessed. This method allows us to investigate bias in *G* as an estimator of γ in a novel way relative to the traditional statistical literature. Because our definition of the population value of γ is sensitive to the nature of the underlying evidence distributions, we refer to this population value as γ_d to distinguish it from the classical definition of γ.

Simulation 1: Gaussian Distributions

In Simulation 1, we investigated the behavior of *G* for the case where evidence distributions are Gaussian in form. Analyses of recognition memory (e.g., Wixted & Stretch, 2004) and metamemory data (e.g., Benjamin & Diaz, 2008) suggest that Gaussian distributions are highly plausible, although signal and noise distributions are not likely to have equal variance. Therefore, our simulations examine both the equal variance case and the case where the signal distribution has greater variance.

Method. Hypothetical recognition memory data were simulated by randomly sampling from two real-valued normal distributions. We adopt a recognition memory context for illustrative purposes only—our analysis applies equally well to any task involving the type of 2 × 2 classification scheme shown in Table 1 or a 2 × *m* scheme (e.g., perceptual discrimination, JOL, FOK) in which there are *m* response categories reflecting, for example, varying levels of confidence. In this context, each distribution represented variation among items with respect to memory strength (or any variable that might constitute evidence for prior occurrence). The two distributions differed in their mean value; the distribution with the larger mean was the source of studied items, and the distribution with the smaller mean produced the nonstudied items. To simulate responses to these items for the 2 × 2 case, we established a criterion on the memory strength dimension such that any item with a strength value falling above the criterion was classified as old, and all other items were classified as new. We used a range of criterion values and computed *G* at each one to assess possible changes in *G* as response bias varied. The same studied and nonstudied item values were used to simulate a recognition test with a 6-point rating scale (where 1 indicated a high-confidence new response and 6 indicated a high-confidence old response), representing a 2 × 6 case. This case also fits experiments on JOLs and metamemory that typically use a confidence scale of this form and report *G* as a measure of accuracy in subjects' estimates of their learning or their memory predictions.

In separate runs of this simulation, we examined the behavior of *G* at three different levels of true discrimination accuracy (actual distance between distribution means). In addition, we examined the impact on *G* of unequal variance in the two distributions. Different versions of the simulation were run in which the ratio of standard deviation in the nonstudied distribution to standard deviation in the studied distribution was 1.0 (equal variance) or 0.6 (wider studied distribution). A standard deviation ratio of 0.6 is similar to what is seen in recognition memory experiments (Benjamin, 2005; Glanzer, Kim, Hilford, & Adams, 1999), where the variance often is greater in the signal (studied item) distribution. Thus, the simulation was run six times, representing each combination of three levels of discrimination accuracy and two standard deviation ratios. Figure 1 presents a summary of the flowchart for the simulations that we report.

For each run, we computed the "actual" population value of γ_d by randomly sampling many pairs of items, with each pair consisting of one item drawn from the studied distribution and another

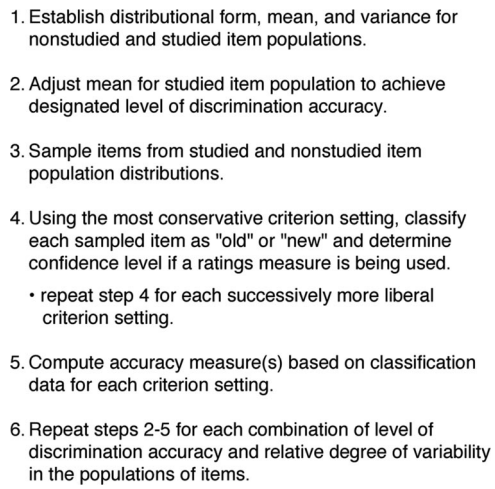

Figure 1. Flowchart showing the steps used in generating the simulation results.

drawn from the nonstudied distribution.[1] In keeping with the probabilistic interpretation of *G* described by Nelson (1984), we computed *V* as the proportion of the sampled pairs for which the studied item had the larger strength. Note that ties will almost never occur in this simulated situation because every pair consists of a studied and a nonstudied item, each consisting of a real-valued number with high precision. Therefore, one of the numbers in a pair is nearly certain to have a greater value than the other. The actual value of γ was then computed as $2V - 1$. In each run, we adjusted the distance between distribution means to produce a particular actual value for γ_d. Details of the evidence distributions, response criteria, and the particular accuracy measures that were computed are shown in Table 2. For the old/new task, each simulated item was compared, in turn, with a decision criterion and then classified into a 2×2 table like the one in Table 1. "Old" responses were made when an item's strength was above the criterion, and "new" responses were made otherwise. To simulate the old/new task with confidence ratings, we constructed a set of five response criteria, separated by equal-sized steps and centered at the criterion used for simulation of the old/new response task. For example, for a false-alarm rate of .05, the original response criterion was a *z* value of 1.645 (cutting off the upper .05 of the standard normal distribution of new items). Two additional criteria were established above and below this point by moving up or down in increments of 0.4 standard deviations. Thus, the full set of five criteria for an overall false-alarm rate of .05 was 0.845, l.245, 1.645, 2.045, and 2.445, corresponding to the boundaries between confidence-rating categories ranging from "sure new" to "sure old". To compute a ratings-based *G*, we classified the simulated items using these sets of response criteria and applied the algorithm described by Nelson (1986a).

In addition to computing *G* and ratings *G*, we also computed G^* and ratings G^* as well as three variants of *G* that represent different methods for correcting for ties. These methods involve including pairs of trials representing one or another type of tie in the denominator of Equation 6. Specifically, Kim (1971) proposed a correction, $d_{y.x}$, in which pairs of observations tied on the predictor variable (e.g., stimulus class, such as studied vs. nonstudied item type, $J_a = K_a$) but differentiated on the criterion variable (e.g., subject responses such as "old" vs. "new") are added to the denominator of Equation 6. For example, working with the cell labels in Table 1, each item in cell *a* could be combined with each item in cell *c* to yield $7 \times 3 = 21$ pairs that are in the same stimulus class (studied items) but distinct with respect to the subject's response (one is classified as old and the other as new). Similarly, another set of pairs tied with respect to stimulus class but differing in the subject's response can be formed from cells *b* and *d*. Thus, the denominator for Equation 6 would become $ad + bc + ac + bd$. Analogously, Somers (1962) proposed a measure, $d_{y.x}$, in which pairs tied on the criterion variable, but not on the predictor variable, are included in the denominator. This correction modifies the denominator of Equation 6 to be $ad + bc + ab + cd$. Finally, a correction advocated by Wilson (1974), *e*, includes both sets of pairs from the Kim and Somers corrections in the denominator, that is, all pairs that are not tied on both the predictor and criterion variables.

Results. The results of Simulation 1 for the equal-variance case are presented in Figure 2, which displays, in the upper panel, the values for *G* calculated as in a 2×2 design, *G* based on confidence ratings, and three versions of *G* corrected for ties. These observed values can be compared with the actual value of γ_d, shown as a horizontal line in each panel. The lower panel shows the obtained values for G^* computed in the 2×2 case and with confidence ratings. The associated population value for G^*, shown as a horizontal line in each of the lower panels, was computed from the actual value of γ_d.

The observed values of *G* (uncorrected for ties) and G^* consistently overestimated the population value, as determined by γ_d, and, critically, the amount of this overestimation varied across false-alarm rates. For example, at $\gamma_d = .4$, the observed *G* exceeded actual γ_d by an amount that varied from .22 when false-alarm rate = .05 to .13 when false-alarm rate = .30. This variation in observed *G* is problematic because it means that even when actual γ_d is constant, conditions that lead to different response biases, and thus different false-alarm rates, may yield spurious differences in *G* as a measure of accuracy. The same is clearly true for G^*. Thus, it is not so much the fact that observed *G* and G^* deviate from the population values determined by γ_d but the fact that this deviation varies across false-alarm rates that is problematic. The correction proposed by Kim (1971) seems to vary somewhat less than *G* across false-alarm rates, with good stability when $\gamma_d = .6$.

The variation in all versions of observed gamma values across changes in false-alarm rate was even more pronounced when the underlying distributions differed in variability, as the results in Figure 3 show. Under these conditions, all of the measures showed a marked decrease in magnitude as false-alarm rate increased. For example, the range in observed *G* across false-alarm rates of .05 to .50 is over .36 for an actual γ_d of .4. Even Kim's (1971) correction fares poorly under these circumstances and is worse than *G* when

[1] We realize that this method of computing the population value of γ_d may not provide the precisely correct value for γ_d that could be found with an analytic solution, but the simulation method is adequate for our purposes.

Table 2
Details of Simulations

Attribute	Simulation 1	Simulation 2	Simulation 3	Simulation 4
Distribution shape and parameters	Gaussian, $\mu_n = 0$, $\sigma_n = 1$, μ_s varies, $\sigma_s = 1$ or 1.67, $\sigma_n/\sigma_s = 1.0$ or 0.6	Rectangular, $\mu_n = 0.5$, $\sigma_n = 0.29$, μ_s varies, $\sigma_s = 0.29$ or 0.48, $\sigma_n/\sigma_s = 1.0$ or 0.6	Gaussian, $\mu_n = 0$, $\sigma_n = 1$, μ_s varies, $\sigma_s = 1$ or 1.67, $\sigma_n/\sigma_s = 1.0$ or 0.6	Gaussian, $\mu_f = 0$, $\sigma_f = 1$, $\mu_r = 1.02$, $\sigma_r = 1.67$
γ_d	.4, .6, or .8	.4, .6, or .8	.4, .6, or .8	.4
Number of items sampled	200,000 items from each distribution	200,000 items from each distribution	3 sets of 50,000 simulated subjects; each set with 16, 64, or 256 old and new items	1,000 simulations of two groups of 20 subjects with 64 remembered and 64 forgotten items each
Response task(s)	Old/new, old/new + confidence ratings	Old/new	Old/new, old/new + confidence ratings	Remember/forget, remember/forget + confidence ratings
False-alarm rates (determine response criteria)	.05, .10, .30, .50, .70, .90, .95	.05, .10, .30, .50, .70, .90, .95	.05, .10, .30, .50	.05, .30
Accuracy measures	*G*, ratings *G*, G^*, ratings G^*, *G* with Kim (1971), Somers (1962), and Wilson (1974) corrections	*G*, G_c	*G*, log-linear *G*, ratings *G*, G^*, log-linear G^*, ratings G^*	*G*, ratings *G*, *d'*, d_a

Note. n = nonstudied, s = studied; for Simulation 4, f = forgotten, r = remembered. γ_d = population value of gamma specific to the underlying evidence distributions; d' = signal detection measure of accuracy assuming equal-variance evidence distributions; *G* = the standard Goodman–Kruskal gamma coefficient; G^* = a modified measure of gamma introduced by Benjamin and Diaz (2008); G_c = corrected gamma coefficient assuming rectangular evidence distributions; d_a = signal detection measure of accuracy allowing for unequal-variance evidence distributions.

$\gamma_d = .8$. These variations in observed gamma values across false-alarm rate reflect a dramatic interaction with response bias.

Computing *G* and G^* from confidence ratings provides for more fine-grained distinctions among items and generally reduces the number of observed ties. Indeed, Figures 2 and 3 reveal that *G* and G^* based on ratings produce smaller overestimates of actual γ_d than when computed from a binary classification task. Spellman et al. (2008) pointed out that, in general, a pair of items close to one another on the relevant stimulus dimension (e.g., similar JOLs) is more likely than a pair that is further apart on the stimulus dimension (e.g., dissimilar JOLs) to be a discordant pair when memory judgments are obtained (i.e., the item given the higher JOL rating is not remembered, whereas the item given a lower rating is remembered). With a binary classification task, however, pairs close together on the stimulus dimension are likely to be classified together, resulting in a tied pair, which is excluded from calculation of *G*. Thus, fewer discordant pairs are generally obtained when computing *G* on the basis of a binary classification task rather than a ratings task. Moreover, ratings-based *G* varies to a smaller extent as the false-alarm rate changes. Despite this improvement, however, even ratings-based *G* systematically deviated from actual γ_d and varied as a function of false-alarm rate, particularly under conditions of unequal variance (a common occurrence in recognition memory studies). Thus, although using confidence ratings can improve the accuracy and consistency of estimated *G,* systematic deviations from actual γ_d and variation with response bias (false-alarm rate) continue to pose problems for the interpretation of this measure. The same problems apply to G^*.

Simulation 2: Rectangular Distributions

We have proposed that the primary source of the discrepancy between observed *G* and actual γ_d is the loss of information associated with pairs of items that are tied with respect to the subject's responses or the stimulus categories. Moreover, we argue that the variation in *G* across changing response criteria is the result of a shift in the number of item pairs that constitute ties and in the relative numbers of those tied pairs that would have been concordant or discordant pairs had subjects been able to make more fine-grained discriminations. For example, in Simulation 1, consider the case in which the standard deviation ratio is 0.6 and $\gamma_d = .6$ (see Figure 2). Note the substantial drop in *G* as the false-alarm rate moves from .10 to .50. That drop in *G* is accompanied by a 17% increase in the number of tied pairs and a change in the ratio of concordant ties to discordant ties from 1.7 to 3.2. The larger proportion of concordant ties associated with the larger false-alarm rate is partly responsible for the reduction in *G*. The improvement in the behavior of *G* when finer grained distinctions in responding were introduced (ratings *G*), thereby reducing the number of ties, is consistent with this proposal, although the variability in estimated *G* with response bias was by no means eliminated. Crucially, the change in the number of ties and in the ratio of concordant to discordant pairs treated as ties across changes in response bias will depend on the shape of the evidence distributions from which the items are drawn.

To provide additional evidence for our claims about how ties affect the behavior of *G*, we turned to a simulation in which data were generated from rectangular distributions. This distribution shape was chosen specifically because it afforded a straightforward means of determining the proportions of tied pairs that should be considered as concordant or discordant. In Appendix B, we show how the computation of *G* should be modified to accomplish this goal, and we develop a corrected version of *G* that we refer to as G_c for the special (and probably unrealistic case) of equal-variance rectangular distributions. Computation of G_c includes all observations and does not ignore ties. Rather, tied pairs are classified as concordant or discordant in proportion to what

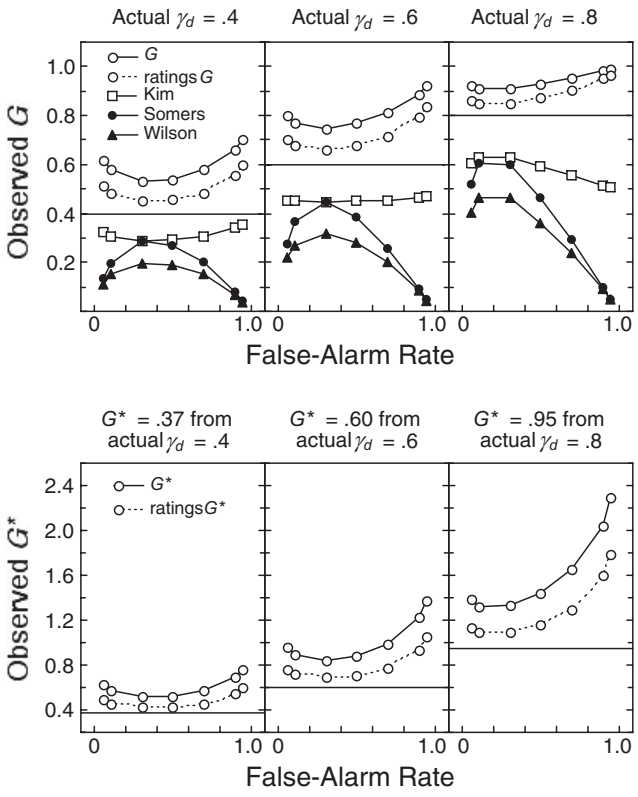

Figure 2. Observed gamma values computed from the data generated in Simulation 1 for Gaussian distributions with equal variance. The upper panels show five versions of observed gamma values: G (the standard Goodman–Kruskal gamma coefficient) based on a binary (e.g., old/new) classification, G based on confidence ratings, and three versions of G that correct for ties, as proposed by Kim (1971); Somers (1962), and Wilson (1974). The lower panels show observed values for G^* computed from a binary classification and from confidence ratings. Each panel represents a different actual value for γ_d, or for G^* derived from γ_d, which is indicated by a horizontal line, where γ_d = population value of gamma specific to the underlying evidence distributions and G^* = a modified measure of gamma introduced by Benjamin and Diaz (2008).

would be expected from overlapping rectangular distributions (see Figure B1).

Method. The validity of G_c was tested in Simulation 2. We again simulated a recognition memory test in which subjects respond "old" or "new" to studied and nonstudied items, but this time the evidence distributions were rectangular in form. Details of the distributions and other aspects of Simulation 2 are shown in Table 2.

Results. The results of this simulation are shown in Figure 4. Note, first, that G changes substantially as false-alarm rate changes, generally decreasing with rising false alarms, then increasing again to form a U-shaped function. It is also the case that at lower, more realistic false-alarm rates, the observed values of G associated with very different actual γ_d values cluster close together. This relative constancy of observed G over changes in true γ_d is especially obvious when the variance of the studied distribution is larger than the variance of the nonstudied distribution (lower panels of Figure 4). In comparison with the results of Simulation 1, in which Gaussian distributions were used to generate the data, the problems with G as a measure of discrimination accuracy are exaggerated.

As expected, G_c provides a much closer approximation to actual γ_d than does the standard measure of G when distributions have equal variance. In that case, G_c performs perfectly, as long as hit rates do not reach 1.0. With ceiling performance on hits (which occurs as the response criterion becomes more liberal and the false-alarm rate approaches 1), G_c drops substantially. This plunge occurs because G_c is directly related to the size of the difference in hit and false-alarm rates (see Equation B10), and this difference approaches 0 when the hit rate reaches ceiling and the false-alarm rate moves toward 1.0. With unequal variance in the studied and nonstudied distributions, G_c changes dramatically across the full range of false-alarm rates, illustrating the fundamental dependence of the validity of G_c on assumptions regarding the evidence distributions (i.e., equal variance, rectangular).

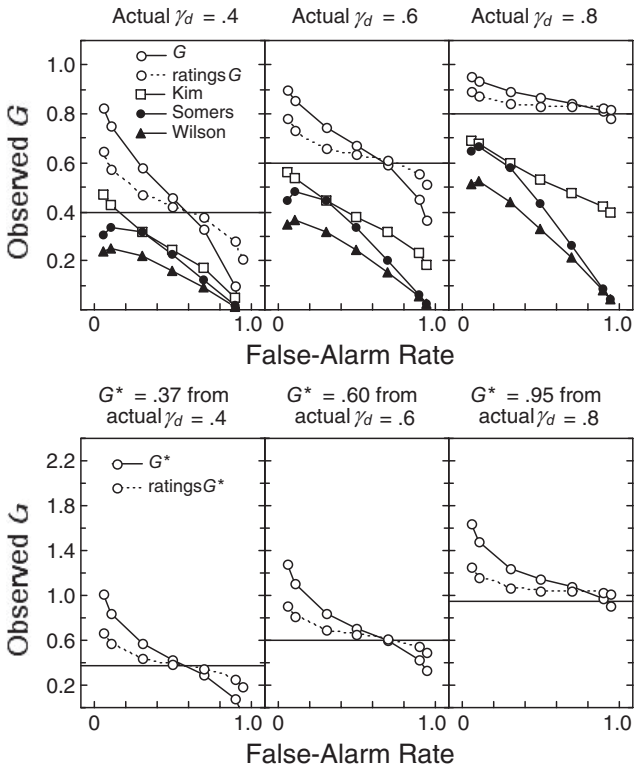

Figure 3. Observed gamma values computed from the data generated in Simulation 1 for Gaussian distributions with unequal variance. The ratio of standard deviation in the noise distribution relative to the signal distribution was 0.6. The panels are arranged as in Figure 2. The upper panels show five versions of observed gamma values: G (the standard Goodman–Kruskal gamma coefficient) based on a binary (e.g., old/new) classification, G based on confidence ratings, and three versions of G that correct for ties, as proposed by Kim (1971), Somers (1962), and Wilson (1974). The lower panels show observed values for G^* computed from a binary classification and from confidence ratings. Each panel represents a different actual value for γ_d, or for G^* derived from γ_d, which is indicated by a horizontal line, where γ_d = population value of gamma specific to the underlying evidence distributions and G^* = a modified measure of gamma introduced by Benjamin and Diaz (2008).

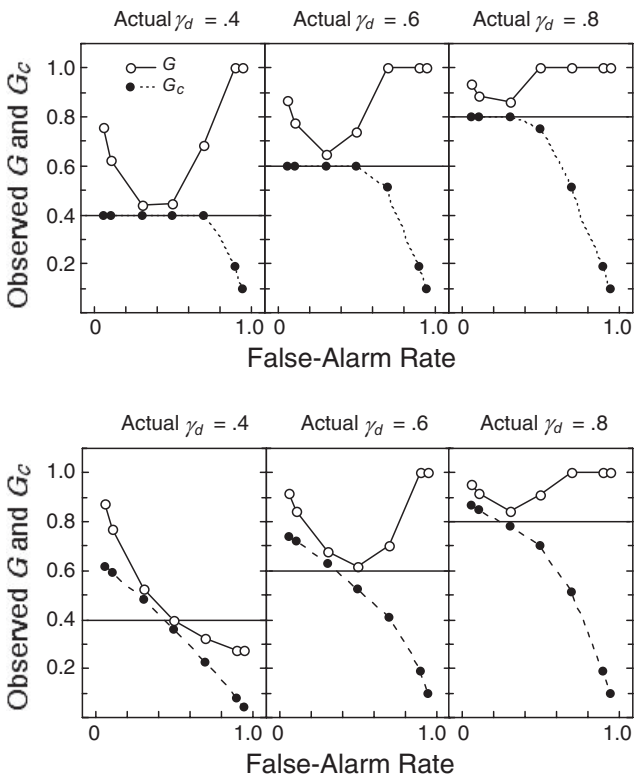

Figure 4. Observed *G* (the standard Goodman–Kruskal gamma coefficient) values computed from the data generated in Simulation 2 for rectangular distributions using Equation 6 and the corrected version of *G* derived in Appendix B, G_c. The upper panels show the equal-variance case, and the lower panels present the data for unequal-variance distributions, where the ratio of noise to signal distribution standard deviations is 0.6. Each panel represents a different actual value for γ_d (population value of gamma specific to the underlying evidence distributions), which is indicated by a horizontal line.

The results of Simulation 2 confirm our claim that the bias in *G* that we have demonstrated is due to the fact that the computation of *G* ignores tied pairs of observations. When correct proportions of these ties can be defined as concordant versus discordant pairs, *G* can be computed with perfect accuracy. This computation, however, requires specific information about the shape of the distributions from which observations are drawn, particularly in the regions occupied by tied observations, and about the relative variance of the distributions. To reinforce this point, we computed G_c for the data generated in Simulation 1. The results are shown in Figure 5. When applied to data generated by Gaussian distributions, the estimated values of G_c show substantial variation across changes in the false-alarm rate while γ_d is held constant. As with the simulation of data from rectangular distributions, the large fall in G_c as the false-alarm rate approaches 1 in the equal-variance case (upper panel of Figure 4) is due to *H* hitting ceiling. Unlike the situation with rectangular distributions, however, G_c changes substantially across lower values of the false-alarm rate even when the distributions have equal variance. It is clear that G_c is useful only when the specific distributional assumptions underlying its derivation hold. We do not, therefore, advocate G_c as a replacement for *G,* and we suspect that it would be rare for distributions underlying actual data to conform to rectangular shape. Our introduction of G_c here was specifically intended to demonstrate the reason for the variation in observed values of *G* as response criterion changes.

Simulation 3: Population γ_d and Estimation Accuracy

We have identified two problems with *G* computed from sample data when it is used as an estimate of discrimination accuracy. First, *G* systematically deviates from the true population value of γ_d, even when a correction for tied observations is included in the computation of *G*. Second, *G* varies, sometimes substantially, as a function of response bias even though the population value γ_d is constant across these changes. Although the concern regarding response bias is demonstrated here for the first time, the traditional statistics literature previously has addressed the properties of *G* as an estimator of γ, defined in the classical manner.

In the statistics literature, interest has centered on the accuracy with which sample-based *G* estimates γ. Goodman and Kruskal (1963, 1972) demonstrated that with asymptotically large samples,

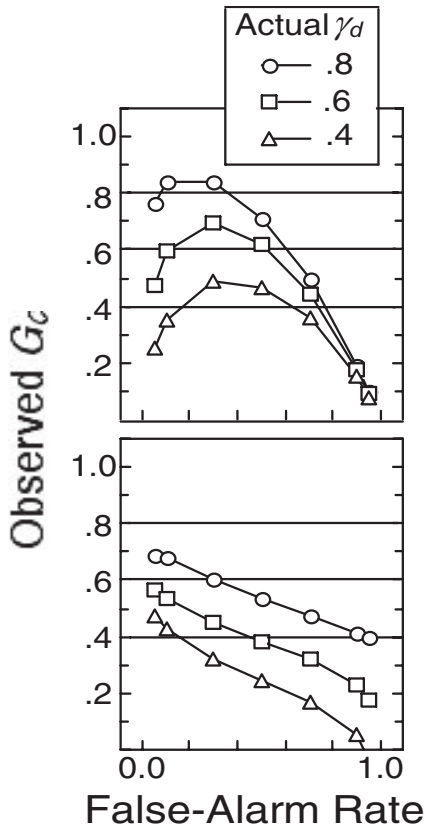

Figure 5. Observed G_c (corrected gamma coefficient assuming rectangular evidence distributions) values computed from the data generated in Simulation 1 for Gaussian distributions. The upper panel presents results for the equal-variance case, and the lower panel displays results for a noise/signal standard deviation ratio of 0.6. Each symbol type shows the value of G_c computed when a different actual value of γ_d (population value of gamma specific to the underlying evidence distributions) was in effect. Actual values of γ_d are also indicated by the horizontal lines in each panel.

G is an unbiased estimate of γ. For small to moderate sample sizes and the 2×2 and 2×3 cases, however, the distribution of sample G values has rather high variability and is irregular and skewed, resulting in a biased estimate of γ. Moreover, the distribution of G converges very slowly to the ideal form reported by Goodman and Kruskal for the asymptotic case (Gans & Robertson, 1981a, 1981b; see also Lui & Cumberland, 2004; Rosenthal, 1966; Woods, 2007).

Method. In Simulation 3, we examined bias and variability in observed values of G relative to γ_d by simulating data for a large number of subjects, each of whom contributes a G value determined by a defined number of observations, as is done in the statistical literature when assessing the convergence of observed G to actual γ. We also included G^* (Benjamin & Diaz, 2008) in these simulations to compare its convergence properties to those of G. Details of this simulation are shown in Table 2. The numbers of trials on which each simulated subject's data were based matched the numbers used by Verde, Macmillan, and Rotello (2006) in their investigation of alternative measures of discrimination accuracy, thereby allowing us to compare our results with those they reported.

For simulation of the old/new task, we computed G and G^* for each subject but excluded any subjects for whom these measures were undefined (e.g., G is undefined if $H = F = 0$; G^* is undefined if $G = 1.0$). To reduce the number of observations lost due to extreme outcomes, following Verde et al. (2006), we also computed G and G^* based on a log-linear transformation of hit and false alarm rates, such that $H = (T_h + 0.5)/(T + 1)$ and $F = (T_f + 0.5)/(T + 1)$, where T_h and T_f are the number of hits and false alarms, respectively, and T is the number of signal trials and the number of noise trials. Loss of observations due to extreme outcomes was not a problem in Simulations 1 and 2 because in those cases, we were not simulating individual subjects whose data were based on relatively few observations. Rather, those simulations were based on tens of thousands of trials for each computation of a measure, so ceiling or floor values (1 and 0) were never encountered.

Results. For each run of the simulation, the mean and standard deviation was computed for each measure across the simulated subjects, excluding any of those subjects for whom a measure was undefined. These values are shown in Figure 6 as a function of number of trials and false-alarm rate for the case of equal-variance distributions and $\gamma_d = .6$. The results for γ_d values of .4 and .8 are shown in Appendix C. As in Simulation 1, both G and G^* overestimate the actual value of γ_d, and both measures vary with changes in false-alarm rate even though true accuracy is constant. Increasing the number of observations on which each simulated subject's measure is based clearly reduced the variability among subjects, but it did not lead to convergence of G or G^* to the actual value of γ_d, nor did it reduce, in any substantial way, the variation in measures of G or G^* as a function of false-alarm rate. It is difficult to compare these results with those reported in the statistics literature, because, in addition to our different definitions of the true value of γ, there are many other ways in which our simulations differed. Gans and Robertson (1981a) reported the mean and standard deviation of estimated G for a few different values of γ calculated with Equation 6 for samples sizes of 10, 30, or 50 trials. For their Condition I, the value of γ was .75 and the false-alarm rate was .10; for 50 simulated trials, the resulting mean

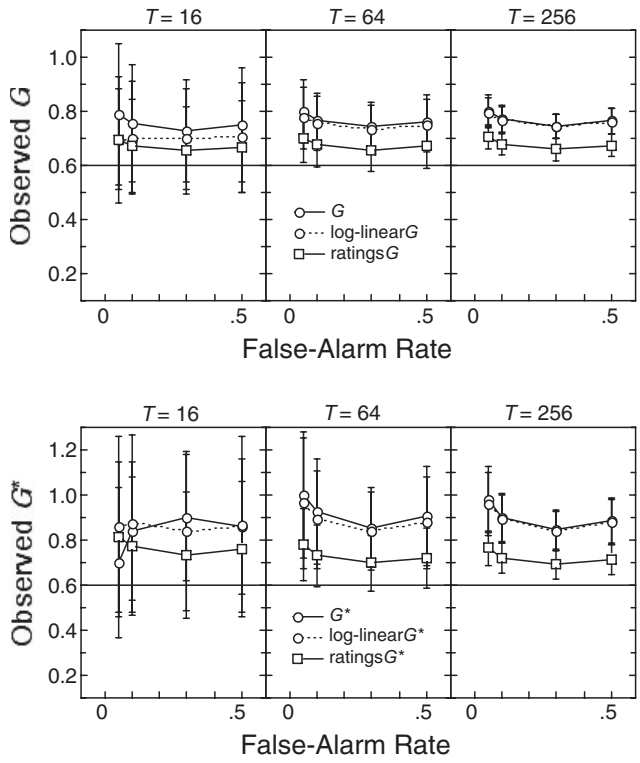

Figure 6. Observed mean gamma values computed from the data generated in Simulation 3 for Gaussian distributions with equal variance. The upper panels show three versions of observed gamma values: G (the standard Goodman–Kruskal gamma coefficient) based on simple hit and false-alarm rates from a binary (e.g., old/new) classification, G based on a log-linear conversion of hit and false-alarm rates, and G based on confidence ratings. The lower panels show corresponding observed values for G^* (a modified measure of gamma introduced by Benjamin & Diaz, 2008). Each panel represents a different number of trials (T) on which each simulated subject's data were based. The actual value for γ_d (population value of gamma specific to the underlying evidence distributions) was .6 and is indicated by a horizontal line in each panel. A γ_d value of .6 corresponds to a G^* value of .6. Error bars show the standard deviation for each measure.

estimated G (.74) was quite similar to the true value, a result that can be understood to arise from the use of Equation 6 to calculate both estimated G and γ. For the same condition, Gans and Robertson reported the standard deviation of estimated G to be 0.15. We can best compare those results with our equal-variance simulation, in which $\gamma_d = 0.8$, false-alarm rate = .10, and 64 trials were run per subject (reported in Appendix C). Our simulations revealed a larger bias in mean estimated G (.91) stemming from our more accurate definition of the population value (γ_d), and a much smaller standard deviation (0.04), which presumably was the result of the particular characteristics of our evidence distributions (equal-variance Gaussian distributions).

Figure 7 shows observed G and G^* values for the unequal-variance case with $\gamma_d = .6$. (Corresponding results for $\gamma_d = .4$ and .8 are shown in Appendix C.) Here, the influence of response bias on observed G measures is even more striking than in the equal-variance case, as was true in Simulation 1. Although for both the

equal- and the unequal-variance cases, the *G* estimates seem to converge toward the true value of γ_d as the false-alarm rate increases, Simulation 1 shows that this is a deceiving trend in that *G* estimates increase sharply once false-alarm rate exceeds .5 in the equal-variance case (see Figure 2), and they decrease markedly in the unequal-variance case, leading to underestimates of γ_d (see Figure 3).

Comparable simulations reported by Verde et al. (2006) using the signal detection measure A_z, which is closely related to the d_a measure presented earlier in the introduction (see Appendix A), show that A_z has much smaller variability than does *G*. For example, across a range of false-alarm rates, with actual accuracy similar to $\gamma_d = .6$, and equal-variance Gaussian distributions, Verde et al. obtained standard deviations of the A_z estimate ranging from approximately .08 to .02 for sample sizes that varied from 16 to 256. In our simulations with *G*, the corresponding standard deviations ranged from approximately .20 to .05 (ratings *G* was only slightly better and log-linear *G* was worse). The sampling variability for G^* was even larger than that for *G*, in part because G^* takes on a wider range of possible values. Thus, Simulation 3, in conjunction with the Verde et al. simulations, shows that sampling variability for *G* is generally larger than for A_z, a measure based on signal detection theory.

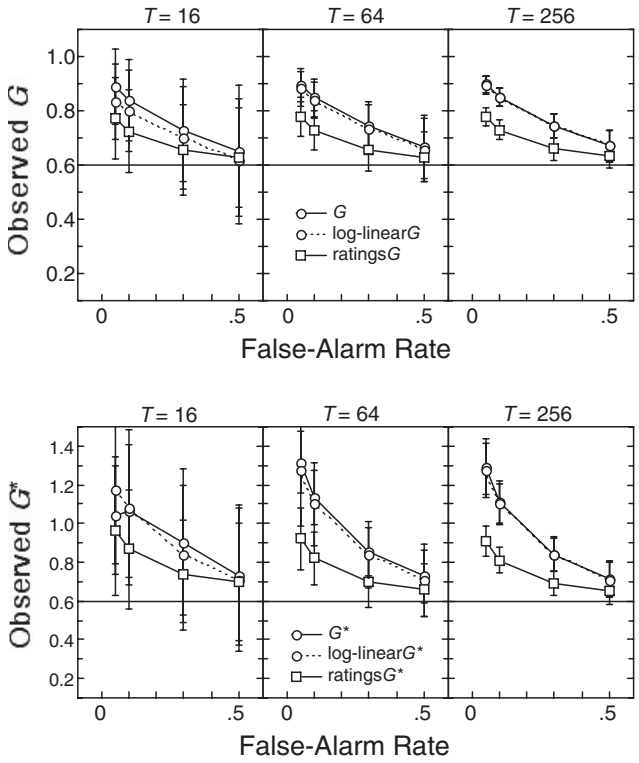

Figure 7. Observed mean gamma values computed from the data generated in Simulation 3 for Gaussian distributions with unequal variance. The ratio of standard deviation in the noise distribution relative to the signal distribution was 0.6. The panels are arranged as in Figure 6: The upper panels show three versions of observed gamma values: *G* (the standard Goodman–Kruskal gamma coefficient) based on simple hit and false-alarm rates from a binary (e.g., old/new) classification, *G* based on a log-linear conversion of hit and false-alarm rates, and *G* based on confidence ratings. The lower panels show corresponding observed values for G^* (a modified measure of gamma introduced by Benjamin & Diaz, 2008). Each panel represents a different number of trials (T) on which each simulated subject's data were based. The actual value for γ_d (population value of gamma specific to the underlying evidence distributions) was .6 and is indicated by a horizontal line in each panel. A γ_d value of .6 corresponds to a G^* value of .6. Error bars show the standard deviation for each measure.

A Signal Detection Alternative to G

Desirable properties of an accuracy measure are independence from response bias, low statistical bias, and a low standard error. Our first three simulations showed that *G* and ratings *G* fail on all counts. In addition, a successful measure must be consistent with the observed form of the empirical receiver operating characteristic (ROC). Swets (1986a) showed that, for a range of disciplines, empirical ROCs are consistent with underlying evidence distributions that are approximately Gaussian in form and unequal in variance (like the recognition ROCs in Figure 8 and the metamemory ROCs in Figure 9). Here, too, *G* fails because it is most consistent with equal-variance logistic distributions (Swets, 1986b). Requiring an accuracy measure to be consistent with unequal-variance Gaussian distributions eliminates a number of commonly used measures, including percentage or proportion correct, recognition scores corrected for bias (i.e., *Pr* = hit rate – false-alarm rate), and A' (Verde et al., 2006; Rotello, Masson, & Verde, 2008). Even d' fails these tests, as it does not always have low statistical bias or a low standard error (Miller, 1996), and it is independent of response bias only when the data are drawn from equal-variance Gaussian distributions (Macmillan & Creelman, 2005; Rotello et al., 2008).

In contrast, the accuracy measure A_z, which estimates the area under the ROC for data drawn from any distributions that can be monotonically transformed to equal- or unequal-variance Gaussian distributions, satisfies all of our criteria: It has very low statistical bias, has a low standard error, and is not affected by the standard deviation ratio of the underlying distributions (Macmillan, Rotello, & Miller, 2004; Verde et al., 2006). Indeed, Swets (1986a) argued that the consistency of the form of the empirical ROCs across fields "supports the use of the accuracy index A_z" (p. 196). The variable d_a is a distance-based measure that shares the desirable properties of A_z, including independence from response bias (see Macmillan et al., 2004; Rotello et al., 2008), and is monotonically related to it (see Equation A1). Thus, use of either A_z or d_a is well justified.

Simulation 4: A Simulated Metacognitive Experiment

In Simulation 4, we demonstrate that with realistic numbers of observations and subjects, *G* and ratings *G* are likely to produce artifactual indications of accuracy differences between conditions that vary only with respect to response bias. Moreover, we demonstrate that the signal detection–based measure d_a avoids this problem. We base our simulation on the observation that a number of experiments on metamemory have shown that the estimates of the FOK or JOLs are influenced by factors such as the delay between study exposure and the metamemory decision on each item. *G* (or ratings *G*) is smaller when a JOL for an item is made at the same time that the item is studied, relative to when JOLs are

collected after even a brief delay (for a review, see Schwartz, 1994). Subject-related factors have also been shown to influence *G* in theoretically interesting ways. For example, Souchay, Moulin, Clarys, Taconnat, and Isingrini (2007) reported that older adults' FOK judgments on episodic memory tasks were less accurate than those of younger adults': Ratings *G* was lower for the older subjects. The authors' interpretation was that older adults are less able to use a recollective memory process that is essential for accurate episodic FOK judgments. Another explanation of this age difference in ratings *G* is that it is the result of a response bias effect. That is, the older adults in Souchay et al.'s Experiment 2 used a more liberal response criterion than did the younger adults (false-alarm rate = 0.22 for older adults and 0.16 for younger adults).[2] The results of our Simulation 1 (Figures 2 and 3) showed that both *G* and ratings *G* decrease as the response criterion becomes more liberal (at least over the range of false alarms exhibited by the two age groups in the Souchay et al. study). It is reasonable to ask, then, whether the difference in the observed gamma correlations is a real effect based in a memory or metamemory process or an effect that is driven by the different response criteria used by the younger and older subjects. Given only the data reported by Souchay et al., we cannot definitively separate accuracy effects from response bias effects.

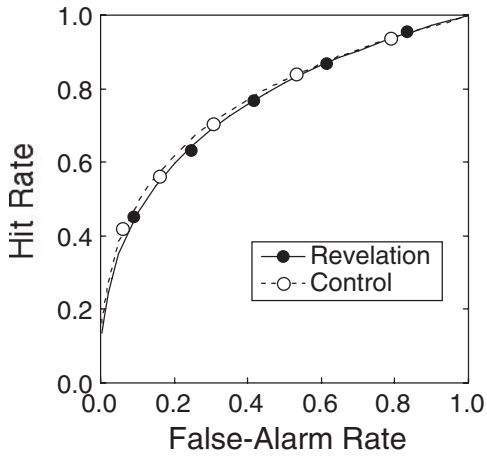

Figure 9. Receiver operating characteristics for the data from Verde and Rotello (2003). Curves are the best fitting functions derived from the data for each condition.

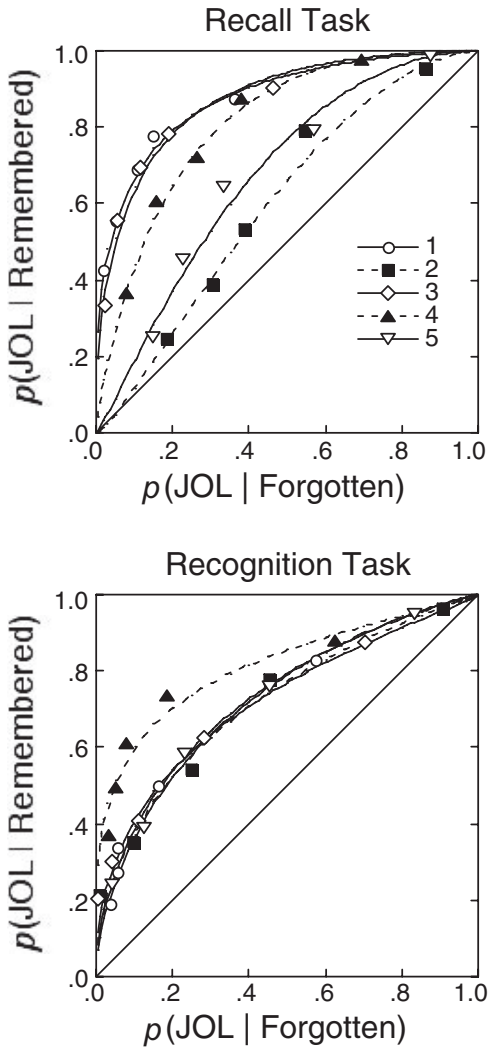

Figure 8. Receiver operating characteristics for the recall and recognition data from Weaver and Kelemen (2003). Curves are the best fitting functions derived from the data for each of five study conditions (defined in the text). JOL = judgment of learning.

In this section, we simulate the consequences of a response-bias difference across conditions in a metacognitive experiment. The resulting data are reported for *G*, ratings *G*, and two signal-detection based statistics, d' and d_a. Other measures related to *G* that we examined in earlier simulations, such as G^*, show the same pattern of behavior as *G*, both in Simulation 4 and in our evaluation of actual empirical data presented below. The advantage of using a simulation is that we know the two simulated conditions have a common level of true accuracy but different response biases. Thus, we can evaluate the probability that significance tests with each accuracy measure would lead the researcher to the incorrect conclusion that the two conditions differ in accuracy (e.g., that *G* differs reliably between groups). In other words, Simulation 4 estimates the Type I error rate for each measure under conditions of changing response bias (see also Rotello et al., 2008). In the next section, we consider how these same accuracy measures fare with actual empirical data from two published experiments.

Method. In Simulation 4, we assumed that there are two experimental conditions in which each subject has the same true accuracy level of $\gamma_d = 0.4$ but that subject in the conservative condition have an average false-alarm rate that is lower than that of subjects in the liberal condition. Details of the evidence distri-

[2] Most metamemory experiments use recall or cued recall to assess memory performance. Of those that use recognition, most report only the overall percentage of correct memory judgments, from which response bias cannot be determined. Of course, the absence of information about response bias does not eliminate the problems created by differences in response bias.

butions and other aspects of the simulation are shown in Table 2. Note that in this simulation, larger variance was assumed for the distribution of remembered items than for the distribution of forgotten items.

Results. If G and ratings G were unbiased estimators of true γ_d, then both would have average values of around 0.4. Table 3 shows that both measures overestimate the true value, and the amount of overestimation is larger in the conservative condition. In addition, all of the 1,000 t tests (one per simulated experiment) declared that the two conditions differed significantly, both on G and on ratings G. In other words, the Type I error rate for these measures was 1.0. Similar results were reported by Rotello et al. (2008) for a wide range of simulated experimental conditions.

Use of the signal detection measure d' to assess memory accuracy in this simulation also led to a Type I error rate of 1.0 (see Table 3 and Rotello et al., 2008) because d' is confounded with response bias when the underlying distributions are of unequal variance (Macmillan & Creelman, 2005). In contrast, the d_a measure, which generalizes d' to the unequal-variance situation (see Equation 2), yielded a Type I error rate near the standard alpha level of 5%. Moreover, the mean d_a value in each condition is 0.75, which is essentially identical to the theoretical value of 0.74 for these distributions.

The implications of Simulation 4, as well as the related simulations reported by Rotello et al. (2008), are quite broad. Use of G, ratings G, or d' to measure metacognitive performance may mislead researchers about the nature of the differences between conditions. When the underlying metamemory or memory distributions have unequal variance and only response bias differs between conditions, all of these measures misattribute those bias differences to accuracy effects. Only d_a provides an appropriate description of metamemory accuracy.

Two Numerical Examples With Real Data

Using simulations, we have demonstrated that G is associated with four disadvantages relative to accuracy measures based on signal detection: (a) It is modulated by varying response criteria even when actual accuracy is constant; (b) it does not asymptotically converge on the true value of γ_d; (c) its sampling variability is large; and (d) it yields an unacceptably high Type I error rate when accuracy is constant across conditions but response bias differs. We now consider how G and ratings G, as well as the signal-detection–based measures d' and d_a, fare in two real data sets. For our evaluations, we chose examples in which confidence ratings were collected so that the form of the underlying evidence distributions, and therefore true sensitivity, could be estimated accurately. The availability of confidence ratings also allows computation of both ratings G and d_a. We selected one example from the metamemory literature (Weaver & Kelemen, 2003) in which G is used most heavily and a second example from the recognition memory literature where d' dominates (Verde & Rotello, 2003).

Metamemory example. Weaver and Kelemen (2003) asked subjects to make JOLs on cue–target word pairs (e.g., *ELEPHANT–sunburn*) and later to recall the target word in response to the cue or to recognize the correct cue–target pair from a set that included 6 lure pairs. The theoretical motivation for the authors' study was to test whether the accuracy of the JOLs reflected transfer-appropriate monitoring. That is, if the JOLs were generated under conditions that more closely mimicked the test conditions, would they be more accurate? To find out, Weaver and Kelemen collected the JOLs in five conditions. In Condition 1, the JOL probe consisted of only the cue word (*ELEPHANT–*?) from each pair. This condition mimicked a cued recall task. In Condition 2, the JOL probe consisted of the cue–target pair (*ELEPHANT–sunburn*). In Condition 3, the JOL probe was the cue word (*ELEPHANT–*?), but it was shown in the context of 6 lure pairs (*ELEPHANT–diamond, ELEPHANT–hillside, ELEPHANT–sugar,* etc.); this condition reflected a mixture of cued-recall and recognition tasks. Condition 4 was similar to Condition 3, except that the entire cue–target pair (*ELEPHANT–sunburn*) was presented with the lures; this condition mimicked the recognition task. Finally, Condition 5 was identical to Condition 4 except that the correct cue–target pair was marked (*ELEPHANT–sunburn****).

Weaver and Kelemen (2003) reported ratings G correlations of the JOLs in each condition with the eventual memory performance, predicting that Condition 1 would show the highest ratings G when computed using cued recall data and that Condition 4 would show the highest ratings G when computed using recognition data. The results, shown in the ratings G column of Table 4, do not support the transfer-appropriate monitoring hypothesis. Ratings G was higher in Condition 1 than in Conditions 2 or 5 when recall data were used, but it was equal to that in Conditions 3 and 4. When the recognition data were used, Condition 4's ratings G was among the lowest of all. For these reasons, Weaver and Kelemen rejected transfer-appropriate monitoring as a possible theoretical account of metamemory judgments, concluding that alternative theories better accounted for their results.

Table 3
Results of Simulation 4: A Simulated Metacognition Experiment in Which Conditions Differ Only in Response Bias

	Measure					
Simulated condition	H	F	d'	G	Ratings G	d_a
Conservative	0.35	0.05	1.33	0.82	0.64	0.75
Liberal	0.62	0.30	0.83	0.57	0.47	0.75
Type I error rate over 1,000 simulated experiments			1.00	1.00	1.00	0.06

Note. Values for d' are based on log-linear corrections for 0s and 1s in the hit and false-alarm rates of individual simulated subjects. $H = p(\text{FOK} \geq 50 \mid \text{remembered})$; $F = p(\text{FOK} \geq 50 \mid \text{forgotten})$, where FOK (feeling of knowing) is rated on a 100-point scale. d' = signal detection measure of accuracy assuming equal-variance evidence distributions; G = the standard Goodman–Kruskal gamma coefficient; d_a = signal detection measure of accuracy allowing for unequal-variance evidence distributions.

Table 4
Reanalysis of Data From Weaver and Kelemen (2003)

Memory task/ JOL condition	Measure					
	H	*F*	*d′*	*G*	Ratings *G*	d_a
Recall						
1*	0.69	0.11	1.74	0.90	0.84	1.64
2	0.53	0.39	0.41	0.29	0.44	0.40
3	0.69	0.11	1.71	0.89	0.83	1.56
4	0.72	0.26	1.23	0.76	0.84	1.29
5	0.64	0.34	0.79	0.56	0.57	0.66
Recognition						
1	0.34	0.06	1.18	0.79	0.53	0.87
2	0.54	0.25	0.78	0.56	0.45	0.86
3	0.41	0.11	0.98	0.69	0.59	0.84
4*	0.61	0.08	1.69	0.90	0.49	1.27
5	0.58	0.23	0.93	0.64	0.37	0.85

Note. JOL (judgment of learning) conditions marked with an asterisk were predicted by Weaver and Kelemen (2003) to produce the best metacognitive performance. $H = p(\text{JOL} \geq 50 \mid \text{remembered})$; $F = p(\text{JOL} \geq 50 \mid \text{forgotten})$, where JOL is rated on a 100-point scale. d' = signal detection measure of accuracy assuming equal-variance evidence distributions; G = the standard Goodman–Kruskal gamma coefficient; d_a = signal detection measure of accuracy allowing for unequal-variance evidence distributions.

We used the JOL ratings to create metamemory ROCs, employing a method that Benjamin and Diaz (2008) described in detail (see also Nelson, 1984). Analogous to the recognition ROCs, metamemory ROCs plot accurate responses (remembered items) on the *y*-axis and errors (forgotten items) on the *x*-axis as a function of JOL rating. Thus, the first point on each ROC reflects metamemory accuracy for items that subjects were sure they would remember (JOL = 100%), the second point indicates responses to items that subjects were quite confident about (JOL ≥ 80%), and so forth. The resulting metamemory ROCs are shown in Figure 8; curves higher in the space reflect greater metamemory accuracy. There are two striking effects. First, Conditions 1 and 3 fall higher than the others when the recall data are considered; second, Condition 4 falls well above the others when recognition data are used. Weaver and Kelemen predicted this pattern of metamemory accuracies on the principle of transfer-appropriate processing but did not observe it with ratings *G*.

We computed d', d_a, and nonrating *G* for these data, assuming that a hit was a remembered item to which subjects had assigned a JOL of at least 50%, and a false alarm was a forgotten item given a similar JOL. Thus, $H = p(\text{JOL} \geq 50 \mid \text{remembered})$ and $F = p(\text{JOL} \geq 50 \mid \text{forgotten})$. Because the metamemory ROCs in Figure 8 are roughly consistent with underlying distributions that are unequal-variance Gaussian in form, and because response bias differences are evident across conditions, we expected that d_a would provide the best description of metamemory accuracy, as it did in Simulation 4. In other words, we expected d_a to show a pattern consistent with the varying heights of the metamemory ROCs across JOL conditions. The resulting values, shown in Table 4, are consistent with that expectation.[3] The implication of this analysis is startling. Had Weaver and Kelemen used d_a rather than ratings *G* to assess metamemory accuracy, they would have reached the opposite theoretical conclusion, finding support for their preferred theory (transfer-appropriate monitoring) and against the alternatives.

Recognition memory example. Verde and Rotello (2003) collected confidence ratings in three within-subject recognition memory experiments. In each experiment, subjects studied a list of words and then were given a recognition test that consisted of both studied and nonstudied items. In the revelation conditions, subjects were asked to solve an anagram immediately prior to making their recognition judgment for a particular item (the anagram and test word were unrelated); in the control conditions, there were no anagrams. The substantive question of interest was whether memory sensitivity was affected by the revelation task. In each experiment, the hit and false-alarm rates were both higher in the revelation condition than in the control condition (see Table 5). These response rates can be used to compute single-point sensitivity statistics. In the literature on the revelation effect, *d'* is used most commonly. Its application suggests that accuracy is decreased by the revelation task, significantly so when considered across experiments, $t(81) = 2.09$, $p < .05$. Application of *G* supports that conclusion, $t(81) = 1.71$, $p = .09$, but use of ratings *G* reduces the difference to a nonsignificant level, $t(81) = 1.06$, $p > .25$. On the basis of *d'* results like these, the revelation effect previously had been attributed to an increase in the familiarity of the test items during the revelation task (see Hicks & Marsh, 1998, for a meta-analysis).

In this case, however, only ratings *G* accurately describes the effect of the revelation task on memory accuracy. Verde and Rotello (2003) used the confidence rating data to estimate the standard deviation ratios of the distributions underlying the nonstudied and studied items, which was about 0.8. When the standard deviation ratio is less than 1, larger values of d' are estimated for more conservative response criteria that result in lower hit and false-alarm rates (i.e., the control conditions). Analogous problems exist for *G* (and ratings *G*; see Figures 2 and 3 and Table 3), which is consistent with equal-variance logistic distributions (see Swets, 1986b). Verde and Rotello computed the more appropriate statistic d_a and found no sensitivity differences across conditions, $t(81) = 0.06$. Consistent with this conclusion, we show in Figure 9 the ROCs for the revelation and control conditions on the basis of confidence-rating data aggregated across the three experiments (these data are available in the Appendix of the Verde & Rotello article). Note that the two curves in Figure 9 fall at a nearly identical height in the space, indicating very similar levels of discrimination accuracy. Use of the more appropriate statistic d_a provided theoretical insight on the revelation literature: Familiarity-based explanations of the effect were eliminated for revelation designs in which the anagram and the test item were

[3] Bill Kelemen very kindly provided the data from this experiment, but the ROCs were extremely unstable at the individual subject level because there were too few trials per condition. Therefore, statistical comparison of these measures across JOL conditions was not possible, and we based our estimates of *G* and the signal detection measures on the group data. The individual data do allow clarity on one curious aspect of these data, which is that ratings *G* does not appear to capture the pattern of performance as well as *G*. A number of subjects achieved perfect recall or recognition in one condition or another, and *G* cannot be computed from data of this form because all items pairs are ties (see the program in Nelson, 1986a). Thus, these subjects were omitted from the analyses reported by Weaver and Kelemen (2003). When ratings *G* is calculated on the group data, however, the resulting values mimic the order of the ROCs in Figure 8.

Table 5
Reanalysis of Data From Verde and Rotello (2003)

Experiment/ condition	Measure					
	H	*F*	*d′*	*G*	Ratings *G*	d_a
Experiment 1						
Revelation	0.77	0.41	1.06	0.63	0.59	1.12
Control	0.71	0.31	1.19	0.68	0.61	1.08
Experiment 2						
Revelation	0.78	0.46	0.96	0.60	0.57	1.02
Control	0.71	0.32	1.10	0.65	0.59	1.07
Experiment 3						
Revelation	0.76	0.39	1.07	0.65	0.60	1.08
Control	0.69	0.28	1.17	0.68	0.60	1.06
Overall						
Revelation	0.77	0.42	1.04	0.63	0.58	1.08
Control	0.70	0.31	1.16	0.67	0.61	1.09

Note. H = hit rate; F = false-alarm rate. d' = signal detection measure of accuracy assuming equal-variance evidence distributions; G = the standard Goodman–Kruskal gamma coefficient; d_a = signal detection measure of accuracy allowing for unequal-variance evidence distributions.

unrelated (as in Figure 9), but, in concert with response bias explanations, were supported for revelation designs in which the anagram and the test item were identical (Verde & Rotello, 2003, 2004).

The apparent sensitivity effect revealed by use of *d'* and *G* reflects susceptibility to Type I errors that arises when the assumptions underlying a sensitivity measure are violated (see Rotello et al., 2008). Although one might guess that use of confidence ratings to estimate *G* would alleviate the confound between response bias and apparent sensitivity in these experiments, that method of calculating *G* addresses, albeit not completely, the issue of ties rather than the violation of the equal-variance assumption. As a consequence, ratings *G* does not consistently lead to accurate comparisons of discriminability across conditions, as can be seen in Figures 2 and 3 as well as Table 3.

Conclusion and Recommendations

The elegant probabilistic interpretation of *G* developed by Nelson (1984, 1986b) provided an appealing basis for computing discrimination accuracy or association. This interpretation has supported the application of *G* in assessments of metacognition and metamemory performance. Indeed, the use of ratings *G* in these domains is standard practice. We have demonstrated, however, that *G* computed from simulated sample data does not converge on the actual value of γ_d and also systematically varies with response bias so that artifactual effects, due only to response bias and not to genuine differences in accuracy, may arise if *G* is used as a measure of accuracy. These artifacts are exaggerated in, but not restricted to, the common situation in which the underlying strength distributions have unequal variances.

Nelson (1986b) challenged the foundational signal detection theory assumption that accuracy and response bias are necessarily independent and argued instead that this independence must be empirically established, not merely assumed. In his view, situations that lead to shifts in decision threshold may well be associated with changes in discrimination accuracy. Thus, the behavior of *G* across variations in false-alarm rates that might be observed empirically could reflect valid changes in discrimination accuracy. As our simulations show, however, the inconstancy of observed values of *G* over changes in response criterion readily arises even when the true population value it estimates is unchanged. In Simulations 3 and 4, we have also shown that these problems occur with realistic numbers of trials and samples of subjects. Notably, we have identified the reason for this discrepancy between observed *G* and the actual value of γ_d. The problem lies with the fact that in typical data sets, a portion of the data (i.e., ties) are ignored when *G* is computed. The exclusion of ties from the computation leads to systematic bias in the values of *G* that are generated. Only when knowledge of the underlying distributions is available can ties be properly treated in the computation of *G* and this bias be avoided (see Appendix B and Simulation 2). Consequently, the behavior of *G*, like other measures of discrimination accuracy such as d' and even A' (see Grier, 1971), very much depends on the characteristics of the underlying evidence distributions. (For treatments of the distributional assumptions underlying A' and its lack of independence from response bias, see Macmillan & Creelman, 1996; Pastore, Crawley, Berens, & Skelly, 2003; Smith, 1995; Snodgrass & Corwin, 1988; and Verde et al., 2006.)

The results we have presented reveal a bias in *G* as a measure of discrimination accuracy that emerges even under circumstances that are deemed most typical in psychological studies (i.e., the underlying distributions are Gaussian). We have shown elsewhere that relative to d', this bias in *G* makes it more susceptible to Type I errors for both Gaussian and rectangular distributions (Rotello et al., 2008). Thus, we do not concur with Nelson's (1986b) recommendation to prefer *G* over d' as a measure of accuracy when distributional information is not available. Rather, we conclude that d' is a less problematic, albeit not ideal, single-point measure of discrimination accuracy (see also Schooler & Shiffrin, 2005).

It is clear that information about the nature of underlying distributions is critical in selecting the most appropriate measure of accuracy. Our recommendation is that whenever possible, researchers obtain multiple-point measures of discrimination accuracy. This objective can be accomplished through procedures such as having subjects provide a confidence rating when making a binary response classification (e.g., old/new in a recognition memory task). In studies of metacognition, it is typical that subjects are asked to provide some form of confidence rating when making predictions about, for example, future memory performance. With multiple points available, one can construct an ROC and, under the assumption of Gaussian distributions (or distributions that can be monotonically transformed to Gaussian distributions), estimate the relative variance in the signal and noise distributions (see Appendix A). With this information, an accuracy measure grounded in signal detection theory, d_a, can be computed. As shown in Equation 2, this measure is a generalization of d' that is sensitive to possible differences in the variances of the signal and noise distributions. Researchers, such as those in the areas of metacognition and metamemory, who typically compute a ratings-based version of *G* can instead compute d_a with no need to change data collection procedures. Even though the validity of d_a depends

on the assumption that underlying distributions are Gaussian or can be monotonically transformed to that shape, this measure is preferable to *G*, which we have shown here to vary systematically with response criterion under a variety of distributional assumptions. Moreover, examination of the empirical ROCs can provide some indication of whether the Gaussian assumption is justified.

One could develop a correction for *G* that takes into account distributional properties, as we have done for the case of equal-variance rectangular distributions. As we have shown, however, such corrections are highly specific to particular distributional parameters and potentially break down with moderate to high false-alarm rates. We recommend d_a as a more robust alternative and one that reacts properly to differences in the variance of signal and noise distributions as well as changes in response bias.

References

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Benjamin, A. S. (2005). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about purportedly nonparametric measures. *Memory & Cognition, 33,* 261–269.

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73–94). New York: Psychology Press.

Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the confidence–accuracy relationship across domains. *Journal of Experimental Psychology: Applied, 5,* 76–88.

Criss, A. H., & Shiffrin, R. M. (2004). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition, 32,* 1284–1297.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545–565.

Eng, J. (n.d.). *ROC analysis: Web-based calculator for ROC curves.* Retrieved August 23, 2007, from http://www.jrocfit.org

Freeman, L. C. (1986). Order-based statistics and monotonicity: A family of ordinal measures of association. *Journal of Mathematical Sociology, 12,* 49–69.

Gans, L. P., & Robertson, C. A. (1981a). Distributions of Goodman and Kruskal's gamma and Spearman's rho in 2 × 2 tables for small and moderate sample sizes. *Journal of the American Statistical Association, 76,* 942–946.

Gans, L. P., & Robertson, C. A. (1981b). The behavior of estimated measures of association in small and moderate sample sizes for 2 × 3 tables. *Communications in Statistics—Theory and Methods, 10,* 1673–1686.

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

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119,* 159–165.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–764.

Goodman, L. A., & Kruskal, W. H. (1963). Measures of association for cross classifications. III: Approximate sampling theory. *Journal of the American Statistical Association, 58,* 310–364.

Goodman, L. A., & Kruskal, W. H. (1972). Measures of association for cross classifications. IV: Simplification of asymptotic variances. *Journal of the American Statistical Association, 67,* 415–421.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75,* 424–429.

Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1105–1120.

Kao, Y.-C., Davis, E. S., & Gabrieli, J. D. E. (2005). Neural correlates of actual and predicted memory formation. *Nature Neuroscience, 8,* 1776–1783.

Kim, J. O. (1971). Predictive measures of ordinal association. *American Journal of Sociology, 76,* 891–907.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135,* 36–69.

Lui, K.-J., & Cumberland, W. G. (2004). Interval estimation of gamma for an $r \times s$ table. *Psychometrika, 69,* 275–290.

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

Masson, M. E. J., & Hicks, C. L. (1999). The influence of selection for response on repetition priming of word identification. *Canadian Journal of Experimental Psychology, 53,* 381–393.

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

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson, T. O. (1986a). BASIC programs for computation of the Goodman–Kruskal gamma coefficient. *Bulletin of the Psychonomic Society, 24,* 281–283.

Nelson, T. O. (1986b). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100,* 128–132.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: "The delayed-JOL effect". *Psychological Science, 2,* 267–270.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 384–413.

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

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

Rosenthal, I. (1966). Distribution of the sample version of the measure of association, gamma. *Journal of the American Statistical Association, 61,* 440–453.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70,* 389–401.

Schooler, L. J., & Shiffrin, R. M. (2005). Efficiently measuring recognition performance with sparse data. *Behavior Research Methods, 37,* 3–10.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review, 1,* 357–375.

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.

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

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review, 27,* 799–811.

Souchay, C., Isingrini, M., Clarys, D., Taconnat, L., & Eustache, F. (2004). Executive functioning and judgment-of-learning versus feeling-of-knowing in older adults. *Experimental Aging Research, 30,* 47–62.

Souchay, C., Moulin, C. J. A., Clarys, D., Taconnat, L., & Isingrini, M. (2007). Diminished episodic memory awareness in older adults: Evidence from feeling-of-knowing and recollection. *Consciousness and Cognition, 16,* 769–784.

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2008). Measuring memory and metamemory: Theoretical and statistical problems with assessing learning (in general) and using gamma (in particular) to do so. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 95–116). New York: Psychology Press.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false-alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics, 68,* 643–654.

Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 739–746.

Verde, M. F., & Rotello, C. M. (2004). ROC curves show that the revelation effect is not a single phenomenon. *Psychonomic Bulletin & Review, 11,* 560–566.

Weaver, C. A., III, & Kelemen, W. L. (2003). Processing similarity does not improve metamemory: Evidence against transfer-appropriate monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 1058–1065.

Wilson, T. P. (1974). Measures of association for bivariate ordinal hypotheses. In H. M. Blalock (Ed.), *Measurement in the social sciences* (pp. 327–342). Chicago: Aldine.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11,* 616–641.

Woods, C. M. (2007). Confidence intervals for gamma-family measures of ordinal association. *Psychological Methods, 12,* 185–204.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75,* 579–652.

Appendix A

Methods for Estimating Ratio of Noise and Signal Distribution Standard Deviations

The assumption of equal-variance signal and noise distributions may be violated, as Ratcliff, Sheu, and Gronlund (1992) have shown for recognition memory tasks. In such cases, a variant of d', d_a, can be computed as shown in Equation 2 (Simpson & Fitter, 1973). To compute d_a, it is necessary to estimate s, the size of the standard deviation of the noise distribution relative to the standard deviation of the signal distribution (i.e., the ratio of the standard deviations). This can be done in a number of ways. First, s may be determined by the nature of artificially constructed stimuli (e.g., Ashby & Gott, 1988). Alternatively, s can be estimated empirically by constructing a receiver operating characteristic (ROC) based on pairs of H and F values computed from rating responses (e.g., confidence ratings for "old" and "new" responses in a recognition memory task). The ROC is then replotted on binormal probability coordinates to create a *z*ROC (see Figure A1). If the noise and signal distributions are Gaussian, then the *z*ROC will form a straight line with slope equal to s (Swets, 1986a). Both d_a and s (as well as other characteristics of the ROC) can easily be computed within SYSTAT's SDT module; Macmillan and Creelman (2005, Chapter 3) describe the necessary commands.

An alternative to d_a may also be computed, which is an estimate of the area under the ROC (Swets & Pickett, 1982):

$$A_z = \Phi\left(\frac{d_a}{\sqrt{2}}\right), \tag{A1}$$

where Φ is the cumulative normal distribution function. A_z may be estimated with SYSTAT or a convenient Web-based program (Eng, n.d.) available at http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html. A_z has been shown to have slightly better statistical properties than d_a (Macmillan et al., 2004).

Appendix B

Derivation of G_c

Consider a population of nonstudied items and a population of studied items, both of which form rectangular distributions of equal variance, as shown in Figure B1. The response criterion for old/new decisions is indicated by a vertical line. The four indicated regions correspond to the response cells depicted in Table 1, in which, for example, a is the number of studied items correctly classified as old. Note that region a extends from the rightmost border of the distribution leftward to the response criterion and therefore overlaps with all of Region b. Similarly, Region d (correct rejections) overlaps with all of region c (misses).

In a forced-choice design, each trial typically presents one studied item and one nonstudied item, and the subject must choose

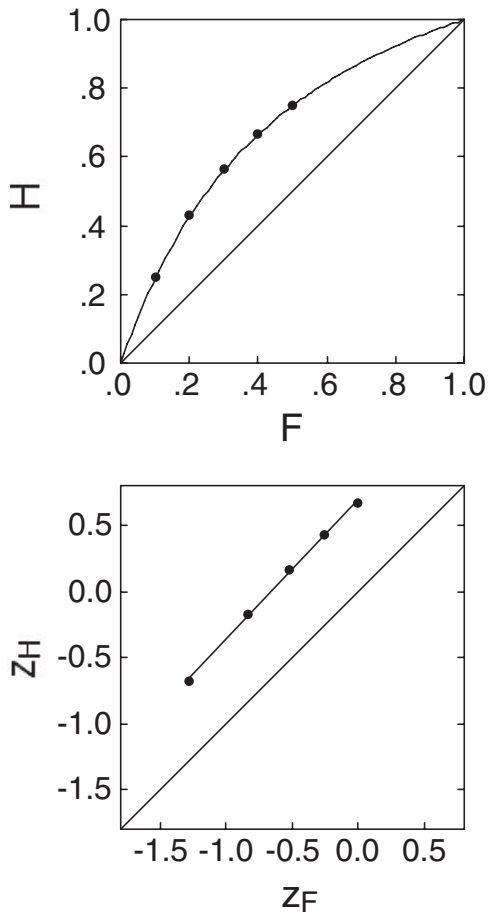

Figure A1. Receiver operating characteristic (ROC) based on hypothetical data from a recognition memory experiment (upper panel) and the same data plotted on binormal coordinates to form a *z*ROC (lower panel). H = hit rate; F = false-alarm rate.

which was presented previously. Thus, items from Region *a* could be paired with items from region *d*, which would lead to accurate responding because the regions do not overlap in strength and fall on opposite sides of the decision criterion. The probability of this pairing of test probes is $\left(\frac{a}{a+c}\right)\left(\frac{d}{b+d}\right)$, the product of their respective sampling probabilities, and the probability of a correct response for those probes is 1.0 (assuming a noise-free system). Alternatively, items from region *b* could be paired with items from region *c*, which would result in inaccurate responding because the regions fall on opposite sides of the decision criterion and their relationships to that criterion are reversed (e.g., the studied item falls in the "new" response region). Thus, the probability of this pairing of test probes is $\left(\frac{c}{a+c}\right)\left(\frac{b}{b+d}\right)$, where all such pairings lead to erroneous responses.

It is also possible that a studied–nonstudied test pair might consist of items from region *b* (for the nonstudied item) and the region of the studied distribution that entirely overlaps with *b*. The probability of such a pair is $\left(\frac{b}{b+d}\right)\left(\frac{b}{a+c}\right)$. For such pairs, it is equally likely that the studied item or the nonstudied item will have greater strength, so choosing the stronger item will lead to a correct response for 50% of these pairs. Analogously, the studied–nonstudied pair might consist of items drawn from region *c* (from the studied distribution) and the region of the nonstudied distribution that entirely overlaps with *c*. The probability of such a pair is $\left(\frac{c}{a+c}\right)\left(\frac{c}{b+d}\right)$, and the expected accuracy rate is again 50%.

Finally, consider two remaining types of studied–nonstudied pairs that could be sampled: Those in which one member of the pair falls in (*a*–*b*) and the other falls in *b* and those in which one member falls in region (*d*–*c*) and the other falls in *c*. These pairs could be answered accurately because in all cases the studied member of the pair has greater strength than the nonstudied member; these trials have probabilities of $\left(\frac{a-b}{a+c}\right)\left(\frac{b}{b+d}\right)$ and $\left(\frac{c}{a+c}\right)\left(\frac{d-c}{b+d}\right)$, respectively.

Taking all of these cases together, the probability of a correct response is

$$p(\text{correct}) = \frac{ad + \frac{1}{2}b^2 + \frac{1}{2}c^2 + (a-b)b + (d-c)c}{(a+c)(b+d)} \quad \text{(B1)}$$

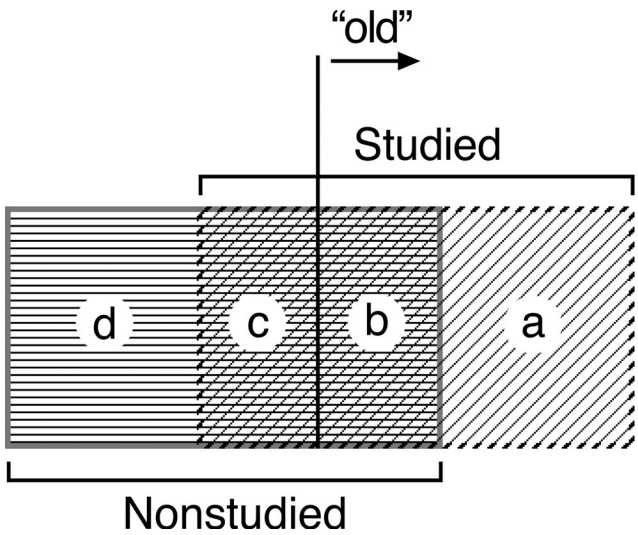

Figure B1. Hypothetical rectangular distributions of memory strength. The distribution of studied items, outlined with dashed lines, consists of Regions a (diagonal lines) and c (cross-hatched); the distribution of nonstudied items, outlined with gray lines, consists of regions d (horizontal lines) and b (cross-hatched). Note that Region a overlaps with all of Region b, and that Region d overlaps with all of Region c. The vertical line separating Regions b and c is the response criterion; items lying above the criterion generate "old" responses.

(*Appendixes continue*)

and the probability of an error is

$$p(\text{error}) = \frac{bc + \frac{1}{2}b^2 + \frac{1}{2}c^2}{(a + c)(b + d)} \quad \text{(B2)}$$

Nelson (1984) showed that $G = 2V - 1$, where $V = \frac{ad}{ad + bc}$. This calculation considers the probabilities of only some of the possible correct (concordant) and incorrect (discordant) outcomes, namely, *ad* and *bc*, respectively. Substituting Equation B1 for the probability of a correct decision and Equation B2 for the probability of an incorrect decision yields the following equation for *V*, which we term *corrected V* to minimize confusion:

$$\text{Corrected } V = \frac{ad + \frac{1}{2}b^2 + (a - b)b + \frac{1}{2}c^2 + (d - c)c}{ad + \frac{1}{2}b^2 + (a - b)b + \frac{1}{2}c^2 + (d - c)c + bc + \frac{1}{2}b^2 + \frac{1}{2}c^2} \quad \text{(B3)}$$

Given Table 1 or Figure B1, the hit, miss, false-alarm, and correct rejection rates from an old–new design can be calculated as follows:

$$H = \frac{a}{a + c} \quad \text{(B4)}$$

$$1 - H = \frac{c}{a + c} \quad \text{(B5)}$$

$$F = \frac{b}{b + d} \quad \text{(B6)}$$

$$1 - F = \frac{d}{b + d} \quad \text{(B7)}$$

Rearranging terms in Equations B4–B7 reveals that $a = H(a + c)$, $b = F(b + d)$, $c = (1-H)(a + c)$, and $d = (1 - F)(b + d)$. Substituting these values into Equation B3 and simplifying the result yields

$$\text{Corrected } V = 1 + HF - F - \frac{F^2(b + d)}{2(a + c)} - \frac{(1 - H)^2(a + c)}{2(b + d)} \quad \text{(B8)}$$

If the number of studied items ($a + c$) equals the number of nonstudied items on the test ($b + d$), then Equation B6 may be further reduced:

$$\text{Corrected } V = 1 + HF - F - \frac{1}{2}F^2 - \frac{1}{2}(1 - H)^2 \quad \text{(B9)}$$

It should be noted that this correction makes three key assumptions: (a) the number of studied and nonstudied items on the test are equal, (b) the underlying distributions are rectangular, and (c) the distributions have equal variance. Equation B8 may be used if the first assumption is violated. If either the second or the third assumption is violated, then it is no longer the case that the probability of a correct response is 50% when both test items are drawn from Regions *b* or *c*. Instead, the probability would depend on the exact form of the probability density functions for the studied and nonstudied distributions within those regions.

Finally, calculation of *G* proceeds as in Equation 4, substituting *corrected V* for *V*. A bit of algebra shows that, under the three assumptions laid out above,

$$G_c = 2(H–F)–(H–F)^2. \quad \text{(B10)}$$

This correction works very well for low false-alarm rates at any level of true sensitivity and for moderate or even high false-alarm rates when true sensitivity is low (see Figure 4). As true sensitivity rises, however, the observable values of $H - F$ are constrained: As the criterion becomes more liberal, *H* reaches 1 and cannot increase any more, but F continues to increase until it reaches 1. The calculated value of G_c changes in that case because $H - F$ changes, even though the distributions themselves are unchanged.

Appendix C

Results of Simulation 3 for $\gamma_d = .4$ and $\gamma_d = .8$

Number of trials	*F*	*G*		Log-linear *G*		Ratings *G*		*G**		Log-linear *G**		Ratings *G**	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
						$\gamma_d = .4$, equal variance							
16	.05	.61	.46	.47	.35	.50	.27	0.36	0.34	0.53	0.43	0.52	0.33
	.10	.56	.36	.50	.31	.47	.24	0.50	0.36	0.56	0.41	0.49	0.29
	.30	.51	.27	.49	.26	.45	.21	0.56	0.35	0.52	0.33	0.45	0.24
	.50	.52	.28	.49	.27	.45	.21	0.56	0.36	0.53	0.35	0.45	0.25
64	.05	.61	.21	.58	.20	.51	.13	0.66	0.29	0.63	0.30	0.50	0.16
	.10	.57	.17	.56	.16	.48	.12	0.60	0.24	0.58	0.22	0.46	0.13
	.30	.53	.13	.52	.13	.45	.10	0.53	0.17	0.52	0.16	0.43	0.11
	.50	.54	.14	.53	.14	.45	.10	0.54	0.17	0.53	0.17	0.43	0.12
256	.05	.62	.10	.61	.10	.51	.07	0.64	0.15	0.63	0.15	0.49	0.08
	.10	.58	.08	.57	.08	.48	.06	0.58	0.11	0.57	0.11	0.46	0.07
	.30	.53	.07	.53	.07	.45	.05	0.52	0.08	0.52	0.08	0.42	0.06
	.50	.54	.07	.54	.07	.46	.05	0.53	0.09	0.53	0.09	0.43	0.06

(*table continues*)

Table (*continued*)

Number of trials	*F*	*G*		Log-linear *G*		Ratings *G*		G^*		Log-linear G^*		Ratings G^*	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
						γ_d = .8, equal variance							
16	.05	.92	.11	.87	.11	.86	.10	1.17	0.31	1.30	0.39	1.22	0.37
	.10	.90	.10	.87	.10	.85	.10	1.31	0.36	1.31	0.41	1.18	0.35
	.30	.90	.10	.87	.10	.85	.10	1.31	0.36	1.31	0.41	1.18	0.36
	.50	.92	.12	.86	.12	.87	.10	1.09	0.28	1.27	0.37	1.23	0.37
64	.05	.92	.05	.91	.05	.86	.05	1.44	0.27	1.39	0.28	1.16	0.17
	.10	.91	.04	.90	.05	.85	.05	1.37	0.24	1.33	0.22	1.11	0.16
	.30	.91	.04	.90	.05	.85	.05	1.37	0.24	1.33	0.22	1.12	0.16
	.50	.93	.05	.91	.05	.87	.05	1.46	0.27	1.44	0.32	1.19	0.19
256	.05	.92	.02	.92	.02	.86	.02	1.42	0.14	1.40	0.14	1.14	0.08
	.10	.91	.02	.91	.02	.85	.02	1.34	0.11	1.33	0.11	1.10	0.08
	.30	.91	.02	.91	.02	.85	.02	1.34	0.11	1.33	0.11	1.10	0.08
	.50	.93	.03	.93	.03	.87	.02	1.47	0.17	1.44	0.16	1.17	0.09
						γ_d = .4, unequal variance (*SD* ratio = 0.6)							
16	.05	.81	.24	.72	.21	.64	.21	0.76	0.31	0.92	0.40	0.71	0.32
	.10	.73	.23	.68	.22	.57	.21	0.79	0.34	0.83	0.40	0.61	0.29
	.30	.56	.25	.53	.25	.47	.21	0.62	0.36	0.58	0.33	0.47	0.25
	.50	.44	.30	.42	.28	.42	.22	0.47	0.36	0.43	0.33	0.42	0.26
64	.05	.82	.10	.80	.10	.65	.10	1.06	0.28	1.02	0.29	0.68	0.16
	.10	.75	.11	.73	.11	.57	.10	0.88	0.23	0.84	0.22	0.58	0.13
	.30	.57	.13	.57	.12	.47	.10	0.58	0.17	0.57	0.16	0.45	0.12
	.50	.45	.15	.45	.15	.42	.11	0.44	0.17	0.43	0.17	0.40	0.12
256	.05	.82	.05	.82	.05	.65	.05	1.03	0.15	1.02	0.14	0.67	0.08
	.10	.75	.05	.75	.05	.58	.05	0.85	0.11	0.85	0.11	0.57	0.07
	.30	.58	.06	.58	.06	.47	.05	0.58	0.08	0.57	0.08	0.44	0.06
	.50	.46	.07	.46	.07	.42	.06	0.43	0.08	0.43	0.08	0.39	0.06
						γ_d = .8, unequal variance (*SD* ratio = 0.6)							
16	.05	.95	.07	.92	.08	.89	.09	1.43	0.33	1.54	0.40	1.33	0.37
	.10	.93	.08	.90	.08	.87	.09	1.46	0.37	1.46	0.43	1.24	0.35
	.30	.88	.11	.85	.12	.84	.11	1.25	0.37	1.23	0.41	1.15	0.37
	.50	.85	.16	.80	.16	.83	.13	1.00	0.32	1.09	0.40	1.10	0.37
64	.05	.95	.03	.95	.03	.89	.04	1.69	0.28	1.64	0.29	1.28	0.18
	.10	.93	.03	.93	.03	.87	.04	1.52	0.24	1.47	0.23	1.18	0.16
	.30	.89	.05	.88	.05	.84	.05	1.27	0.22	1.24	0.21	1.08	0.16
	.50	.86	.07	.85	.07	.83	.06	1.19	0.26	1.14	0.26	1.07	0.19
256	.05	.96	.02	.95	.02	.89	.02	1.66	0.15	1.64	0.14	1.26	0.08
	.10	.93	.02	.93	.02	.87	.02	1.48	0.11	1.47	0.11	1.16	0.08
	.30	.89	.02	.89	.03	.84	.03	1.25	0.11	1.24	0.10	1.07	0.08
	.50	.87	.04	.86	.04	.83	.03	1.15	0.13	1.14	0.12	1.05	0.09

Note. $F = p(\text{JOL} \geq 50 \mid \text{forgotten})$, where JOL (judgment of learning) is rated on a 100-point scale. G = the standard Goodman–Kruskal gamma coefficient; G^* = a modified measure of gamma introduced by Benjamin and Diaz (2008).

Received July 16, 2008
Revision received October 31, 2008
Accepted November 7, 2008 ■