Behavior Research Methods
2005, 37 (1), 3-10

Efficiently measuring recognition performance with sparse data

LAEL J. SCHOOLER
Max Planck Institute for Human Development, Berlin, Germany

and

RICHARD M. SHIFFRIN
Indiana University, Bloomington, Indiana

We examine methods for measuring performance in signal-detection-like tasks when each participant provides only a few observations. Monte Carlo simulations demonstrate that standard statistical techniques applied to a d' analysis can lead to large numbers of Type I errors (incorrectly rejecting a hypothesis of no difference). Various statistical methods were compared in terms of their Type I and Type II error (incorrectly accepting a hypothesis of no difference) rates. Our conclusions are the same whether these two types of errors are weighted equally or Type I errors are weighted more heavily. The most promising method is to combine an aggregate d' measure with a percentile bootstrap confidence interval, a computer-intensive nonparametric method of statistical inference. Researchers who prefer statistical techniques more commonly used in psychology, such as a repeated measures t test, should use γ (Goodman & Kruskal, 1954), since it performs slightly better than or nearly as well as d'. In general, when repeated measures t tests are used, γ is more conservative than d': It makes more Type II errors, but its Type I error rate tends to be much closer to that of the traditional .05 α level. It is somewhat surprising that γ performs as well as it does, given that the simulations that generated the hypothetical data conformed completely to the d' model. Analyses in which H − FA was used had the highest Type I error rates. Detailed simulation results can be downloaded from www.psychonomic.org/archive/Schooler-BRM-2004.zip.

In this article, we examine methods for measuring performance in signal-detection-like tasks when each participant provides only a few observations. As an example, we couch our discussion in the context of memory recognition experiments, but the points that will be made apply quite generally.

Our results may be of particular relevance to those using the Deese/Roediger–McDermott (DRM) false memory paradigm (Deese, 1959; Read, 1996; Roediger & McDermott, 1995), because such studies often involve small numbers of observations per participant. In a typical experiment, lists of words are presented to participants. The words on these lists (e.g., BED, REST, AWAKE) are all associated to a generator word (e.g., SLEEP), but usually do not include it. When participants are asked to recall the list, they produce the generator at surprisingly high rates. In some versions of the task, participants are asked to make recognition judgments about whether words are old or new—that is, whether the words were or were not on the list. The responses are typically scored as depicted in Table 1. Participants can be correct by saying that an old word is old (a hit) or that a new word is new (a correct rejection). They can make an error by saying that an old word is new (a miss) or that a new word is old (a false alarm).

In such experiments, the hit and false alarm rates for generator words can be excessively high. For instance, Miller and Wolford (1999) observed an average hit rate of .97 for generator words. The problem is that such high false alarm and hit rates lead to undefined d' values, especially when there are few observations per subject. Even with more modest hit and false alarm rates of .89 and .88, Gallo, Roediger, and McDermott (2001) found that "10 of the 24 subjects had both critical hit and false alarm rates of 100%" (p. 584). We use Monte Carlo simulations to explore how best to analyze experiments such as these.

d' ANALYSIS

In a typical recognition experiment, participants are asked to study stimuli (e.g., words), and after some retention interval their memory for the stimuli is tested. The test consists of a mixture of *old*, studied words and *new*, unstudied words. A d' analysis can be applied to situations such as this, when a participant has to distinguish between

This work was begun while L.S. was supported by NSRA Fellowship 1F32HD/MHC7787-01A1 at Indiana University. We thank Joan Snodgrass, George Wolford, and several anonymous reviewers for their comments on the manuscript. Correspondence concerning this article should be addressed to L. Schooler, Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, Berlin 14195, Germany (e-mail: schooler@mpib-berlin.mpg.de).

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

Table 1
The Different Ways in Which a Subject Can Be Correct or Incorrect

	Trial Type	
Response	Old	New
"Old"	hit	false alarm
"New"	miss	correct rejection

a signal embedded in noise and noise alone. In the case of a recognition experiment, the embedded signal and noise alone would be evidence in favor of old and new items, respectively. *Sensitivity* refers to the ability to detect a signal—that is, to distinguish old and new words. Sensitivity alone, however, inadequately captures performance. Two participants with equal sensitivity may display different patterns of results because they have different response biases. Some participants might conservatively respond "old" only when they are quite certain they have studied a word, whereas others might be prone to respond "old" even when they are unsure. The conservative group would tend to have fewer hits (more misses) but more correct rejections (fewer false alarms) than the second, more liberal group.

Researchers commonly use a d' analysis to get a measure of participant sensitivity that is independent of response bias. A d' analysis, originally based on the theory of signal detection (Green & Swets, 1966), rests on the assumption that each stimulus results in a single numerical value that is used as evidence. If this value exceeds a criterion, C, participants respond "signal," and if it falls below this point they respond "noise." Where the criterion is set determines the response bias; extremely low criteria mean that participants are prone to say "signal" on the basis of meager evidence.

The evidence in favor of the old and new words is assumed to be normally distributed with shared variance and means of μ_{old} and μ_{new}, respectively Examples of old and new distributions are plotted in Figure 1. In this example, the means of the new and old distributions are arbitrarily set to 0 and 2, respectively, with a common standard deviation (SD) of 1 and a response criterion (C) of 1.3. A d' analysis estimates the distance between μ_{old} and μ_{new}, measured in SDs, at 2 in Figure 1. This distance and placement of the criterion are unknown but can be determined from the hit and false alarm rates. In Figure 1, the response criterion is .7 SDs below μ_{old}, so $Z_H = -.7$, and it is 1.3 SDs above μ_{new}, so $Z_{FA} = 1.3$, yielding a d' of 2 [i.e., $d' = Z_{FA} - Z_H$; $2 = 1.3 - (-.7)$]. Z_H and Z_{FA} can be estimated from the Z scores corresponding to the observed hit and false alarm rates. In the example, a false alarm rate of .0968 (the area under the new item distribution to the right of the criterion) corresponds to the Z_{FA} value of 1.3, and a hit rate of .75804 (the area under the old item distribution to the right of the criterion) corresponds to the Z_H value of $-.7$. Figure 1 illustrates how the data can be described fully by reporting just the proportion of hits and false alarms, because the criterion divides the old distribution such that P(hit) + P(miss) = 1, and the new distribution such that P(false alarm) + P(correct rejection) = 1.

A problem with using d' as a performance measure is that it is undefined when either the hit rate or the false alarm rate equals 1 or 0, because these probabilities correspond to infinite values of Z_H and Z_{FA}.

TRANSFORMING RAW HIT AND FA RATES FOR A d' ANALYSIS

There are some standard data transformations that allow a d' analysis to be carried out even when the observed hit or false alarm rates equal 1 or 0. The simplest method is to replace empirical hit and false alarm rates of 0 and 1 with fixed values of, say, .99 and .01. A second method replaces 1 with $1 - \frac{1}{2n}$ and 0 with $\frac{1}{2n}$, where n is the number of trials. This captures the intuition that the greater the number of trials, the greater one's confidence that participants' true hit or false alarm rates are close to 1 or to 0. Snodgrass and Corwin (1988) advocated adding 0.5 to the number of responses and 1 to the number of trials. This turns out to be a Bayesian estimate of hit and false alarm rates that results from using a $\beta(.5,.5)$ distribution as the initial prior distribution. Thus, the expected hit rate is expressed as

$$H = \frac{\#\text{"old"} + .5}{\#\text{old trials} + 1}, \quad (1A)$$

where #old trials is the number of trials on which previously seen stimuli were tested and #"old" is the number of these old trials that the participant said were old. Similarly, the expected value for the false alarm rate is expressed as

$$FA = \frac{\#\text{"old"} + .5}{\#\text{new trials} + 1}, \quad (1B)$$

where #new trials is the number of trials on which previously unseen stimuli were tested and #"old" is the number of these new trials that the participant said were old. For the sake of convenience, we term these the *Snodgrass–Corwin correction*. In the subsequent analyses, we focus on this correction because preliminary simulations showed that it slightly outperformed the other corrections and is grounded in Bayesian estimation procedures.

ALTERNATIVE MEASURES OF SENSITIVITY

The problems of conducting a d' analysis when hit and false alarm rates equal 1 or 0 stem from its parametric assumptions. It may be profitable to consider alternative measures that make different assumptions about the processes responsible for hits and false alarms.

Hits − FAs

Perhaps the simplest measure is to use hits minus false alarms (H − FA), which is the appropriate dependent measure if the data are generated according to a double high-threshold model (see Snodgrass & Corwin, 1988, for an extended analysis). In this model, participants

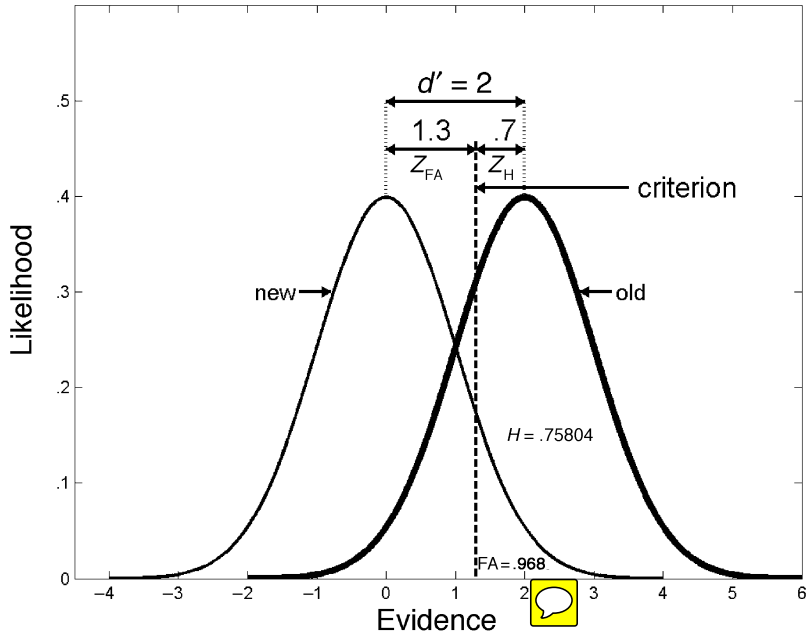

Figure 1. An illustration of the distributions that underlie a *d′* analysis. H, hit; FA, false alarm

perceive stimuli presented for recognition in one of three mutually exclusive ways: They can be certain that a stimulus is new, certain that it is old, or unsure whether it is new or old. In addition, the probability that participants will recognize old items as old is assumed to equal the probability that new items will be recognized as new. This shared probability, *R*, is taken to be a measure of participant sensitivity. False alarms and misses result purely from bias and occur only in response to stimuli about which participants are uncertain. Thus, the probability of a hit is

$$\mathrm{H} = R + (1 - R)B, \tag{2}$$

where *B* is the probability of a participant's saying that an item in the unknown state is old.

Only new stimuli that were not recognized as new can generate false alarms. Therefore, the probability of a false alarm is

$$\mathrm{FA} = (1 - R)B. \tag{3}$$

Substituting Equation 3 into Equation 2 yields an estimate of the sensitivity measure, *R*:

$$R = \mathrm{H} - \mathrm{FA}. \tag{4}$$

Goodman and Kruskal's (1954) γ

Nelson (1984) advocated a probabilistic definition of participant sensitivity, which was developed to analyze feeling-of-knowing experiments but applies to recognition in general. One measure of participant sensitivity, *V*, can be defined as

P(participant's saying A was studied more than B|A was studied more than B).

Nelson (1984), having chosen *V* as the best measure of participant sensitivity, shows that Goodman and Kruskal's (1954) γ is the best dependent measure.

$$\gamma = \frac{\mathrm{H} - \mathrm{FA}}{\mathrm{H} + \mathrm{FA} - 2\mathrm{HFA}} = \frac{(\#\mathrm{H} \times \#\mathrm{CR}) - (\#\mathrm{FA} \times \#\mathrm{M})}{(\#\mathrm{H} \times \#\mathrm{CR}) + (\#\mathrm{FA} \times \#\mathrm{M})}, \tag{5}$$

where #H is number of hits. Nelson contends that analyses based on γ are superior to those based on *d′* scores or on H − FA, because γ has a 1:1 linear relation with *V* that these other measures lack. γ shares with *d′* the limitation that certain combinations of hit and FA rates leave it undefined. This can be seen most easily by looking at the raw data version of γ, which has the expression (#H × #CR) + (#FA × #M) in the denominator. When there are no hits or no correct rejections, and no falese alarms or no misses, the definition of γ calls for division by 0. To avoid division by 0 in the subsequent simulations, we will apply the Snodgrass–Corwin correction (adding 0.5 to the numerator and 1 to the denominator) to each hit rate and each false alarm rate before calculating γ.

BOOTSTRAP HYPOTHESIS TESTING

One method used to get around the distortions introduced by the truncation transforms is to aggregate the hits and false alarms of all the participants and calculate aggregate *d′* scores. A problem with aggregate *d′* scores is that it is not clear how to test the reliability of differences between conditions; next, we propose a method to do this.

Bootstrapping (Efron & Tibshirani, 1991; Mooney & Duval, 1993) is a technique well suited to such problems. A bootstrap analysis relies on a large number of resam-

plings from the data to estimate variability in an arbitrary statistic. Given data vector $\boldsymbol{x}$ (e.g., the participants' responses in a recognition experiment) of size n (e.g., the number of participants), bootstrapping involves generating a new vector, $\boldsymbol{x}_i$, by sampling n times with replacement from $\boldsymbol{x}$, calculating some statistic, $s(\boldsymbol{x}_i)$, on the basis of this sample and repeating the process b times. Hypotheses about the true value of $s(\boldsymbol{X})$ (the population statistic) can be tested by looking at the distribution of the b values of $s(\boldsymbol{x}_i)$. A 95% confidence interval around $s(\boldsymbol{x})$ can be obtained by sorting the b $s(\boldsymbol{x}_i)$ values and finding the values of $s(\boldsymbol{x}_i)$ associated with the 2.5 and 97.5 percentiles. This *percentile method*—perhaps the most straightforward method for constructing a bootstrap confidence interval—requires b values of 1,000 or more. However, when $s(\boldsymbol{x})$—the value of the statistic of interest applied to the original sample—is not close to the median of the bootstrap distribution, this can result in the misplacement of end points of the confidence interval. There are various methods for adjusting the end points of these confidence intervals. One such method, the *bias-corrected and accelerated confidence interval* (BC_a), operates under the assumption that there exists a transformation that brings $s(\boldsymbol{x})$ closer to the median (Martinez & Martinez, 2002). One strength of BC_a is that it does not require knowledge of what this transformation is, but only that it exists. In those cases in which the bootstrap distribution is centered about $s(\boldsymbol{x})$, it reduces to the percentile method.

Our goal in this section was to provide not a complete introduction to bootstrapping, but simply a feel for how the method works, since several excellent introductions already exist (e.g., Mooney & Duval, 1993). Both the article by Zoubir and Boashash (1998) and the book by Martinez and Martinez (2002) are particularly useful, because they have freely available MATLAB functions and examples of how to use the various techniques.

SIMULATIONS

We call the combination of a statistic (e.g., t test) and a measure (e.g., d') a *method*. Here, we use Monte Carlo simulations to explore which, if any, of nine methods (t test, percentile, and BC_a each crossed with d', H − FA, and γ) is best to use when participants provide only a few observations per condition, assuming that the true state of the world is that described by the d' model. It seems likely that the best methods will be those suited to the process that generates the data. Nonetheless, we assume throughout this discussion that the hit and false alarm rates arise from the d' model we have described: equal variances and normal distributions for signal and noise, with participants responding "old" whenever a sample from the appropriate distribution exceeds a criterion chosen by the participant for that condition.

The basic procedure was to generate a population of simulated participants that followed the d' model. These participants were run in two experimental conditions, producing raw data in the form of the proportion of hits and false alarms, which were then analyzed according to each of the nine methods. By varying whether or not the conditions were in fact different, the methods were compared on the basis of the proportions of Type I errors (incorrect rejection of the null hypothesis when it was true) and Type II errors (failure to reject the null hypothesis when it was false). That is, we tracked how often the measures rejected the null hypothesis when the conditions did and did not differ in terms of d'. We next describe these simulations in detail.

Method

Each simulated participant's performance depends on its general aptitude, the characteristics of the condition, and chance.

Simulated participants vary in terms of aptitude. Simulated participants were endowed with an aptitude measured in d'. To model aspects of participant variability, aptitudes were drawn from a normal distribution with a mean of 1.5 and an *SD* of 1.

Simulated conditions vary in terms of d' and bias. The difficulty of a condition for a particular participant was determined by drawing a d' value from a normal distribution with the mean and the *SD* (always equal to 1) associated with that condition, and adding this d' value to the participant's aptitude value.

The conditions varied in terms of how they biased the participants' responses. Such between-conditions (e.g., generator word vs. generated word) criterion shifts have been central to the brisk debate about how best to interpret the underlying cause of the high hit and FA rates for generator words in the DRM paradigm. Miller and Wolford (1999) argued that these rates do not reflect induced false memories but simply result from the use of different criteria for the different kinds of words. Wixted and Stretch (2000) countered that such item-by-item criterion shifts represent more mental effort than the average participant is willing to expend, but their own data demonstrate that participants are willing to shift criteria between lists. Stretch and Wixted (1998) did not find evidence that participants adjust criteria on an item-by-item basis, whereas others did. For instance, Heit, Brockdorff, and Lamberts (2003) demonstrated that in a signal-to-respond recognition experiment, participants adapted their response criteria on a trial-by-trial basis.

To model experiments such as these, we include conditions with neutral, conservative, and liberal response criteria. Thus, we assume that some conditions are unbiasing, some lead participants to say "old," and others lead participants to say "new." In each condition, μ_{new} was arbitrarily set to 0, so setting a criterion halfway between 0 and the participants' overall competence in that condition leads to unbiased performance. Biased conditions were created by setting the response criterion 1 *SD* below or above the unbiased criterion. No random noise was associated with the bias.

Table 2 shows the six conditions that were used. The leftmost column contains a reference number for each condition. Columns 2–5 describe the d' model underlying participant performance in each condition. The total d' of the condition results from random draws from two normal distributions with means μ_p and μ_c, which are the means of the normal distributions that describe the participants' aptitudes and the difficulty of that condition, respectively. Thus, a participant's competence in a particular condition was obtained by adding its aptitude to the difficulty of that condition. For example, if a participant's aptitude was 1.3 and the difficulties drawn from the distributions for Conditions A and B were 2.2 and 1.3, respectively, then its overall competence would be 3.5 in Condition A and 2.6 in Condition B. This two-stage sampling scheme captures correlations in performance due to participant differences, as is seen in the finding that a participant that does well in one condition tends to do well in another.

Table 2
Condition Specifications

	Underlying d' Model				Performance		Proportion of Participants Transformed			
Condition	μ_p	μ_c	$\mu_p + \mu_c$	Bias	H	FA	$P(H) = 1$	$P(H) = 0$	$P(FA) = 1$	$P(FA) = 0$
1	1.5	1	2.5	−1	.988	.401	.93	.00	.14	.28
2	1.5	1	2.5	0	.894	.106	.67	.02	.02	.65
3	1.5	1	2.5	1	.599	.012	.29	.14	.00	.90
4	1.5	2	3.5	−1	.997	.227	.98	.00	.06	.46
5	1.5	2	3.5	0	.960	.040	.82	.00	.00	.79
6	1.5	2	3.5	1	.773	.003	.48	.06	.00	.94

Note—Six conditions were used to generate simulated participants. A description of the d' models underlying participant performance in each condition and the mean levels of performance are provided. The last four columns show, from left to right, the proportions of pseudoparticipants in the three-trials-per-condition comparisons that always hit, never hit, always false alarmed, and never false alarmed. H, hit; FA, false alarm.

The fifth column represents the bias of the condition. Columns 6 and 7 show the mean hit and FA rates. Columns 8–11 present the proportion of participants that had observed hit and false alarm rates of 0 and 1.

Participants' performance. A participant's performance, defined in terms of expected hit and FA rates, was a random function of the characteristics of the condition and the participant's aptitude. The probability of the participant's responding "old" or "new" was a random function of these expected hit and false alarm rates. For example, if a participant's expected hit rate in a particular condition was, say, .90, then for each of the three old trials a random number was drawn from uniform distributions between 0 and 1. If the number was less than .90, then it was counted as a hit; otherwise, it was scored as a miss.

Running a simulated experiment. There were 36 participants in each experiment, generated according to the sampling scheme described above, that were run in two conditions. In one set of simulations, each participant saw 3 old trials and 3 new trials in each of the two conditions; in a second set, the participants saw 10 trials of each kind; and in a third, they saw 20 trials of each kind. These numbers were chosen because they are thought to be illustrative of situations in which each participant provides only a small number of observations in each condition. In the Monte Carlo simulations, each condition was paired with all of the conditions, including itself, yielding 21 comparisons.

Analyzing a simulated experiment. Each of the nine methods was applied in turn to every simulated experiment. For the methods involving t tests, hit and false alarm rates were calculated according to the Snodgrass–Corwin correction for each participant. These rates were then used to calculate the sensitivity measures d', γ, and H − FA. These measures were analyzed with a repeated measures t test; with the null hypothesis (H_0) that the participants' performance, measured in terms of sensitivity, was the same in the two conditions, and with an α level of .05.

For methods involving percentile and BC_a confidence intervals, we define $s(\boldsymbol{x})$ to be $f(\boldsymbol{x}_\mathbf{A}) - f(\boldsymbol{x}_\mathbf{B})$, where $\boldsymbol{x}_\mathbf{A}$ and $\boldsymbol{x}_\mathbf{B}$ are participants' data for Conditions A and B, respectively, and f is a measure of sensitivity (i.e., d', H − FA, or γ). If there are no sensitivity differences between Conditions A and B, then the 95% confidence interval of $f(\boldsymbol{x}_\mathbf{A}) - f(\boldsymbol{x}_\mathbf{B})$, estimated from the bootstrap distribution, should include 0.

For each bootstrap sample, aggregate hit and false alarm rates were calculated for the sample according to the Snodgrass–Corwin transformation before $s(\boldsymbol{x})$ was applied. Such transformations are sometimes necessary, even when aggregated data are used, because d' (and γ) can be undefined. To see why this is the case, consider a condition with a hit rate of say, .97, such as that observed by Miller and Wolford (1999). In experiments with 36 participants, each of whom saw three trials, one would expect that in 3.7% of such experiments the aggregated d' would be undefined, because all the participants would have perfect performance (e.g., $.97^{\{36 * 3\}}$).

Results

The on-line appendix at www.psychonomic.org/archive/Schooler-BRM-2005.zip presents the results of 63 experimental comparisons (21 pairings of the 6 conditions × 3, 10, and 20 trials per condition). Table 3 provides examples of three such experimental comparisons

Table 3
Example Condition Comparisons

C2-Versus-C2 Comparison ($d' = 2.5$; Bias = 0)				
Statistic	Measure	C2 > C2	C2 > C2	Null
t test	d'	.014	.023	.963
t test	γ	.022	.022	.956
t test	H − FA	.015	.026	.959
pctl	d'	.021	.033	.946
pctl	γ	.020	.033	.947
pctl	H − FA	.019	.030	.951
BC_a	d'	.019	.027	.954
BC_a	γ	.040	.048	.913
BC_a	H − FA	.021	.035	.944
C2-Versus-C1 Comparison (For C1, $d' = 2.5$, Bias = −1; for C2, $d' = 2.5$, Bias = 0)				
Statistic	Measure	C2 > C1	C1 > C2	Null
t test	d'	.395	0	.605
t test	γ	.206	.002	.792
t test	H − FA	.416	0	.584
pctl	d'	.029	.050	.921
pctl	γ	.009	.099	.893
pctl	H − FA	.382	.048	.570
BC_a	d'	.027	.045	.928
BC_a	γ	.025	.127	.849
BC_a	H − FA	.350	.047	.603
C5-Versus-C2 Comparison (For C2, $d' = 2.5$, Bias = 0; for C5, $d' = 3.5$, Bias = 0)				
Statistic	Measure	C5 > C2	C2 > C5	Null
t test	d'	.571	0	.428
t test	γ	.444	0	.556
t test	H − FA	.563	0	.438
pctl	d'	.607	0	.393
pctl	γ	.607	0	.393
pctl	H − FA	.595	0	.405
BC_a	d'	.544	0	.456
BC_a	γ	.708	0	.292
BC_a	H − FA	.591	0	.408

based on the simulated participants' responses to three new and three old test stimuli per condition. Each section heading shows the d' level and bias for each of the two conditions being compared. The first and second columns present pairings of a statistic and a measure of sensitivity, which in combination define the method. The third column shows the proportion of the 2,000 runs in which the mean level of performance for Condition C1 exceeded that of Condition C2 and the null hypothesis was rejected at an α level of .05. The fourth column presents the proportion of runs in which the mean level of performance for C2 exceeded that for C1 and the null hypothesis was rejected. The final column shows the proportion of runs in which the null hypothesis could not be rejected.

No method was always better than the others. For example, the bootstrap methods, pctl_d' and BC_a_d', do well in the comparison of C2 versus C1 (equal in d' but differing in bias), in which they correctly accept the null hypothesis 92.1% and 92.8% of the time, respectively. By contrast, the *t*-test methods detect a difference in favor of C2 (a Type I error) 20.6%–41.6% of the time. In experiments involving comparisons between Conditions C5 and C2 (differing in d' but equal in bias), BC_a_γ detects the d' difference in 70.8% of the simulations. The next best methods, pctl_d' and pctl_γ, both detect the difference 60.7% of the time. Furthermore, it is simply not the case that γ is the best measure, since *t*-test_γ, the worst performer, detects the difference only 44.4% of the time.

Experimental comparisons grouped by type. Table 4 summarizes the results by showing the mean proportion of the 2,000 simulated experiments in which the null hypothesis was rejected for different kinds of comparisons: (1) those that differ in terms of d'; (2) those that are truly identical (i.e., the same condition is compared against itself); and (3) those that share the same level of d' but differ in bias. In general, methods that have lower Type I error rates tend to have higher Type II error rates. When these errors are considered to be of equal importance, the pctl_d' method looks best, irrespective of whether the participant saw 3, 10, or 20 trials per condition.

Table 4
Aggregate Performance Mean Experimental Hit (H) and False Alarm (FA) Rates

Statistic	Measure	$H_{A>B}$	$FA_{A=A}$	$FA_{A=B}$	$H_{A>B} - FA_{A=A,B}$
		3 Trials per Participant			
t test	d'	.567	.046	.318	.366
t test	γ	.499	.043	.178	.378
t test	H − FA	.565	.046	.331	.362
pctl	d'	.527	.078	.083	.444
pctl	γ	.482	.082	.107	.381
pctl	H − FA	.576	.053	.341	.362
BC_a	d'	.469	.093	.094	.366
BC_a	γ	.590	.146	.170	.416
BC_a	H − FA	.563	.060	.340	.345
best		BC_a_γ	*t*-test_γ	pctl_d'	pctl_d'
worst		BC_a_d'	BC_a_γ	pctl_H − FA	BC_a_H − FA
		10 Trials per Participant			
t test	d'	.739	.047	.381	.488
t test	γ	.581	.039	.063	.528
t test	H − FA	.713	.047	.502	.416
pctl	d'	.776	.072	.081	.694
pctl	γ	.710	.074	.113	.610
pctl	H − FA	.728	.061	.516	.417
BC_a	d'	.729	.068	.073	.653
BC_a	γ	.790	.120	.158	.638
BC_a	H − FA	.728	.066	.508	.416
best		BC_a_γ	*t*-test_γ	*t*-test_γ	pctl_d'
worst		*t*-test_γ	BC_a_γ	pctl_H − FA	*t*-test_H − FA
		20 Trials per Participant			
t test	d'	.811	.050	.324	.573
t test	γ	.598	.037	.040	.562
t test	H − FA	.744	.049	.544	.424
pctl	d'	.866	.075	.078	.783
pctl	γ	.805	.075	.119	.700
pctl	H − FA	.758	.063	.559	.423
BC_a	d'	.833	.071	.071	.757
BC_a	γ	.864	.124	.163	.707
BC_a	H − FA	.757	.071	.546	.422
best		pctl_d'	*t*-test_γ	*t*-test_γ	pctl_d'
worst		*t*-test_γ	BC_a_γ	pctl_H − FA	BC_a_H − FA

The column headed $H_{A>B}$ shows conditions that differ in terms of d'. For these comparisons, larger proportions correspond to better power (i.e., 1 − the Type II error rate). The bottom two rows of each section show which method performed best and which worst. Focusing on the three-trials-per-condition comparisons, BC_{a}_γ does best, correctly rejecting the null hypothesis in 59% of the comparisons.

The column headed $FA_{A=A}$ shows comparisons of conditions paired against themselves, so these proportions correspond to Type I error rates. Focusing first on the three-trials-per-condition comparisons, the *t*-test methods all achieved Type I error rates of less than .05. The bootstrap methods produce Type I errors above the intended .05 α level. Of particular interest are the bootstrap d' methods, pctl_d' and BC_{a}_d', with Type I error rates of 7.8% and 9.3%, respectively. BC_{a}_γ, the worst offender, has a Type I error rate of 14.6%. When all participants in a condition perform perfectly, the variability in the bootstrap distributions is reduced, contributing, we believe, to the slightly elevated proportion of Type I errors in the A = A comparisons.

The column headed $FA_{A=B}$ shows the results of comparisons between conditions that have shared levels of d' but differ in terms of bias. Since our performance measures are supposed to measure sensitivity free of bias, these proportions are Type I error rates. Again in the three-trials-per-condition comparison, the *t*-test methods have Type I error rates that range from 17.8% to 33.1%. The best performer is pctl_d', with a Type I error rate of 8.3%. The worst performing method, pctl_H − FA, had a Type I error rate of 34.1%.

Overall performance. The final column, headed $H_{A>B} - FA_{A=A,B}$, is a measure of method sensitivity fashioned after the H − FA measure. It discounts a method's power—that is, its tendency to detect differences when they are present ($H_{A>B}$)—by subtracting an aggregate Type I error rate, which we take to be the average of $FA_{A=A}$ and $FA_{A=B}$.

Using this measure, pctl_d' does best, because it makes modest numbers of Type I errors in both the A = A and the A = B comparisons. The *t*-test methods and all the H − FA methods are hurt by their tendency to make Type I errors in the A = B comparisons.

Performance with increasing numbers of trials. The results of simulations based on 10 and 20 trials per condition are shown in the middle and bottom sections of Table 4, respectively. The pattern is quite clear: The overall performance of the pctl_d' and BC_{a}_d' methods rise as the number of trials per participant increases. Their Type I error rates ranged from .068 to .075, whereas their Type II error rates fall. For the methods using H − FA, the Type I error rates rise dramatically, offsetting the modest fall in Type II errors.

Performance of *t*-test methods. Since some researchers may prefer to use a *t* test, which is a more commonly used technique in psychology than bootstrapping, we consider the *t*-test methods separately. An interesting case is *t*-test_γ, in the 20-trials-per-subject condition, since this is the only method to achieve the desired .05 α level for both the A = A and the A = B comparisons. In general, *t*-test_γ's performance improves with the number of trials. Notably, its Type I error rate in the A = B comparisons fall from 17.8% to 4%. Although it makes more Type II errors then the other methods, it makes substantially fewer Type I errors.

Performance weighted by error costs implied by the 5–80 convention. Controlling the proportion of Type I errors is, arguably, more important than controlling the proportion of Type II errors, because Type I errors are typically given greater weight in psychology than Type II errors. As a rule of thumb, Cohen (1988) suggests proportions of .05 and .80 for the Type I error rate and power (1 − Type II error rate) of an experiment. This pairing of error rates, sometimes called the *5–80 convention* (Di Stefano, 2003), implies that the cost of a Type I error is four times that of a Type II error. Table 5 shows the performance of the various methods when the error rates are weighted by the costs suggested by the 5–80 convention. Specifically, the entries equal (1 − Type II error rate) − 4(Type I error rate). These values serve to highlight the conclusions based on the equal weighting of errors—namely, that the H − FA methods do poorly and that their performance deteriorates as the number of trials increases. The performance of pctl_d' does best of all the measures, and it improves with the number of trials. The best of the *t*-test methods is *t*-test_γ.

RECOMMENDATIONS

Our recommendations depend on the preferences of the researcher and come with the caveat that a d' model must underlie the simulated data.

H − FA does poorly, especially in terms of Type I errors when the conditions differ in terms of bias but not in terms of d', regardless of whether the results are analyzed with a repeated measures *t* test or one of the bootstrap methods. This problem gets worse as the number of trials per subject increases. This could well be because the parametric assumptions of this measure are at odds with those of the d' model used to generate the data in the simulations. Nevertheless, those who use H − FA should

Table 5
Performance Weighted by Error Costs Implied by the 5–80 Convention

Statistic	Measure	3 Trials	10 Trials	20 Trials
t test	d'	−.161	−.117	.063
t test	γ	.057	.377	.444
t test	H − FA	−.189	−.385	−.442
pctl	d'	.205	.470	.560
pctl	γ	.104	.336	.417
pctl	H − FA	−.212	−.426	−.486
BC_a	d'	.095	.447	.549
BC_a	γ	−.042	.234	.29
BC_a	H − FA	−.237	−.420	−.477
best		pctl_d'	pctl_d'	pctl_d'
worst		BC_a_H − FA	pctl_H − FA	pctl_H − FA

be aware of the possibility that they could be misinterpreting differences in bias as differences in sensitivity.

Whether Type I and Type II errors are weighted equally or by error costs implied by the 5–80 convention, pctl_d' and BC_a_d' do best. Perhaps pctl_d' should be preferred because it slightly outperforms BC_a_d', especially in the three-trials-per-subject condition, and it is more straightforward.

A researcher who prefers the repeated measures t test over the bootstrap methods should opt for γ. It is less susceptible to Type I errors than d' is, even when each participant sees 20 trials per condition. Although it does produce substantially more Type II errors than does d', in combination with a t test γ is the only measure that approaches the .05 α level customarily used in psychology. Given that the data conformed completely to the d' model, it comes as something of a surprise that γ does so well.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, **58**, 17-22.

Di Stefano, J. (2003). How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology*, **17**, 707-709.

Efron, B., & Tibshirani, R. (1991). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**, 54-77.

Gallo, D. A., Roediger, H. L., III, & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review*, **8**, 579-586.

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

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

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review*, **10**, 718-723.

Martinez W. L., & Martinez A. R. (2002). *Computational statistics handbook with MATLAB*. New York: Chapman & Hall/CRC.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review*, **106**, 398-405.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* London: Sage.

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

Read, J. D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review*, **3**, 105-111.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words that were not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814.

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

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1379-1396.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, **107**, 368-376.

Zoubir, A. M., & Boashash, B. (1998). The bootstrap and its application in signal processing. *IEEE Signal Processing Magazine*, **15**, 56-76.

ARCHIVED MATERIALS

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

To access these files, search the archive for this article using the journal (*Behavior Research Methods*), the first author's name (Schooler) and the publication year (2005).

File: Schooler-BRMIC-2005.zip

Description: The compressed archive file contains the file schooler_appendix as a single .xls file, the contents of which are also available in .pdf and .txt format. This appendix contains the results of the experimental comparisons. The experimental comparisons are separated into 9 sheets. For example, the sheet named "A>B, ntrials=20" includes experiments in which there are d' differences between conditions and each subject participated in 20 trials per condition. A detailed description of the file format starts in the first paragraph of the results section of this article. Table 3 shows the results of three experimental comparisons.

Author's e-mail address: schooler@mpib.berlin.mpg.de

Author's Web site: http://www.mpib-berlin.mpg.de/en/mitarbeiter/home/schooler.htm

(Manuscript received October 9, 2000;
revision accepted for publication July 11, 2004.)