Software to calculate measures of sensitivity and response bias based on detection theory and threshold theory

KOEN VAN DER GOTEN
AND ANDRÉ VANDIERENDONCK
University of Ghent, Ghent, Belgium

This paper describes software that calculates 11 measures of sensitivity and 13 measures of response bias. The functions defined in the program can be stored as independent routines in a program library and can thus be called from other programs. The main program tests parameter limitations, and the calculations remain accurate below chance level. These measures are derived from Gaussian, logistic, and rectangular distributions. The bias measures are based on criterion location, criterion location relative to sensitivity, or likelihood ratio. The user is able to choose between different modes of output and two alternative ways to deal with extreme proportions. A brief description of the use and tenability of each of the measures is included.

Signal detection theory (SDT), originally proposed by Tanner and Swets (1954), has an important advantage over classical psychophysical theories. It allows for the segregation of a pure measure of sensitivity and a measure reflecting effects of cognitive factors—the response bias. Over the years, several variants have been proposed and used. Some indexes were originally thought to be non-parametric—for example, A' , B'_H , and B'' . It is clear now, however, that all measures of bias and sensitivity make distributional assumptions, and that none of them is non-parametric in the sense of being without distributional assumptions. Arguments for this position can be found in Swets (1986a, 1986b) and Macmillan and Creelman (1990, 1991, 1996). Swets (1986b) has evaluated several indexes of sensitivity, and Macmillan and Creelman (1990) did the same for the measures of response bias.

Swets (1986b) proposed one single standard by which a sensitivity statistic can be evaluated: the form of the empirical receiver operating curve (ROC). By this standard, he proposed the index A_z , which is the proportion of the area of the ROC graph that lies beneath an ROC (on ordinary scales) that is assumed to be a straight line on a binormal graph. In order to be able to calculate A_z , one has to obtain a sufficient number of data points to define an

ROC. A computer program developed by Dorfman and Alf (1969) and revised by Dorfman (cited in Swets & Pickett, 1982) is already available for the calculation of this index. This program requires a minimum of 3 points in ROC space (four answer categories). However, to calculate A_z , only 2 points in ROC space are required. Therefore, we incorporated functions that calculate A_z and related measures (d_a , d'_e , Δm , and the slope of the ROC curve) for exactly three categories. Swets (1986b) argued that when insufficient data points are available, some of the measures based on Gaussian and logistic distributions are the best choice. This manuscript describes software for the calculation of these “simple” measures, which may be of use when the necessary conditions for the calculation of A_z are impossible to obtain.

Macmillan and Creelman (1990) have studied the characteristics of bias measures that assume Gaussian distributions (SDT), logistic distributions (choice theory), and rectangular distributions (single and double high-threshold theory). Six criteria were used to evaluate the measures: Indexes were judged superior if (1) their isobias curves were monotonic, (2) they measured bias sensibly at and below chance, (3) their range was independent of sensitivity, (4) they were not distorted by averaging across observers, (5) they lent themselves to statistical analysis, and (6) they were of theoretical importance. By these standards, indexes of criterion location were judged to be the best response bias statistics.

Some programs are available for calculating some measures of sensitivity and bias. For example, Boice and Gardner (1988) described a Turbo Pascal program that calculates five measures, and Macmillan and Creelman (1991) mentioned other programs that calculate a number of measures. We developed a program that calculates 24 of the indexes of sensitivity and bias that were evaluated by Swets (1986b) and by Macmillan and Creelman (1990). Table 1 lists the measures, their assumed distributions, and the names of the routines that generate the indexes. These routines can be stored in a program library and thus can be called from other programs.

The user can choose between two methods to deal with the problem of infinite scores (hit and false-alarm proportions of 0 or 1) in the calculation of the measures. Macmillan and Kaplan (1985), among others, suggested replacing a proportion of 0 by $1/(2N)$ and a proportion of 1 by $1 - 1/(2N)$, where N is the number of noise or noise + signal presentations. The second method, proposed by Snodgrass and Corwin (1988) and frequently used in log-linear analysis, requires the addition of 0.5 to each cell in a 2×2 contingency table that defines the performance of the observer. Row and column totals are increased by 1. In this method, the contingency table is always transformed, irrespective of whether extreme proportions are present. Hautus (1995) has recently studied both methods by means of Monte Carlo simulations. He found that

This research was supported by Grant 33008793 of the Fund for Medical Scientific Research (Belgium) to A. Vandierendonck. The authors are indebted to Neil Macmillan, Michael J. Hautus, and two anonymous reviewers for their valuable comments on an earlier version of this paper. The authors are also indebted to Hans Vandierendonck for his invaluable help in translating the C programs in Pascal. Correspondence should be addressed to A. Vandierendonck, Department of Experimental Psychology, University of Ghent, Henri Dunantlaan 2, B-9000 Ghent, Belgium (e-mail: andre.vandierendonck@rug.ac.be).

Table 1
Indexes of Sensitivity and Bias Calculated by the
Routines, With Their Assumed Distributions
and the Names of the Routines

Distribution	Measure*	Routine
Gaussian	d'	sens_g
	c	loc_g
	c'	relloc_g
	$\beta_{(G)}$	like_g
Logistic	$\log(\beta_{(G)})$	loglike_g
	α	sens_alfa
	b	loc_l
	$\log(\alpha)$	sens_l
	$\log(b)$	logloc_l
	b'	relloc_l
	$\beta_{(L)}$	like_l
	$\log(\beta_{(L)})$	loglike_l
	B'_{H1}	hodos1
	B''	grier
Rectangular	q	sens_q
	$P(c)$	sens_pc
	k	loc_k
	k'	relloc_kp
"Nonparametric"	A'	sens_np
Others proposed	LOR	sens_lor
by Swets (1986b)	Yule's Q	sens_yule
Empirical ROC	slope	roc_slope
	Δm	delta_m
	d_a	roc_da
	d'_e	roc_de
	A_z	roc_az

*The reader is referred to Macmillan and Creelman (1990, 1991, 1996) and to Swets (1986a, 1986b) for additional information about the measures.

the log-linear rule results in less biased estimates of d' and always underestimates population d' . The $1/(2N)$ rule, apart from being more biased, either over- or underestimates population d' .

The main program requires a total of six or eight arguments, depending on the flags set. The first argument selects the desired calculation: G for the calculation of measures based on Gaussian assumptions; L for measures based on logistic assumptions; R for measures based on rectangular distributions; O for the measures suggested by Swets (1986b) and one "nonparametric" index of sensitivity. A prints all preceding measures. Finally, ROC selects the empirical ROC estimates. The second argument sets a flag to determine the way the program handles extreme proportions: $2N$ uses the $1/(2N)$ rule if an extreme case is present, and LL uses the log-linear rule for every proportion.

When the first flag selects measures based on a single point in ROC space (G , L , R , O , or A), in addition to the second flag, four arguments are required: the number of signal presentations, the number of noise presentations, the number of hits, and the number of false alarms. In case ROC is the first flag, the last four arguments are repeated for the second data point in ROC space. The output in the latter case consists of the *slope* of the ROC curve, Δm , d_a , d'_e , and A_z . These measures are preceded by a header describing the input and whether there has been a correction for extreme proportions.

When the A flag is chosen, the output consists of the 8 measures of sensitivity and the 13 indexes of response bias for a single point in ROC space. These measures are printed with a header describing the input, the assumed distribution, and whether there has been a correction for extreme proportions. Bias measures based on relative criterion or likelihood ratio are undefined when the hit ratio equals the false-alarm ratio. The main program mentions this when appropriate.

When accurately stored in a library, the 26 routines return the value of the sensitivity or bias estimate. The routines for a single point in ROC space are called with five arguments: the number of signal presentations, the number of noise presentations, the number of hits, the number of false alarms, and the flag for the correction for extreme proportions ($= 0$ for the $2N$ rule, $\neq 0$ otherwise). The measures based on empirical ROC also require the number of signal presentations, the number of noise presentations, the number of hits, and the number of false alarms for the second point. When the measure is undefined (*supra*), the function returns an arbitrarily large value, *HUGE*, defined in *math.h* for C; the Pascal program uses the same value. Every subroutine that calculates a measure calls the function *ratio*. This function checks for invalid parameters and corrects for extreme proportions according to the flag set. The correction for extreme proportions, when called for (log linear) or when necessary [for $1/(2N)$], is carried out for every measure. This way, the programs will always be computationally correct, and it is up to the researcher to judge whether the results are scientifically meaningful. The routine approximating the inverse of the normal distribution function is the one proposed by Odeh and Evans (1974); for a discussion, see Brophy (1985).

Description and Tenability

Measures based on empirical ROC. There is little doubt that the most appropriate measures for estimating sensitivity and bias are the measures based on ROC, assuming that the necessary conditions to calculate those measures can be met. Swets (1986b) and Swets and Pickett (1982) concluded that the slopes of empirical ROCs varied between 0.5 and 2.0 and that the slope is not predictable from sensitivity. In combination with the fact that sensitivity estimates may be misleading if ROC slopes are not equal to 1, this finding leads to the recommendation that ROCs should be collected whenever possible. When a single measure is desired to characterize an ROC, d_a is a good one, especially in forced-choice designs (see Macmillan & Creelman, 1991). Sometimes a measure expressed as a proportion instead of one expressed as a distance is preferred; A_z is appropriate in those circumstances. Another commonly used distance measure is d'_e , based on the arithmetic average of the standard deviations.

Measures of sensitivity for a single point in ROC space. Although d' and α have different roots, they are very similar as measures of sensitivity. Both are *good* measures of sensitivity (see, e.g., Macmillan & Creelman, 1991). The corrected hit probability q is based on single

high-threshold theory. Threshold models produce non-regular ROCs and are inconsistent with the available data, which show regular ROCs across a wide range of discrimination tasks. With proportion correct $p(c)$, the user assumes that the hit rate and the correct rejection rate are equal. This measure is based on double high-threshold theory and should never be used when better measures are available. One use for this $p(c)$, however, is in adaptive estimation procedures, as described in Macmillan and Creelman (1991). Swets (1986b) pointed out in a review of the literature that Yule's Q and LOR (log odds ratio) have been shown to be transformations of α . After ROC analysis, the author concluded that d' , α , Yule's Q , and LOR would lead to the same conclusions under most conditions. Macmillan & Creelman (1996) have shown that A' depends on two different measures of sensitivity: α and $p(c)$. At low levels of sensitivity, α dominates, whereas at higher levels $p(c)$ prevails. In other words, A' is consistent with logistic distributions at low sensitivities and with rectangular distributions at higher levels. Therefore, A' can hardly be said to be distribution free, as has been assumed by numerous authors.

Measures of response bias for a single point in ROC space. The bias measures based on Gaussian and logistic distributions, like the sensitivity measures, are virtually equivalent. In this short evaluation, we highlight the SDT measures. We can divide measures of response bias into three categories: those based on criterion location (c), criterion location relative to sensitivity (c'), or likelihood ratio ($\beta_{(G)}$). When sensitivity is constant, any of the three measures will do. As noted, a bias measure should minimally have an isobias curve that is supported empirically and that satisfies the monotonicity condition. The measure should also be independent of sensitivity. According to the two latter standards, criterion location c is the best one. Moreover, likelihood ratio and relative criterion are undefined when sensitivity equals zero, and their below-chance performance is nonoptimal. Two additional advantages of criterion location are that it is statistically independent of the sensitivity and that its range does not depend on the range of the sensitivity. From this enumeration it follows that for SDT, c is better than c' and $\beta_{(G)}$, and that for choice theory, b is better than b' and $\beta_{(L)}$. B'' and B'_H are monotonic transformations of the logistic likelihood ratio $\beta_{(L)}$ and are therefore equivalent. B'_H was computed by means of the formula proposed by Hodoss (1970) and extended by Singh and Karney (1987). B'' was calculated on the basis of a formula proposed by Aaronson and Watts (1987). Two other measures that have been thought to be distribution free, the yes-rate k and the error ratio k' , are actually consistent with double high-threshold theory. Macmillan and Creelman (1990) reported that both measures satisfy the monotonicity condition and that they behave well when sensitivity is at chance. Moreover, k is independent of its associated sensitivity measure $p(c)$ and behaves well below chance. The error ratio k' behaves

less well. According to Macmillan and Creelman (1991), however, there is one good reason to prefer SDT's c : When both the sensitivity and the bias are reported, they should derive from the same model, and it has already been shown that d' is superior to $p(c)$.

Availability

The source code (in C or Pascal) of the main program and/or the 26 routines can be obtained by sending an e-mail message to one of the authors (koen.vandergoten@rug.ac.be or andre.vandierendonck@rug.ac.be). Compiled versions for DOS and Solaris 2 are also available by public ftp (ftp.rug.ac.be/, directory pub/expssy).

REFERENCES

- AARONSON, D., & WATTS, B. (1987). Extensions of Grier's computational formulas for A' and B'' to below-chance performance. *Psychological Bulletin*, **102**, 439-442.
- BOICE, R., & GARDNER, R. M. (1988). A computer program to generate parametric and nonparametric signal-detection parameters. *Bulletin of the Psychonomic Society*, **26**, 365-367.
- BROPHY, A. L. (1985). Approximation of the inverse normal distribution function. *Behavior Research Methods, Instruments, & Computers*, **17**, 415-417.
- DORFMAN, D. D., & ALF, E., JR. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals rating-method data. *Journal of Mathematical Psychology*, **6**, 487-496.
- HAUTUS, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior Research Methods, Instruments, & Computers*, **27**, 46-51.
- HODOS, W. (1970). Nonparametric index of response bias for use in detection and recognition experiments. *Psychological Bulletin*, **74**, 351-354.
- MACMILLAN, N. A., & CREELMAN, C. D. (1990). Response bias: Characteristics of detection-theory, threshold-theory, and "nonparametric" indexes. *Psychological Bulletin*, **107**, 401-413.
- MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.
- MACMILLAN, N. A., & CREELMAN, C. 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., & KAPLAN, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, **98**, 185-199.
- ODEH, R. E., & EVANS, J. O. (1974). Algorithm AS 70: The percentage points of the normal distribution. *Applied Statistics*, **23**, 96-97.
- SINGH, N. S., & KARNEY, D. F. (1987). A simple algorithm to obtain nonparametric response bias estimates using Hodoss's method. *Behavior Research Methods, Instruments, & Computers*, **19**, 460-461.
- SNODGRASS, J. G., & CORWIN, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.
- 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., & PICKET, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.
- TANNER, W. P., JR., & SWETS, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, **61**, 401-409.