Three computational versions of proportion correct for use in forced-choice experiments[1]

DENNIS McFADDEN[2]
UNIVERSITY OF TEXAS

Three commonly used measures of performance in forced-choice experiments, all called proportion correct, are compared as response bias, bias in the stimulus outcomes, and level of performance are varied. No position is taken on the question of which is the computational version of choice, but some of the characteristics of the various versions are examined.

Forced-choice techniques are now being adopted by Es in many branches of psychological research. The reason for this is obvious—when used correctly, these techniques permit estimates of the S's ability to discriminate among the stimulus alternatives that are relatively independent of his tendency to use the various response alternatives. Psychophysicists were the first to recognize the confounding of these two factors in data collected with classical psychophysical methods, but awareness of the problem of separating discriminability and response criterion is now becoming widespread among psychologists concerned with phenomena other than sensory processing.

In order to keep the discussion simple, only forced-choice experiments that employ two stimulus alternatives and two response alternatives will be considered. Much of what will be said can undoubtedly be extended to include "m-alternative" experiments, but that is left for the future. So, each trial of the experiments under consideration here produces one of four possible stimulus-response combinations. Let us denote the two classes of stimuli "A" and "B," and the two response alternatives "a" and "b." Clearly, then, when the S is forced to give one of the two responses on every trial, every response contributes to one of the four proportions P(a/A), P(b/A), P(a/B), or P(b/B). These proportions are known as correct acceptances (hits), incorrect rejections (misses), incorrect acceptances (false alarms), and correct rejections, respectively, and they are frequently displayed as the entries in a 2 by 2 stimulus-response matrix. The stimulus alternatives might be words or trigrams that were either seen previously in the experiment or not seen previously (Egan, 1958; McFadden & Greeno, 1968), the presence or absence of a disk or line partially masked by a second disk or line (Schiller & Greenfield, 1969; Parlee, 1969), a visual array in which the elements are either all identical or identical except for one (Donderi & Zelnicker, 1969), a fixed-ratio reinforcement schedule of one magnitude or another (Rilling & McDiarmid, 1965), the presence or absence of a particular digit in a sequence of rapidly presented digits (Eriksen & Collins, 1969), two observation intervals, one of which contains a tone of one frequency and the other a tone of a different frequency (Henning, 1967), or, more familiarly, noise-alone or signal-plus-noise trials. Thus, the response alternatives in such experiments can be "old" and "new," "present" and "absent," "same" and "different," "Interval I" and "Interval II," "yes" and "no," etc. (The knowledgeable reader may wish to glance ahead to Note 3 at this point.)

The idea of presenting so-called "blank trials" or "catch trials" is an old one, but the use of the errors made on such trials in the assessment of the S's performance is a recent development. In the 1950s the theory of signal detectability, or TSD (see Swets, 1964; Green & Swets, 1966), first made clear to psychophysicists the importance of obtaining a reliable estimate of the errors on the "blank trials." Such errors were classically viewed as the hallmark of a poor S, but they are now regarded as indispensable when the objective is a measure of the S's ability to discriminate stimulus alternatives. Indeed, for reasons that will later become obvious, a good S is now defined as one that makes, on the average, as many false alarms as misses. TSD employs estimates of the S's hit rate and of his false-alarm rate to compute the statistic d'. Within the theory, this index of discriminability is independent of the S's response bias, i.e., his tendency to use one of the two responses more frequently than the other. Obviously, this independence is a characteristic of great value, for it allows comparisons of sensitivity between Ss with very different decision criteria.

Briefly, TSD assumes that each of the (here) two stimulus alternatives is represented "inside the S" as a distribution and, further, that the two distributions lie upon a common decision axis. Under the influence of several variables, the S is thought to select some value along this axis as his criterion for response. That is, if the sample value obtained on a particular trial is equal to or greater than the criterion value he has adopted, then he gives one response, and, if it is smaller, he gives the other response. Obviously, the adoption of an extreme value for the criterion will lead to a preponderance of responses of one kind, a response bias. In its simplest form, TSD assumes that the two underlying distributions are both normal with variances that are equal, and it is upon this assumption that the definition of d' lies. Quite simply, this measure is the distance in z-score units between the means of the two underlying distributions. Numerically, it is the z-score distance from the mean of the B distribution to the criterion value, minus the z-score distance from the mean of the A distribution to the criterion. These two distances are estimated for a block of trials from the false-alarm and hit rates, respectively.

As mentioned above, the beauty of d' is that, when the underlying distributions are normals of equal variance, this measure of discriminability is "pure" in the sense that it is unaffected by the particular response criterion adopted by the S. And this is true of no other metric presently in use. Consider, for example, another commonly seen measure of performance, the proportion (percentage) of correct decisions. For two-alternative experiments, this measure is usually expressed as

$$P(C) = P(A)P(a/A) + P(B)P(b/B), \quad (1)$$

where P(A) and P(B) are the a priori probabilities of the two stimulus alternatives (see Egan & Clarke, 1966). Table 1 shows three 2 by 2 matrices, all with the same proportion correct, but with different response biases, P(a). As can be seen, d' is different for each of these matrices; it is a minimum when the S has no response bias, P(a) = 0.5, and it increases as the response bias increases. In Table 2, d' is held constant and proportion correct changes when the three response biases of Table 1 are adopted. Here proportion correct is a maximum when there is no response bias, and it declines as the response bias grows. Clearly, then, equal proportion-correct matrices are not equal-d' matrices and conversely.

If this difference between the two measures is not intuitive, consider a S operating with such an extreme criterion

 Copyright 1970, Psychonomic Journals, Inc., Austin, Texas

Table 1
Equal-Proportion-Correct Matrices for Different Values of Response Bias, P(a). As the bias worsens, d′ increases.

Response Bias P(a)	Stimuli	Responses a	b	P(C)	d′
.50	A	.80	.20	.80	1.68
	B	.20	.80		
.40	A	.70	.30	.80	1.81
	B	.10	.90		
.35	A	.65	.35	.80	2.02
	B	.05	.95		

Table 2
Equal-d′ Matrices for Different Values of Response Bias, P(a). As the bias worsens, proportion correct decreases.

Response Bias P(a)	Stimuli	Responses a	b	P(C)	d′
.50	A	.800	.200	.800	1.68
	B	.200	.800		
.40	A	.685	.315	.785	1.68
	B	.115	.885		
.35	A	.617	.383	.767	1.68
	B	.083	.917		

that nearly *none* of the A distribution lies beyond his criterion, i.e., his hit rate, P(a/A), is nearly zero. It follows that nearly *all* of the B distribution lies below his criterion, making his correct-rejection rate, P(b/B), essentially 1.0, and for equal a prioris, Eq. 1 will yield a proportion correct close to chance performance. And this would be true no matter what the separation between the means of the two distributions, the d′. Less extreme criteria would lead to proportions correct less close to chance, but the point is that this measure is not criterion free.

When the percentage of correct decisions is plotted against some experimental parameter such as a measure of the "strength" of the stimulus, the resulting relation is known as a psychometric function (see Egan, Lindner, & McFadden, 1969). Classically, psychometric functions were plots of a hit rate, a P(a/A), without regard for the corresponding false-alarm rate. Consequently, these functions confound response bias with sensitivity. True psychometric functions, those employing measures of sensitivity as the ordinate, are of great interest to sensory psychologists because they show not only how well a S is discriminating at some stimulus level but also how his discrimination performance changes with changes in "stimulus strength." (Regrettably, many psychologists working in areas other than sensory psychology have yet to realize the value of determining the entire psychometric function instead of just a single point on it.)

A favorite measure of performance for psychometric functions is proportion correct, but, partly because of the effect demonstrated in Tables 1 and 2, there are presently in use at least three different methods for computing proportion correct from a forced-choice data matrix. The purpose of this paper is to give some examples of how these various versions differ. There is little said here that is new to anyone who has worked with data from forced-choice experiments, but perhaps the newcomer can be saved some computations and some anxiety about whether he is using the "correct percent correct."

In practice, Eq. 1 is employed in two ways, depending upon the particular E's interpretation of the quantities P(A) and P(B). That is, many Es program their trial sequences from a source of random numbers without constraints, a procedure which leads to stimulus biases on individual blocks of trials that are not necessarily accurate reflections of the a priori probabilities. Specifically, a block of 100 trials run with the a priori probability of P(A) = P(B) = 0.5 will only rarely result in exactly 50 trials of each stimulus alternative. When the obtained proportions of the two stimulus alternatives are different from the a priori probabilities, some Es use the obtained proportions for P(A) and P(B) in Eq. 1. This is easily shown to be algebraically equivalent to computing the proportion correct directly from the raw frequencies in the 2 by 2 data matrix. That is, when the obtained stimulus proportions are used for P(A) and P(B), Eq. 1 becomes

$$P(C)_1 = \frac{f_{(a/A)} + f_{(b/B)}}{f_{(a)} + f_{(b)}}, \qquad (2)$$

where $f_{(a/A)}$ is the number of responses a on A trials, etc.

Other Es ignore the fact that the a priori probabilities of the two stimulus alternatives are sometimes quite different from the obtained proportions. In computing proportion correct, these investigators use the a priori probabilities for P(A) and P(B) in Eq. 1. The most commonly used value for the a prioris is equality, i.e., P(A) = P(B) = 0.5. It is worth noting that when this value is adopted and the E uses the a priori probabilities for P(A) and P(B), Eq. 1 becomes

$$P(C)_2 = \frac{1}{2}[P(a/A) + P(b/B)]. \qquad (3)$$

Needless to say, if the E does not use an "independent-trials process" to program his trials but instead employs a procedure that constrains his stimulus outcomes to be equal, then for that experiment there is no difference between Eq. 2 and Eq. 3.

J. P. Egan introduced and named the last measure of performance to be considered here, maximum proportion correct or $P(C)_m$ (see Egan, 1965; Green & Swets, 1966). The motivation behind this measure is clearly illustrated in Table 2, where it is shown that the other estimates of proportion correct are not independent of the S's response bias. That is, equal-d′ matrices can yield different proportions correct. This is not true of $P(C)_m$ because this measure is obtained through a transformation of the d′ associated with the 2 by 2 data matrix. Specifically, a normal table is entered with a z-score of (d′/2), and the corresponding area is taken as $P(C)_m$. This value is the proportion correct that would have been obtained had the S adopted a symmetric criterion, P(a) = P(b) = 0.5, and had the underlying distributions both been equal-variance normals with a separation between the means of d′. Considering Table 2 again, $P(C)_m$ is .80 for all three matrices because the d's are equal. In computing $P(C)_m$ the E is, in effect, moving the S's criterion to the point of intersection of the two underlying distributions, the criterion value at which proportion correct is maximum given equal a prioris (Eq. 3) or equal obtained proportions (Eq. 2).

Many Es feel compelled to perform this "correction for bias," even though they know that the correction is typically small for trained Ss. They seem to feel that there is something unacceptable about crediting a S with a level of performance "below the one he could have achieved had he had no response bias." This last phrase is in quotations to indicate that such an attitude shows a deep commitment to the assumptions of TSD. Other Es are hesitant to "correct" their data because they are unwilling to allow theoretical commitments to affect their data, even if the effect is small. The purpose of this paper is to examine the discrepancies among the various computational versions of proportion correct. The discrepancies are examined first as a function of the S's response bias and then as a joint function of his response bias and of the bias in the stimulus proportions.

Table 3 shows six 2 by 2 matrices for each of five levels of performance as measured by d′. The response alternatives, a and b, are indicated at the top of each column of matrices, and the stimulus alternatives, A and B, are indicated at the left for each row of matrices. All six of the matrices in a given column have the same d′ and, hence, the same $P(C)_m$; the six matrices differ in the number (proportion) of each of the two responses given, i.e., in

Table 3

Each Column of Matrices Represents a Different Value of d′ and $P(C)_m$, and Each Row Represents a Different Value of Response Bias, P(a). The proportion correct shown can be regarded as either $P(C)_1$ or $P(C)_2$. Decimal points were omitted throughout the body of the table.

	$P(C)_m$ =	.950		.850		.750		.650		.550	
	d′ =	3.29		2.07		1.35		0.77		0.25	
P(a)		a	b	a	b	a	b	a	b	a	b
.10	A	200000	800000	19800	80200	189	811	161	839	122	878
	B	000018	999982	00177	99823	013	987	039	961	078	922
	P(C) =	600		598		588		561		522	
.20	A	4010	5990	3900	6100	356	644	302	698	235	765
	B	0002	9998	0094	9906	043	957	098	902	165	835
	P(C) =	700		690		656		602		535	
.30	A	5990	4010	570	430	508	492	429	571	343	657
	B	0012	9988	029	971	092	908	171	829	257	743
	P(C) =	799		770		708		629		543	
.40	A	7930	2070	729	271	640	360	544	456	448	552
	B	0067	9933	072	928	160	840	256	744	352	648
	P(C) =	893		828		740		644		548	
.45	A	8830	1170	794	206	697	303	599	401	500	500
	B	0177	9823	106	894	202	798	301	699	400	600
	P(C) =	933		844		748		649		550	
.50	A	950	050	850	150	750	250	650	350	550	450
	B	050	950	150	850	250	750	350	650	450	550
	P(C) =	950		850		750		650		550	

the response bias, P(a). The bottommost matrix in each column is what is known as the symmetric matrix because the hit and the correct-rejection rates are equal, as are P(a) and P(b). As the eye ascends each column, it encounters matrices that have increasingly large response biases toward the response b, i.e., P(a) declines. For each of the symmetric matrices in Table 3, the computed proportion correct is equal to the $P(C)_m$ for that column of matrices, but for all the other matrices in a column, the computed proportions correct are smaller than the corresponding $P(C)_m$. It is just such discrepancies that impel some Es to "correct for bias" by computing $P(C)_m$ instead of $P(C)_1$ or $P(C)_2$. For all of the calculations associated with Table 3 (and Tables 1 and 2 as well), it was assumed that either the a prioris were equal and Eq. 3 was used to compute proportion correct or that the obtained proportions were equal and Eq. 2 was used. Thus, the proportions correct shown in this table can be interpreted either as $P(C)_1$ or as $P(C)_2$, as a function of response bias.

As can be seen from Table 3, the magnitude of the discrepancy between the computed proportion correct and $P(C)_m$ for any value of response bias is different for the different columns, i.e., for the different levels of performance as measured by d′. For example, for the column d′ = 3.29 and $P(C)_m$ = .950, the values of proportion correct for the two extreme response biases, .10 and .50, differ by 35%, whereas that difference is only 2.8% in the column d′ = 0.25 and $P(C)_m$ = .550. Said differently, a given response bias leads to a discrepancy between computed proportion correct and $P(C)_m$ that is greater the greater the d′. This is a well-known fact, and it is an encouraging one because it is typically only when discrimination is difficult that trained Ss adopt response biases, not when discrimination is relatively easy. Yet the more likely a response bias, the less effect it has.

Again, for Table 3 the a priori probabilites and/or the obtained stimulus proportions were assumed to be equal. That assumption will now be relaxed, and the effects of both a response bias and a bias in the stimulus proportions will be examined. $P(C)_1$ will be considered first and then $P(C)_2$. If a S has a bias for the response b on a block of trials in which there are indeed more B trials, it is intuitive that $P(C)_1$ will be larger than if the S's bias were for the response a. In order to illustrate the combined effects of stimulus bias and response bias on the value of $P(C)_1$, Figs. 1a-1e have been prepared. The values on which the curves in Figs. 1a-1e are based were obtained in part by operating on the equal-d′ matrices shown in Table 3. For example, the family of curves shown in Fig. 1a was derived from the column in Table 3 labeled $P(C)_m$ = .950. For each matrix, the obtained stimulus proportions were systematically varied, the value of P(a) calculated, and then the value of $P(C)_1$ computed. In all figures, the ordinate is $P(C)_1$ and the parameter on the curves, P(A), is the obtained stimulus proportion (the a prioris were assumed to be equal). Thus, each curve shows, for a given value of the obtained proportions, how $P(C)_1$ varies as a function of the S's response bias, P(a).

What these figures illustrate is that the value of $P(C)_1$ computed for a block of trials is a pronounced function both of the obtained stimulus proportions and of the response bias of the S on that block of trials. Consider first the effect of a response bias on $P(C)_1$ when there is no bias in the stimulus proportions; this is shown in each figure by the curve designated P(A) = 0.5. These curves are symmetric around P(a) = 0.5, and they show several things: the more response bias a S demonstrates, the more his $P(C)_1$ will differ from $P(C)_m$; the magnitude of this difference, for a particular value of P(a), is greater the greater the d′; and in no case is it possible for $P(C)_1$ to be greater than $P(C)_m$ when the obtained stimulus proportions are equal.

The relationship becomes more complicated, however, when the stimulus proportions are biased. If the S responds without bias in the face of a bias in the stimulus proportions, $P(C)_1$ will be smaller than $P(C)_m$ by an amount that is dependent upon both the severity of the stimulus bias and the difficulty of the discrimination. If the S demonstrates a response bias that "follows" the bias in the stimulus proportions, as is likely when the discrimination is easy, then the resulting $P(C)_1$ may be greater than $P(C)_m$ (compare Green & Swets, 1966, p. 409). On the other hand, if his response bias is opposite to the direction of the stimulus bias, it is possible for $P(C)_1$ to be below the nominal value for chance performance in the forced-choice paradigm we are considering. Again it should be emphasized that for each point on each curve in these figures, there exists a 2 by 2 matrix having a d′ and a $P(C)_m$ equal to that for all other points

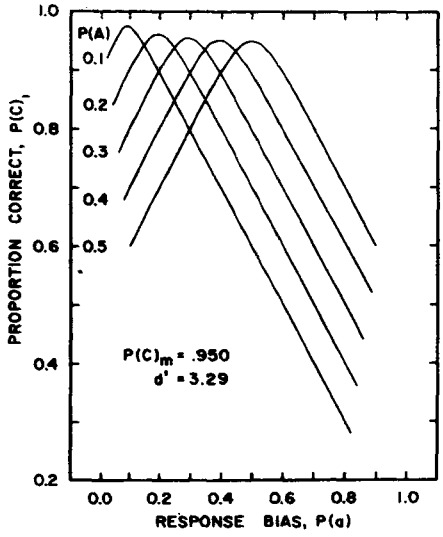

Figure 1a.

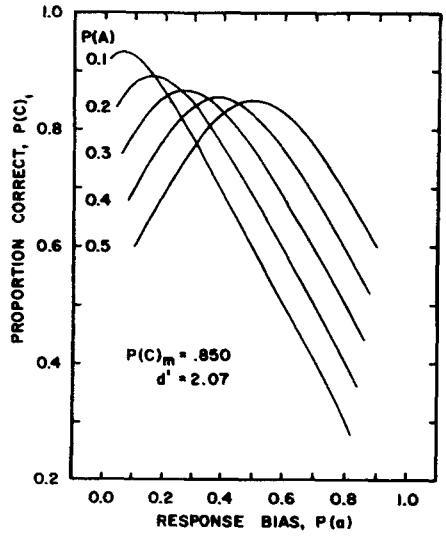

Figure 1b.

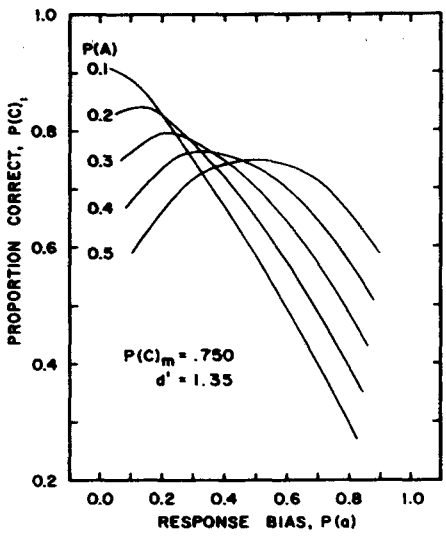

Figure 1c.

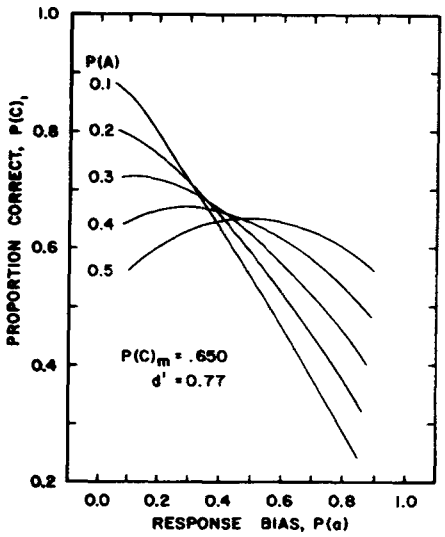

Figure 1d.

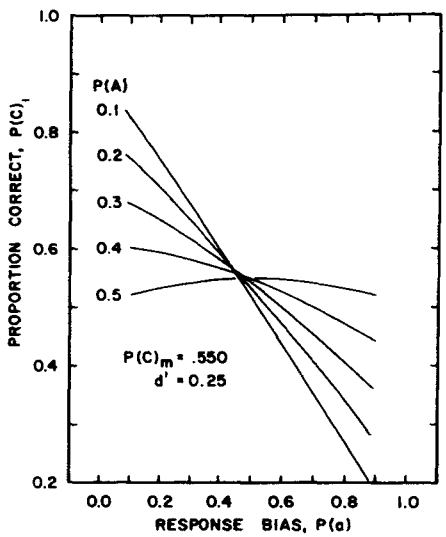

Figure 1e.

Figs. 1a-1e. $P(C)_1$ as a function of response bias, P(a), and of bias in the obtained stimulus proportions, P(A), for each of five values of $P(C)_m$. Each point on each curve is based upon a 2 by 2 data matrix that has a $P(C)_m$ which is identical to that for all other points on all other curves in that same figure. Clearly, $P(C)_1$ can be greater than, equal to, or smaller than $P(C)_m$, and it can be smaller than the value for chance performance, 0.5.

on all other curves in that same figure, but the $P(C)_1$ for that matrix is given by the ordinate value of the point. Values of P(A) greater than 0.5 are not shown in Figs. 1a-1e because they would be redundant; having a bias of 0.1 for response a and a bias in the stimulus proportions of 0.1 for stimulus A is identical to having a response bias of 0.9 and a stimulus bias of 0.9. That is, each curve could be reflected around the value P(a) = 0.5 to obtain a curve whose parameter would be the complement of that of the first curve.

Since block-by-block variations in the obtained stimulus proportions are ignored by the measure $P(C)_2$, one might expect that the differences between $P(C)_2$ and $P(C)_m$ would be less dramatic than those shown for $P(C)_1$ and $P(C)_m$ in Figs. 1a-1e. Figures 2a-2e confirm this expectation. The families of curves shown in these figures are based on the matrices of Table 3, just as were the curves in Figs. 1a-1e. The only difference is that after manipulating the obtained stimulus proportions and calculating P(a), the value for $P(C)_2$, not $P(C)_1$, was computed. The parameter on the curves, P(A), is the stimulus bias. The curves for P(A) = 0.5 in Figs. 2a-2e are, of course, identical to those shown in Figs. 1a-1e; they are produced again in order to facilitate comparison. Perhaps the most obvious difference between Figs. 1a-1e and Figs. 2a-2e is that, unlike $P(C)_1$, $P(C)_2$ never exceeds $P(C)_m$; it can only be equal to or smaller than $P(C)_m$. It shows in general a much smaller range of variation than does $P(C)_1$; indeed, because of this fact some of the curves had to be omitted in Figs. 2d and 2e. Finally, for these matrices, $P(C)_2$ never drops below the level of chance performance. Again it should be pointed out that exactly the same data matrices were used to obtain the curves for $P(C)_1$ and for $P(C)_2$.

One justification for presenting all of these tables and figures is that they can be used by an E to quickly estimate, for a particular data matrix, the discrepancy between the version of proportion correct he prefers to compute and the other two versions. Obviously, only rarely will all of the parameters of the matrix of data be exactly equal to one of those used to construct the figures, but the range shown should be adequate for "ballpark comparisons" for any data matrix.[3]

DISCUSSION

Now that the various computational versions of proportion correct have been compared as a function of several variables, the inevitable question of which version is "the best" can be faced. Clearly there can

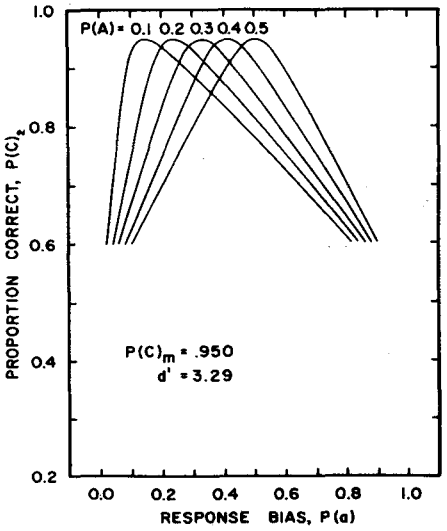

Figure 2a.

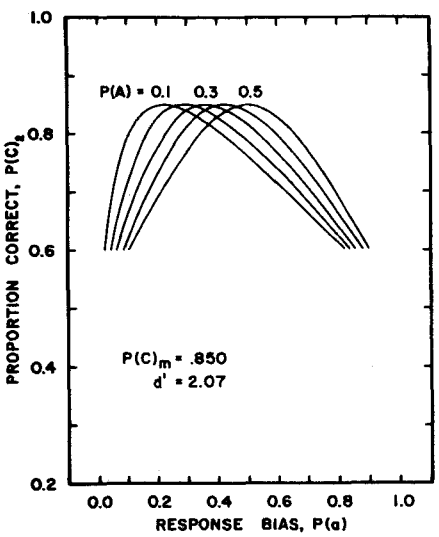

Figure 2b.

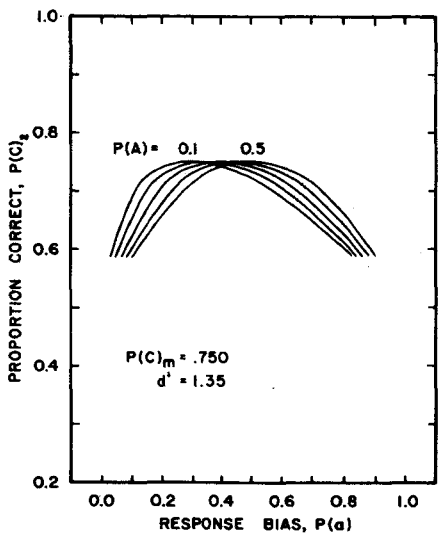

Figure 2c.

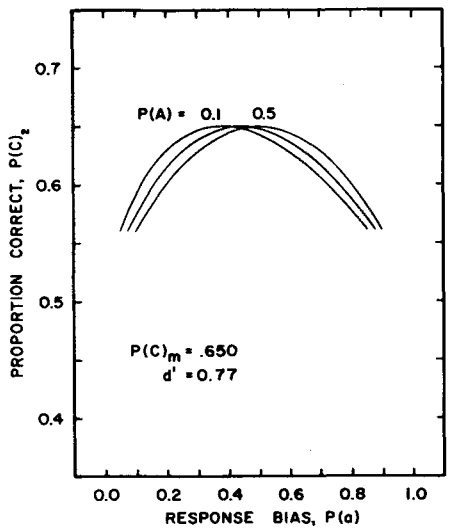

Figure 2d.

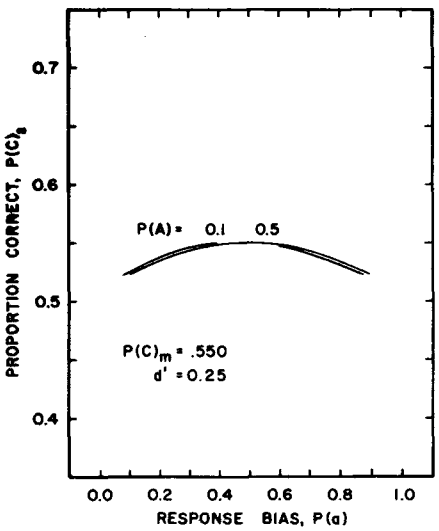

Figure 2e.

Figs. 2a-2e. $P(C)_2$ as a function of response bias, P(a), and of bias in the obtained stimulus proportions, P(A), for each of five values of $P(C)_m$. Each point on each curve is based upon a 2 by 2 data matrix that has a $P(C)_m$ which is identical to that for all other points on all other curves in that same figure. The matrices are the same as those used for Figs. 1a-1e. Unlike $P(C)_1$, $P(C)_2$ can only be equal to or smaller than $P(C)_m$, and, for the matrices considered, it is never below the value for chance performance. Note that in Figs. 2d and 2e the ordinate is expanded, and the entire family of five curves still could not be shown.

be no simple answer to such a question. The decision as to which version is the most appropriate will have to be made by each E for each experiment after careful consideration of such nonindependent factors as his reasons for doing the experiment, the theories on which his data bear, his willingness to allow his present theoretical commitments to affect the form of his basic data, and certain details of the paradigm. The most that can be done here is to point out some of the characteristics of the different computational versions in an attempt to aid the E in his choice.

In one sense the entire discussion up to this point has assumed, if only implicitly, the validity of TSD and of its various assumptions. That is, the figures and tables were constructed by holding d' constant, varying response bias and/or stimulus bias, and computing the various proportions correct. Such an approach might imply that d' and $P(C)_m$ are the measures of choice, the criteria against which other measures should be judged. Such an implication was not and is not intended. The figures and tables were based upon these measures only for reasons of convenience, not to champion them, for, in its simplest form, TSD is now recognized to be in error regarding some details. Most crucial to the present discussion is the discrepancy between theory and data on the question of the nature of the underlying distributions. Simple TSD assumes that they are normal distributions of equal variance, but in many experiments the data (specifically, the receiver operating characteristics, or ROCs) indicate distributions of unequal variance (see Green & Swets, 1966). The pity is that if the underlying distributions are not of equal variance, the value of d' is different for different decision criteria, i.e., it is no longer a criterion-free measure of discriminability. Continued use of d' in the face of evidence challenging the assumptions upon which this measure is based may appear foolish (or worse) to many readers. That it is still widely employed is partly a reflection of the fact that, in some areas of psychology, the d' of TSD is *more* appropriate than any of the other measures available, that is, it is *more nearly* "correct" than any of the alternative measures. (Certainly typical ROC data are better fitted by equal-d' functions than by equal-$P(C)_2$ functions, for example, since the latter are unit-slope straight lines in a ROC space with linear coordinates.)

But this is beside the point. All that need be noted here is that the computation of $P(C)_m$ as an attempt to "correct for

bias" presupposes the validity of the assumptions of simple TSD, some of which are acknowledged to be wrong in detail in certain experimental situations. If an E recognizes this problem but still feels that the use of $P(C)_m$ is justified because it is in some sense better than all of his present alternatives, and/or that its use is required in order to put his data into a form appropriate for the theoretical arguments he wants to make, then well and good. All that can be asked is that he be aware of his assumptions.

The E who is hesitant about making the assumptions required to compute $P(C)_m$ in a particular situation, but who feels that something like this version should be used, can be confident that the less theory-bound $P(C)_2$ will differ from $P(C)_m$ less, on the average, than will $P(C)_1$. As noted above, $P(C)_2$ can only be equal to or smaller than $P(C)_m$ and, particularly at moderate and low levels of performance, it is quite immune to fluctuations in the obtained stimulus proportions.

It is difficult to say much that is positive about $P(C)_1$. It is, at the same time, the most pure and the most primitive of the three versions discussed. Figs. 1a-1e show that the value of $P(C)_1$ can vary enormously as a function of response bias and of stimulus bias and that it can indicate below chance performance when the other versions of proportion correct indicate quite high discriminability. It does have one characteristic that might prove a virtue in some situations, namely, that averaging $P(C)_1$ over several blocks of trials, each taken at the same value of the stimulus parameter, is equivalent to cumulating across blocks the raw frequencies in each of the four cells of the data matrix and then computing $P(C)_1$.

At high levels of performance, a S will sometimes have either no false alarms or no misses (or both) on some blocks of trials. Since d' is infinite in such situations, $P(C)_m$ cannot be calculated for such blocks of trials. This is, of course, a disadvantage that neither of the other versions has. Es who prefer to use $P(C)_m$ typically compute one of the other versions when faced with a matrix containing a zero.

$P(C)_1$ and $P(C)_2$ differ in the way they treat block-by-block discrepancies between the a priori probabilities of the stimulus alternatives and the obtained stimulus proportions–$P(C)_1$ ignores the a prioris and $P(C)_2$ ignores the obtained proportions. It is intuitive that $P(C)_1$ is inherently the more variable version because it is based not only on an estimate of the S's sensitivity, but also on an estimate of the a priori probabilities. Of course, the computation of $P(C)_2$ or $P(C)_m$ involves no such estimates of the a prioris, and, as a consequence, these versions contain one less source of variance than $P(C)_1$.

In some ways, good Ss and poor Ss can be likened to $P(C)_1$ and $P(C)_2$. In many experiments, particularly psychophysical experiments, a good S is defined as one with a keen memory for the classes of input to be discriminated but no memory for the responses he has made on previous trials. That is, he does not adopt exotic response strategies, succumb to gambler's fallacy, etc., but he treats each trial as an independent event. Prior to the onset of every trial, the probability of his giving one response is equal to that of his giving the other (assuming the a prioris are equal). In a sense, such a S ignores the obtained proportions. He may show a response bias at the end of a biased block of trials, but his bias is not the result of a change in his tendency to give one of the two responses; it is simply a reflection of the bias in the stimuli on that block. Such a S could not show a large bias opposite to the stimulus bias. A poor S, on the other hand, is one whose tendency to give the two responses is biased as a result of one influence or another, and, furthermore, his bias may fluctuate from block to block and even within a block. Such a S might begin a block of trials with a predisposition toward one of the responses, say due to an extreme stimulus bias toward that stimulus alternative on the previous block, but as the block progresses, his criterion might move in one direction or the other if a stimulus bias began to develop. In a sense such a S ignores the a prioris and, clearly, the data from a S like this will be more variable than those from the good S discussed above.

Thus, $P(C)_1$ is the version with the most inherent variability, and some Ss are characterized as poor because of their more variable decision behavior, not necessarily because of any difference in their sensitivities. Taking these two facts together, an E using Ss that he knows to be less than highly trained or motivated might be apprehensive about the possibility of adding even more variance to his data by using $P(C)_1$.

As noted above, there is no difference between Eq. 2 and Eq. 3 when a trial sequence is determined not by an "independent-trials process" but by a procedure that results in equal numbers of each of the two stimulus alternatives on every block of trials. This practice warrants comment. Whereas such programming may be convenient and even necessary for proper counterbalancing in certain psychological experiments, in most psychophysical experiments it is generally a bad practice because the a priori probability of the two stimulus alternatives changes trial by trial when the stimulus outcomes are constrained to be equal. If the S knows and uses this information, then the measure of performance computed for him for that block of trials will be contaminated by a nonsensory variable of the sort modern psychophysics tries to eliminate. As an example, consider a block of trials in which a S receives by chance many more A trials than B trials during the first half of the block. If he knows there will be an equal number of both stimuli by the time the block is over, then he also knows that the probability of receiving a B trial is now much greater than that of receiving an A trial. Clearly, by simply increasing his rate of responding "b" during the second half of the block, he can guarantee himself some correct decisions he would not otherwise obtain. If equal stimulus outcomes are indispensable in some experiment, the E probably ought to make the trial sequences even more pseudorandom by preventing large local imbalances in the presentation rates so that the Ss will be less likely to adopt response strategies based on such imbalances.

Not to be minimized in this discussion is the experience of many Es who have concerned themselves with the question of which computational version of proportion correct to use and whose approach has been the strong-arm one of computing all three versions for every block of trials for every S in the experiment. The typical outcome of such heroic undertakings is differences in the *mean* data of about 5%-8%. Such differences are not insignificant in this era of precision psychophysics; on the other hand, they are not enormous differences, particularly in some other areas of psychology. Hopefully, as time passes, better and better measures will be developed, and/or we will learn how to run better experiments, and the problem of choosing among these versions of proportion correct will disappear or at least be minimized. Until then, the tables and figures presented here can facilitate comparisons among the various versions of proportion correct.

REFERENCES

DONDERI, D. C., & ZELNICKER, D. Parallel processing in visual same-different decisions. Perception & Psychophysics, 1969, 5, 197-200.

EGAN, J. P. Masking-level differences as a function of interaural disparities in intensity of signal and of noise. Journal of the Acoustical Society of America, 1965, 38, 1043-1049.

EGAN, J. P. Recognition memory and the operating characteristic. Technical Note AFCRC-TN-58-51, Hearing and

Communication Laboratory, Indiana University, 1958.
EGAN, J. P., & CLARKE, F. R. Psychophysics and signal detection. In J. B. Sidowski (Ed.), *Experimental methods and instrumentation in psychology*. New York: McGraw-Hill, 1966. Pp. 211-246.
EGAN, J. P., LINDNER, W. A., & McFADDEN, D. Masking-level differences and the form of the psychometric function. Perception & Psychophysics, 1969, 6, 209-215.
ERIKSEN, C. W., & COLLINS, J. F. Visual perceptual rate under two conditions of search. Journal of Experimental Psychology, 1969, 80, 489-492.
GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.
HENNING, G. B. Frequency discrimination in noise. Journal of the Acoustical Society of America, 1967, 41, 774-777.
McFADDEN, D. Masking-level differences with continuous and with burst masking noise. Journal of the Acoustical Society of America, 1966, 40, 1414-1419.
McFADDEN, D., & GREENO, J. G. Evidence of different degrees of learning based on different tests of retention. Journal of Verbal Learning & Verbal Behavior, 1968, 7, 452-457.
PARLEE, M. B. Visual backward masking of a single line by a single line. Vision Research, 1969, 9, 199-205.
RILLING, M., & McDIARMID, C. Signal detection in fixed-ratio schedules. Science, 1965, 148, 526-527.
SCHILLER, P. H., & GREENFIELD, A. Visual masking and the recovery phenomenon. Perception & Psychophysics, 1969, 6, 182-184.
SWETS, J. A. (Ed.), *Signal detection and recognition by human observers*. New York: Wiley, 1964.

NOTES

1. The preparation of this paper was supported in part by a grant from the Graduate School of the University of Texas. Preliminary drafts were read and criticized by J. P. Egan, D. S. Emmerich, H. Ermey, D. J. Foss, E. R. Hafter, G. H. Jacobs, and L. A. Jeffress.

2. Address: Department of Psychology, Mezes Hall, University of Texas, Austin, Texas 78712.

3. In psychophysics there are currently two popular forced-choice paradigms that yield data for which 2 by 2 matrices are appropriate, the single-interval or yes-no method and the two-interval forced-choice or 2IFC method. Both involve two stimulus and two response alternatives, that is, they could both be called *two-alternative* methods. The difference is that in the yes-no method the S is presented, on every trial, with a single sample drawn either from one distribution or the other, and his task is to decide on the origin of that sample, whereas in the 2IFC method he is presented with two samples on every trial, one drawn from each distribution, and he must decide which of the two temporal intervals contained the sample from the target distribution. Simple statistics reveal that the performance of an ideal S operating in the two tasks should be related as $\sqrt{2}\, d'_{yn} = d'_{2IFC}$, and this relation has been shown to hold fairly well for human observers in several detection situations. I have been intentionally vague in this paper about which of these two paradigms I was discussing, because I feel I am discussing both of them. If an E is running an analog to a 2IFC experiment and he has reason to believe that the square-root-of-two relation holds for his task, then he may choose to make this correction before computing proportion correct. Such manipulations are sometimes necessary in making an argument (e.g., see McFadden, 1966), but they are irrelevant to the point of this paper. What is said here about the various versions of proportion correct will be true for any *two-alternative* data matrix. All calculations shown in this paper happen to be based on yes-no statistics. The term "forced-choice" is used by some writers for 2IFC and multiple-alternative paradigms only, not for the yes-no method. This distinction has not been respected in this paper.

(Accepted for publication April 3, 1970.)