CPC, 1996, 15 (1), 3-28

Hit-rate bias in mental testing

Peter H. Schönemann[1] and William W. Thompson[2]

[1]*Purdue University and* [2]*University of Virginia, U.S.A.*

Abstract

New results are given concerning a form of test bias which, in the past, with few exceptions (esp. Cole, 1973; Hartigan & Wigdor, 1989), seems to have been largely ignored. We call it "hit-rate bias" because it is defined as the discrepancy between the hit-rates (= probability that a qualified testee passes the test) in a low- and a high-scoring group. Typically, it favors the high-scoring group. In contrast to Cole (1973), our focus is on binary criteria, such as college graduation. In the first, theoretical part, we present a (Hit-Rate Bounds) Theorem which underscores that raising predictor standards is not equivalent to raising criterion standards, as some believe. Instead, it typically increases hit-rate bias. We then derive and tabulate a simple approximation for estimating hit-rates as a function of validity, base-rate, and admission quota. In the empirical portion of the paper, we evaluate the extent of hit-rate bias in practice by re-analyzing a number of data sets involving the SAT, the ACT, and the GATB. Finally, we discuss how the addition of test scores to high school record affects hit-rate bias in predicting college graduation. We find it increases the bias.

Key words: Psychometrics, test theory, selection, test validity, base rate problem, test bias, hit-rate bias, point-biserial correlation. SAT, ACT, GATB, high school record.

---

1. Department of Psychological Sciences, Purdue University, 1364 Psychological Sciences Building, West Lafayette, IN 47907-1364, U.S.A. (e-mail: phs@psych.purdue.edu).
2. Abt Associates Inc., Bethesda, MD, 20814, U.S.A. (e-mail: bill_thompson @abtassoc.com).

INTRODUCTION

Although the problem of test bias has received considerable attention in recent decades, "there has been little agreement about what constitutes bias" (Cole, 1973, p. 237). In her paper, Cole discussed six different prediction models, each of which implies a different conception of bias, and its converse, fairness. In one of them, the Conditional Probability Model, "fairness" is defined as the equality of two conditional probabilities: equally qualified candidates in terms of criterion performance should have an equal chance of passing the test regardless of group membership. We agree with Cole that this is "an intuitively meaningful and defensible type of fairness" (p. 254), and with Hartigan and Wigdor (1989) that it would be unfair "to place the greater burden of prediction error on the shoulders of the lower scoring group" (p. 258). In this paper, we present some new theoretical and empirical results bearing on this particular bias problem.

In accord with conventional Signal Detection Theory, which defines "Hit-Rate" as the conditional probability that a device correctly identifies a signal, we refer to this type of bias as "Hit-Rate Bias". In this instance, the signal to be detected is a qualified testee, and the device is the mental test. Our focus – in contrast to Cole's (1973) and Hartigan and Wigdor's (1989) – will be on binary criteria, specifically, on college graduation, which is of considerable practical interest. Binary (truly dichotomous, as opposed to dichotomized) criteria have an additional advantage of being unambiguously defined, thus obviating quibbles about proper cut-offs to define "qualified" and "unqualified" subjects.

We shall usually communicate our numerical results as percentages rather than probabilities. To avoid ambiguity, we always use capital letters for percentages, reserving lower case letters for analogously defined probabilities. Thus, a "Hit-Rate" (*HR*) is the conditional percentage,

$$HR := 100\ \text{Prob}(\text{subject passes test} \mid \text{subject will graduate}) \qquad (1)$$

The vertical stroke, "|", is read "given that". We denote a "Hit-Rate Proportion", *HR*/100, by "*hr*". By "Hit-Rate Bias" we mean consistent hit-rate differences between two groups, an "Advantaged (e.g., high income) Group", and a "Disadvantaged (low income) Group". To measure the extent of the bias, we shall use the "Hit-Rate Bias Ratio" (*HRB*),

$$HRB := HR(\text{Advantaged Group}) \,/\, HR(\text{Disadvantaged Group}) \qquad (2)$$

If this ratio is 2, then for every one qualified testee of the disadvantaged group passing the test, two qualified testees of the advantaged group

will pass it. To anticipate, we will find that for White/Black comparisons, *HRB* is usually in the neighborhood of 1.7. Concretely, this means that qualified Blacks (those who would graduate if they were admitted), face steeper odds of being admitted than qualified Whites.

We should note that, in one sense, this type of bias is no-one's fault, because it is a direct mathematical consequence of the configuration sketched in Figure 1. The underlying assumptions are:

(a) The advantaged group outscores on average – for whatever reasons – the disadvantaged group both on the test and on the criterion;

(b) The test has imperfect predictive validity;

(c) The same cut-off, c, is used for both groups to define "pass".

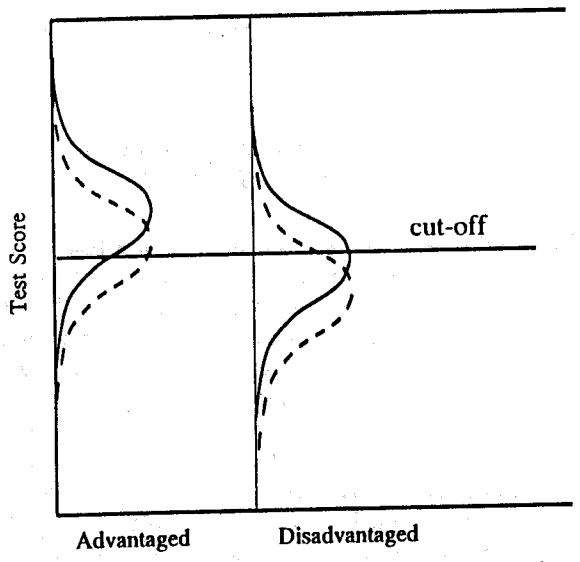

Figure 1. Distribution of test scores for advantaged and disadvantaged groups.

Notes: The solid distributions represent qualified subjects for both the advantaged and disadvantaged groups. The dashed distributions represent unqualified subjects for both the advantaged and disadvantaged groups.

If these conditions are met, as they often are, then it is intuitively obvious from Figure 1 that the proportion of qualified members passing the test will be smaller for the disadvantaged than for the advantaged group, and that the proportion of unqualified testees passing the test will be larger for the advantaged than for the disadvantaged group. In short,

low validity tests favor advantaged groups at the expense of disadvantaged groups.

We agree with Cole (1973) and Hartigan and Wigdor (1989) that this constitutes a form of bias which deserves more attention than it has received so far. If it is ignored, then it will aggravate existing social polarization by placing additional hurdles in the path of disadvantaged groups seeking equal economic opportunity.

Cursory study of Figure 1 also shows that the extent of this type of bias worsens as the cut-off, c, is raised in the mistaken belief that this is equivalent to "raising standards". Instead, as Figure 1 clearly shows, raising predictor cut-offs tends to eliminate a higher percentage of qualified members of the disadvantaged group than qualified members of the advantaged group, thus aggravating hit-rate bias. In this paper, we will sharpen these observations and assess the extent of hit-rate bias in practice.

BASIC DEFINITIONS: BASE-RATES, QUOTAS, AND HIT-RATES

We defer all derivations and proofs to the Appendices, leaving it up to the readers to consult them at their own discretion. Here we only review a few technical terms needed for understanding the body of the paper. They are summarized in Table 1. Whenever one has to choose between two uncertain events, four possible outcomes can occur, as shown in Table 1a. Two of the choices are correct and two of the choices are errors: "Qualified" subjects (q) who pass the test are called "Hits" or "True Positives". We denote their frequency "TP". Qualified subjects who fail the test are called "Misses" or "False Negatives" (FN). Unqualified subjects who fail the test are called "True Negatives" (TN), and unqualified subjects who pass the test are called "False Alarms" or "False Positives" (FP). Now let $TOTAL := TP+TN+FP+FN$ be the total frequency. We call the percentages

$$BR := 100\,(TP+FN)/TOTAL \tag{3}$$

the "Base Rate", and

$$Q := 100\,(TP+FP)/TOTAL \tag{4}$$

the (admission) "Quota" of the test. The corresponding proportions, when needed, will be written "br" or "q". The Total Percent Correct (Validity Rate, in some of the older literature),

$$\%C := 100\,(TP+TN)/TOTAL, \tag{5}$$

will be used as an intuitively straightforward measure of the predictive quality of a test. One obtains Table 1b by normalizing the columns of Table 1a. The resulting conditional percentages have special names: The "Hit-Rate",

$$HR := 100\ TP/(TP+FN) = 100\ \text{Prob(Pass}\,|\,\text{Graduate)}, \tag{6}$$

and the "Miss-Rate",

$$MR := 100\ FN/(TP+FN) = 100\ \text{Prob(Fail}\,|\,\text{Graduate)}. \tag{7}$$

Note that $HR+MR = 100$. Similarly, from the first column of Table 1a, one obtains the analogous percentages for the unqualified groups: The "True Negative Rate"

$$TNR := 100\ TN/(FP+TN) = 100\ \text{Prob(Fail}\,|\,\text{Drop out)}, \tag{8}$$

and the "False Positive" or "False Alarm Rate",

$$FR := 100\ FP = 100\ \text{Prob(Pass}\,|\,\text{Drop out)}. \tag{9}$$

They also sum to 100. We denote the corresponding proportions "*hr*", "*fr*", "*q*", and "*br*". In statistics, *hr* is called "power" and *fr* "significance level".

TABLE 1.
Definitions and notation.

Our notation is designed to assist nonspecialists by retaining symbols suggestive of the content. Our terminology is imported from signal detection theory. Statistical test theory employs a parallel set of terms. The definitions refer to the following joint frequency table:

Table 1a. Joint frequency table.

	Criterion (usually: graduation)		
Decision	Did Not Graduate (NG)	Graduated (GR)	
Pass Test (*P*)	False Positives (FP) ("False Alarms")	True Positives (TP) ("Hits")	*FP+TP*
Fail Test (*F*)	True Negatives (TN)	False Negatives (FN) ("Misses")	*TN+FN*
	FP+TN	*TP+FN*	*TOTAL*

From the joint frequencies Table 1a, three percentages can be derived which will be used throughout this paper:

Quota = $Q := 100\ (FP + TP) / TOTAL$ (T1.1)

Base Rate = $BR := 100\ (TP + FN) / TOTAL$ (T1.2)

Total Percent Correct = $\%C := 100\ (TP + TN) / TOTAL$ (T1.3)

Table 1b. Conditional percentages.

Table 1a can be converted into a table of conditional probabilities by dividing the elements in each column by their respective column sums. For convenience, the resulting fractions will be transformed into conditional percentages by multiplying them by 100. One thus obtains

	(Column) Conditional percentages	
	NG	*GR*
P	False Alarm Rate (*FR*)	Hit-Rate (*HR*)
F	True Negative Rate (*TNR*)	Miss-Rate (*MR*)
	100%	100%

The definitions of the terms in Table 1b are as follows:

Hit-Rate = $HR := 100\ TP / (TP + FN)$ (T1.4)

Miss-Rate = $MR := 100\ FN / (TP + FN)$ (T1.5)

False-Alarm-Rate = $FR := 100\ FP / (FP + TN)$ (T1.6)

Lower case abbreviations (%c, hr, mr, fr, tnr, q, br) will be used to denote the corresponding proportions:

Joint proportions

fp	tp	q
tn	fn	1-q
1-br	br	1

Conditional proportions

fr	hr
tnr	mr
1	1

HIT-RATE BOUNDS THEOREM

Calling *hr* a "Hit-Rate Proportion", we assert:

A Hit-Rate Proportion never exceeds the ratio of the Quota over Base-Rate:

$$hr \leq q/br \tag{10}$$

While the proof (given in Appendix 1) is straightforward, the result merits attention because it bears directly on the widespread, yet erroneous, belief that raising test cut-offs (e.g., of the SAT or ACT) is equivalent to "raising standards". The theorem shows that this is a fallacy because raising test cut-offs will introduce largely more misses, especially for larger base-rates and low validity tests. To illustrate this point, suppose that the *BR* is 80% and the selected cut-off results in a quota (*Q*) equal to 60%. Then the largest possible hit-rate is 100(60/80) = 75%. If we raise the cut-off to reduce *Q* to 40%, then the hit-rate cannot exceed 50%. For an even "higher (admission) standard" of 20%, the hit-rate remains below 25%.

The fallacy results from confusing test performance with criterion performance. This is an easy mistake to make, because psychologists are usually quite cavalier in their use of the loaded term "measure". They claim IQ tests "measure intelligence". However, when pressed, they cannot even define "intelligence" (e.g., "The fact that the concept of intelligence is at present not clearly defined is not troublesome", Jensen, 1983, p. 314). As long as the validity of a test is below 1, it never "measures" the criterion. It only measures itself, in the sense that it produces a numerical score which reflects the test performance. This score may or may not be useful for *estimating* criterion performance which is entirely different from *measuring* it. In particular, if the test validity is low, as it usually is, then the criterion estimate will be poor, even if the criterion – in our case college graduation – can be measured with perfect accuracy on a binary scale: either the student does or does not graduate. The semantic gap between test performance and criterion performance widens in direct proportion to the decline in validity. This is why raising predictor standards in no way is equivalent to raising criterion standards. For low validity tests it usually just makes it harder to estimate the criterion.

APPROXIMATE HIT-RATES AS A FUNCTION OF VALIDITY, BASE-RATE AND QUOTA

For strictly binary criteria, such as college graduation, the validity is given by the Point Biserial Correlation, r_{pb} which, in the present context, can be written

$$r_{pb} = d' \sqrt{br(1-br)}, \tag{11}$$

where $d' := (\mu_q - \mu_u)/\sigma$ denotes the standardized mean difference, familiar from signal detection theory, between the conditional distributions of the qualified and the unqualified subjects, and br is the base-rate proportion ($:= BR/100$).

Equation 11 can be used to express hr and br implicitly. Cursory experimentation suggests an iterative solution may not be easy. It is, therefore, of considerable practical interest to know that simple, explicit approximations for HR and FR can be obtained by linearizing the cumulative normal distribution in its middle range (around 50%). We found these approximations are quite close to the true values over the parameter range normally encountered in practice ($.3 < br$, $q < .7$, $r < .5$). As shown in Appendix 2, such linearization leads to

$$hr = q + r \sqrt{(1-br)/br} \,/\, 3 \tag{12}$$

and

$$fr = q - r \sqrt{br/(1-br)} \,/\, 3. \tag{13}$$

Simulations showed that 98.8% of the estimates of hr fall within .04 of the true values. Outside the range stated above, the accuracy falls off.

The first two column blocks in Table 2 indicate the robustness of our hr approximation for various values of q, br, and r. The header values for r (.1 to .5), together with the stated values for q and br, were used to approximate hr and fr according to the above linearized formulae. We then converted these conditional proportions into joint proportions from which we computed the phi-coefficients shown in the first column block of Table 2. The second block gives the approximations for the point-biserials based on the approximate hr's and fr's. On comparing them with the generating values listed as column headings, one finds that the linearized approximations reproduce the generating values closely.

As a fringe benefit, one further finds that the point-biserial correlation can be approximated over the same range by

$$r_{pb} = 3 \sqrt{(q-fr)(hr-q)}. \tag{14}$$

This obviates the need for normal tables to find point-biserial validities in the lower range.

Table 2.
Robustness indices, ϕ versus r_{pb}, and Base-Rate Problem.

		Phi (ϕ)	Point-biserial (r_{pb})	$\%c$ - max(br, 1-br)
br	*q*	*r*: .1 .2 .3 .4 .5	.1 .2 .3 .4 .5	.1 .2 .3 .4 .5
.3	.3	.1 .1 .2 .3 .4	.1 .2 .3 .4 .5	-.1 -.1 .0 .0 .0
	.4	.1 .1 .2 .3 .3	.1 .2 .3 .3 .4	-.1 -.1 -.1 .0 .0
	.5	.1 .1 .2 .3 .3	.1 .2 .3 .3 .4	-.2 -.1 -.1 -.1 .0
	.6	.1 .1 .2 .3 .3	.1 .2 .3 .4 .5	-.2 -.2 -.1 -.1 -.1
	.7	.1 .1 .2 .3 .4	.1 .2 .3 .5 .6	-.3 -.2 -.2 -.2 -.1
.4	.3	.1 .1 .2 .3 .4	.1 .2 .3 .4 .5	.0 .0 .0 .1 .1
	.4	.1 .1 .2 .3 .3	.1 .2 .3 .3 .4	.0 .0 .0 .1 .1
	.5	.1 .1 .2 .3 .3	.1 .2 .2 .3 .4	-.1 .0 .0 .0 .1
	.6	.1 .1 .2 .3 .3	.1 .2 .3 .4 .5	-.1 -.1 .0 .0 .0
	.7	.1 .1 .2 .3 .4	.1 .2 .3 .4 .6	-.1 -.1 .0 .0 .0
.5	.3	.1 .2 .2 .3 .4	.1 .2 .3 .4 .5	.0 .1 .1 .1 .2
	.4	.1 .1 .2 .3 .3	.1 .2 .3 .3 .5	.0 .1 .1 .1 .2
	.5	.1 .1 .2 .3 .3	.1 .2 .3 .3 .4	.0 .1 .1 .1 .2
	.6	.1 .1 .2 .3 .3	.1 .2 .3 .3 .5	.0 .1 .1 .1 .2
	.7	.1 .2 .2 .3 .4	.1 .2 .3 .4 .5	.0 .1 .1 .1 .2
.6	.3	.1 .1 .2 .3 .4	.1 .2 .3 .4 .6	-.1 -.1 .0 .0 .0
	.4	.1 .1 .2 .3 .3	.1 .2 .3 .4 .5	-.1 -.1 .0 .0 .0
	.5	.1 .1 .2 .3 .3	.1 .2 .2 .3 .4	-.1 .0 .0 .0 .1
	.6	.1 .1 .2 .3 .3	.1 .2 .3 .3 .4	.0 .0 .0 .0 .1
	.7	.1 .1 .2 .3 .4	.1 .2 .3 .4 .5	.0 .0 .0 .1 .1
.7	.3	.1 .1 .2 .3 .4	.1 .2 .3 .5 .6	-.3 -.2 -.2 -.2 -.1
	.4	.1 .1 .2 .3 .3	.1 .2 .3 .4 .5	-.2 -.2 -.1 -.1 -.1
	.5	.1 .1 .2 .3 .3	.1 .2 .3 .3 .4	-.2 -.1 -.1 -.1 .0
	.6	.1 .1 .2 .3 .3	.1 .2 .3 .3 .4	-.1 -.1 -.1 .0 .0
	.7	.1 .1 .2 .3 .4	.1 .2 .3 .4 .5	-.1 -.1 -.0 .0 .0

Table 3 gives hit- and miss-rates based on these approximations. To calculate a hit-rate bias ratio (*HRB*), one first looks up the *HR*s as a function of *r*, *BR*, and *Q*, for both groups, and then forms their ratio. To guard against overestimates caused by slight inaccuracies in the *HR* approximations, we suggest rounding up small denominators. To illus-

trate, let $BR = 60\%$ and $Q = 70\%$ for the advantaged group, and $BR = 40\%$ and $Q = 30\%$ for the disadvantaged group, and the validity $r = .4$ for both groups. Then one finds from Table 3 that the *HR*'s are 81% and 46%, respectively. To err on the conservative side, we round up the denominator to 50%, to obtain a $HRB = 81\%/50\% = 1.6$.

Table 3.
Approximate Hit-Rates and False-Alarm Rates for binary criteria.

		Hit-Rates					False Alarm Rates				
BR	*Q*	*r:* .1	.2	.3	.4	.5	.1	.2	.3	.4	.5
30	30	35	40	45	50	55	28	26	23	21	19
	40	45	50	55	60	65	38	36	33	31	29
	50	55	60	65	70	75	48	46	43	41	39
	60	65	70	75	80	85	58	56	53	51	49
	70	75	80	85	90	95	68	66	63	61	59
40	30	34	38	42	46	50	27	25	22	19	16
	40	44	48	52	56	60	37	35	32	29	26
	50	54	58	62	66	70	47	45	42	39	36
	60	64	68	72	76	80	57	55	52	49	46
	70	74	78	82	86	90	67	65	62	59	56
50	30	33	37	40	43	47	27	23	20	17	13
	40	43	47	50	53	57	37	33	30	27	23
	50	53	57	60	63	67	47	43	40	37	33
	60	63	67	70	73	77	57	53	50	47	43
	70	73	77	80	83	87	67	63	60	57	53
60	30	33	35	38	41	44	26	22	18	14	10
	40	43	45	48	51	54	36	32	28	24	20
	50	53	55	58	61	64	46	42	38	34	30
	60	63	65	68	71	74	56	52	48	44	40
	70	73	75	78	81	84	66	62	58	54	50
70	30	32	34	37	39	41	25	20	15	10	05
	40	42	44	47	49	51	35	30	25	20	15
	50	52	54	57	59	61	45	40	35	30	25
	60	62	64	67	69	71	55	50	45	40	35
	70	72	74	77	79	81	65	60	55	50	45

POTENTIAL HIT-RATE BIAS CONFOUNDS: BASE-RATES AND QUOTAS

In their provocative book, *The Case Against the SAT*, Crouse and Trusheim (1988) conclude, on the basis of a thorough analysis of large data sets:

(a) "The SAT thus acts much like a supplement to high school rank with zero validity that rejects additional blacks" (p. 107); and

(b) "with respect to admission outcomes for low income students [the SAT acts] much as it does for blacks" (p. 131).

In view of the severity of this indictment, it seems prudent to make sure that it is not the result of some spurious confound unrelated to the merits of the test. To anticipate, we will find that some qualifications of the previous statements are indeed warranted, but that the primary conclusion survives intact and is in fact corroborated by independent data sets.

Before we turn to a review of some large data sets to assess the extent of hit-rate bias in practice, we, therefore, briefly digress to discuss two potential confounds which could becloud the interpretation of hit-rate bias, especially for low validity tests. Such tests are the rule, not the exception, especially as the prediction interval lengthens. As Humphreys (1968) has shown, ACT validities drop from .48 for freshman GPA (FGPA) to .16 for senior GPA. Similarly, Crouse and Trusheim (1988) found that the validity of the SAT for the college graduation criterion is in the low 20s.

Base-rate confounds

For low validity tests, the base-rate problem is especially grave. It is apparently not universally realized that a test with positive validity may be worse than useless, namely when betting on the more frequent criterion outcome ensures a larger percentage of correct decisions than use of the test-score. Concretely, if one defines "mentally retarded" by an IQ more than two standard deviations below the mean (or "genius" by an IQ more than two standard deviations above the mean), then simply betting on "normal" (or, equivalently, admitting randomly on the premise that every-one is "normal") ensures over 97% correct decisions. A test would have to have unrealistically high validity to override such a high percentage of correct decisions obtainable from high base-rates alone.

In spite of the self-evident importance of this base-rate problem, and the indisputable fact that the validities of college entrance exams are

modest at best (see, e.g., Donlon, 1984, p. 160), we have been unable to locate quantitative assessments of the extent of the base-rate problem in practice. We, therefore, computed the percentages of correct decisions obtained (a) with predictions based on the test and (b) with predictions based on the maximum base-rate alone. The results are given in the third block of Table 2 in terms of the differences $\%C/100 - \max(br, 1\text{-}br)$. Positive values mean the test improves over random admissions, and negative values that it makes matter worse.

As can be seen, only for even base-rate splits ($br = .5$) do tests uniformly improve over random admissions. For 2:3 base-rate splits, only tests whose validities exceed .4 will improve over random admissions, while tests with validities below .2 are worse than no test at all. For 3:7 splits (as found in Crouse and Trusheim's 1988 chapter on Black/White comparisons, see below), no test in the realistic validity range ($r < .5$) is likely to improve over random admissions. Thus it appears that the base rate problem is much more severe in practice than its neglect by most theorists might suggest. For a notable exception to this rule see Meehl and Rosen (1955).

Quota confounds

Since *HRB*s involve a comparison of two groups which, in general, will have different admission quotas and base-rates, it is important to understand that smaller quotas tend to induce smaller hit-rates (see Hit-Rate Theorem). This is relevant for hit-rate comparisons of single versus compound predictors (e.g., between high-school record alone versus high-school record plus entrance test) when the criterion, such as FGPA, is continuous. If one retains the same cut-off to define "pass", then combining both predictors will induce smaller quotas so that the effect of adding another predictor can no longer be distinguished from the effect of the induced tighter quotas. An obvious remedy for this problem is re-adjustment of the quotas of the composite so that they uniformly dominate the quotas for the single predictor.

EMPIRICAL OVERVIEW OF HIT-RATE BIAS IN PRACTICE

Crouse and Trusheim: SAT, Black/White contrast

The results of our re-analysis of Crouse and Trusheim's Black/White data are shown in Table 4. The joint frequencies *TP, TN, FP*, and *FN*

(see Table 1 for definitions) were reconstructed from the proportions in the bottom half of Crouse and Trusheim's Table 5.8 (p. 104). In view of the high base-rates (77.4% for Whites, 65.7% for Blacks), it comes as no surprise that High School Rank (HSRNK) alone is already worse than purely random admissions (cf. Table 2) for all three groups. Adding the SAT to HSRNK depresses the *%C* still further (see Table 4c). Thus, these particular data are not conclusive evidence against the SAT, because the differential bias effect could be due to the severely skewed base-rates, rather than any intrinsic flaw in the SAT.

Table 4.
Hit-rate bias confound in Crouse and Trusheim Black/White data.

Table 4a. Joint frequency tables for White, Black, and total using HSRANK.*

	White			Black			Total		
	NG	*GR*	Total	*NG*	*GR*	Total	*NG*	*GR*	Total
P	322	1343	1665	44	106	150	366	1449	1815
F	127	195	322	25	26	51	152	221	373
	449	1538	1987	69	132	201	518	1670	2188

Table 4b. Joint frequency tables for White, Black, and total using HSRANK+SAT.*

	White			Black			Total		
	NG	*GR*	Total	*NG*	*GR*	Total	*NG*	*GR*	Total
P	322	1323	1645	429	87	116	351	1410	1761
F	127	215	342	40	45	85	167	260	427
	449	1538	1987	69	132	201	518	1670	2188

* The criterion variable was graduation and the decision variable was predicted GPA (pass > 2.5). Data from Crouse and Trusheim (1988, p. 104, Table 5.8).

Table 4c. Summary statistics for joint frequency tables.

	White		Black		Total	
	HSRNK	+SAT	HSRNK	+SAT	HSRNK	+SAT
N	1,987	1,987	201	201	2,188	2,188
BR	77.4	77.4	65.7	65.7	76.3	76.3
Q	83.8	82.8	74.6	57.7	83.0	80.5
%C	74.0	73.0	65.2	63.2	73.2	72.1
HR	87.3	86.0	80.3	65.9	86.8	84.4

Note: $HRB_{HSRNK} = 1.1$; $HRB_{HSRNK+SAT} = 1.3$.

However, Crouse and Trusheim present additional data which strengthens their case. Their Table 5.5 (p. 100) gives two joint distributions of College GPA predicted (a) from HSRNK alone, and (b) from HSRNK plus SAT, for Whites and for Blacks. They concluded that, for these data, adding the SAT increases bias against Blacks in the sense that Whites are screened in and Blacks are screened out.

One potential problem with this analysis is that predicted college GPA invites quota confounds since it is used with the same cut-off (2.5) in the single and the composite predictor case. As the authors note, adding the SAT to HSRNK tightens the quotas of the composite predictor relative to those of the single predictor.

In order to control for this potential quota confound, we pooled both joint distributions and then located a series of cut-offs to equalize the quotas for single and composite predictors as much as possible, making sure that all composite predictor quotas exceed those for HSRNK alone (thereby biasing the outcome in favor of the composite predictor). The results of our re-analysis confirm part of Crouse and Trusheim's main conclusion, that the addition of the SAT screens in more Whites and screens out more Blacks (See Table 5a). On average across all quotas, addition of the SAT admits 153 (6.9%) additional White students, and it rejects 23 (9.9%) additional Blacks (Table 5b). Since, according to the authors (p. 94), roughly 700,000 Whites and 75,000 Blacks took the SAT in 1972, these estimates translate nationwide into roughly 48,000 Whites who owe their admission to the addition of the SAT over and above HSRNK, and 7,400 Blacks who were denied admission on the basis of the SAT composite but would have been admitted on the basis of HSRNK alone.

Table 5.
Admission bias in Crouse and Trusheim Black/White data after correcting for quota confound.

Table 5a. Effects on admissions when adding SAT to HSRANK.

Total quota		Whites admitted		Blacks admitted	
HSRANK	+SAT	HSRNK	+SAT	HSRNK	+SAT
40.4	48.2	929	1134	67	48
54.6	59.1	1235	1381	103	69
65.2	69.2	1437	1596	126	100
73.8	77.4	1662	1770	147	127
80.3	83.3	1802	1890	168	152

Notes: N = 2,220 Whites and N = 232 Blacks. BR_{white} = 79.8%; BR_{black} = 54.7%; BR_{total} = 77.4%; The decision variable was Predicted GPA (Pass > 2.5). Data from Crouse and Trusheim (1988, p. 100, Table 5.5).

Table 5b. Change in admissions after adding SAT to HSRANK.

White gains (%)	Black losses (%)
205 (8.8)	-19 (8.2)
146 (6.6)	-34 (15.7)
123 (5.6)	-26 (11.2)
108 (5.1)	-20 (8.6)
88 (4.0)	-16 (6.9)

While the reanalysis of Crouse and Trusheim's joint distributions in their Table 5.5 strengthens the plausibility of their main point, it does not completely prove it, because the joint distributions do not involve the criterion. On the basis of these data it is not possible to determine what percentage of Whites screened in are qualified, and what percentage of Blacks screened out are unqualified. To unambiguously settle the problem of differential hit-rate bias, one needs joint distributions which involve the criterion. Before returning to this problem, we briefly review two other data sets bearing on hit-rates generally.

Crouse and Trusheim: SAT, low/high income contrast

Base-rate confounds are no longer a problem for the High/Low Income data Crouse and Trusheim discuss in Chapter 6 of their book (Table 6.5, summarized in our Table 6). In this case, the base-rates are near 50%. HSRNK substantially improves the *%C* over random admissions.

Table 6.
Hit-rate bias in Crouse and Trusheim Income data.

Table 6a. Joint frequency tables for High, Low Income groups using HSRANK alone.*

	High Income			Low Income			Total		
	Low GPA	*High GPA*	Total	*Low GPA*	*High GPA*	Total	*Low GPA*	*High GPA*	Total
P	480	832	1312	154	159	313	634	991	1625
F	180	77	257	54	14	68	234	91	325
	660	909	1569	208	173	381	868	1082	1950

Table 6b. Joint frequency tables for High, Low Income groups using HSRANK+SAT.*

	High Income			Low Income			Total		
	Low GPA	*High GPA*	Total	*Low GPA*	*High GPA*	Total	*Low GPA*	*High GPA*	Total
P	458	830	1288	124	150	274	582	980	1562
F	202	79	281	84	23	107	286	102	388
	660	907	1569	208	173	381	868	1082	1950

* The criterion variable was College GPA (qualified > 2.5) and the decision variable was predicted GPA (pass > 2.5). Data from Crouse and Trusheim (1988, p. 130, Table 6.5).

Table 6c. Summary statistics for joint frequency tables.

	High Income		Low Income		Total	
	HSRNK	+SAT	HSRNK	+SAT	HSRNK	+SAT
N	1,569	1,569	381	381	1,950	1,950
BR	58.0	58.0	45.4	45.4	55.5	55.5
Q	83.6	82.1	82.2	71.9	83.3	80.1
%C	64.5	65.8	55.9	61.4	62.8	64.9
HR	91.5	91.3	91.9	68.7	91.6	90.6

Note: $HRB_{HSRNK} = 1.0$; $HRB_{HSRNK+SAT} = 1.3$.

However, the quota confound is still a potential problem because the authors retain the same cut-off (predicted FGPA = 2.5) for FGPA predicted from HSRNK and from HSRNK+SAT. As a result, the quotas are uniformly smaller for the composite, which – in view of the hit-rate bounds theorem – in turn depresses the hit-rates of the composite. Concretely, for HSRNK the quota is near 83% for both groups (see Table 6c). When the SAT is added, it changes little for the High Income group, but drops 10% for the Low Income group. As a result, the hit-rates change little for the High-Income group, but drop 20% for the Low Income group. Adding the test to HSRNK raises *HRB* from 1.0 to 1.3. This bias increase must be weighed against a 2% increase in *%C* for the total group.

The 20% drop in the Low Income *HR* seems to lend support to Crouse and Trusheim's main contention that adding the SAT to HSRNK penalizes the Low Income group. However, a skeptic might attribute the *HR* drop to the lower quotas induced by the Predicted FGPA. Before the case against the SAT can be considered conclusive, one first has to control the potential quota confound induced by their continuous FGPA criterion.

Hartigan and Wigdor: GATB, Black/White contrast

Hartigan and Wigdor (1989) present a small data set of 136 carpenters who took the General Aptitude Test Battery (GATB). The criterion was job performance ratings by superiors. We reproduce their joint frequency data in Table 7a and the summary statistics in Table 7b.

One striking fact – not commented on by the authors – is the base-rate problem for Whites: their *%C* on the basis of the GATB is exactly the same as the *%C* for random admissions (summary Table 7b). For Blacks, on the other hand, the test improves prediction of job rating considerably (raising the *%C* from 35.6% to 71.7%). The fact that both within group validities are approximately equal is in some quarters interpreted as a sign that the test is "unbiased". However, this is far from true for our hit-rate definition of "bias", because the bias ratio is *HRB* = 1.69. Hartigan and Wigdor (1988, p. 260) conclude:

Table 7.
Hit-Rate Bias in Hartigan and Wigdor GATB data.

Table 7a. Joint frequency tables for White, Black, and total using GATB.

	White carpenters			Black carpenters			Total		
Rated:	*Lo*	*Hi*	Total	*Lo*	*Hi*	Total	*Lo*	*Hi*	Total
P	9	60	69	5	8	13	14	68	82
F	11	11	22	24	8	32	35	19	54
	20	71	91	29	16	45	49	87	136

Note: The criterion variable was Job Performance (Good or Poor) and the decision variable was the GATB (Pass or Fail). Cutpoints had been arbitrarily chosen. Data from Hartigan & Wigdor (1989).

Table 7b. Summary statistics for joint frequency tables using the GATB.

	White	Black	Total
N	91	45	136
BR	78.0	35.6	64.0
Q	75.8	28.9	60.3
%C	78.0	71.1	75.7
HR	84.5	50.0	78.2
FR	45.0	17.2	28.6
Validity	.49	.47	.51

Note: $HRB_{GATB} = 1.7$.

"At this point in history, it is certain that the use of the GATB without some sort of score adjustments would systematically screen out blacks, some of whom could have performed satisfactorily on the job. Fair test use would seem to require at the very least that the inadequacies of the technology should not fall more heavily on the social groups already burdened by the effects of past and present discrimination."

DIFFERENTIAL HIT-RATE BIAS FOR COMPOSITE PREDICTORS: NCAA DATA

We now return to the, as yet, unresolved question of differential hit-rate bias which we were unable to settle because the previous data sets were too small or weakened by confounds.

More recently, a fairly large data set has become available which does lend itself to settling the differential hit-rate bias question Crouse and Trusheim raised in their book. These data were collected under the auspices of the National Collegiate Athletic Association (NCAA) to examine the effects of proposed changes in college admission standards for student athletes. The main results are summarized in Table 8.

Table 8.
Hit-Rate Bias in NCAA Student-Athletic data.

Table 8a. Incremental Hit-Rate Bias when SAT is added to HSGPA.

Total Q	White *HR*s			Black *HR*s			*HRB*		
	GPA	SAT	Com-posite	GPA	SAT	Com-posite	GPA	SAT	Com-posite
89	96	99	98	84	78	76	1.14	1.26	1.29
81	92	95	94	70	60	63	1.32	1.58	1.51
69	85	88	88	50	39	47	1.68	2.27	1.87
60	77	80	81	42	28	37	1.82	2.87	2.12
51	69	73	72	34	23	24	2.03	3.13	3.02
39	55	60	59	21	16	13	2.62	3.89	4.51
29	43	48	47	12	9	9	3.48	5.11	5.01
19	30	32	32	4	5	5	7.76	6.89	6.94
10	16	16	17	2	2	2	6.69	7.02	10.75

Notes: N = 1,373 Whites and N = 303 Blacks who took the SAT; BR = 62% for Whites and BR = 43% for Blacks taking the SAT.

Table 8b. Incremental Hit-Rate Bias when ACT is added to HSGPA.

Total Q	White *HR*s			Black *HR*s			*HRB*		
	GPA	ACT	Com-posite	GPA	ACT	Com-posite	GPA	ACT	Com-posite
93	98	99	99	92	91	95	1.06	1.09	1.04
87	94	98	96	89	67	81	1.05	1.45	1.18
83	93	96	95	84	64	80	1.09	1.49	1.19
79	90	93	93	81	56	72	1.11	1.66	1.29
75	88	91	91	80	52	63	1.11	1.77	1.45
71	86	88	87	75	44	55	1.14	2.02	1.60
60	77	80	80	64	39	42	1.20	2.55	1.90
55	71	75	75	55	29	39	1.30	2.81	1.92
48	65	68	68	42	29	30	1.55	3.33	2.30
40	57	60	60	30	19	23	1.92	3.47	2.52
35	50	51	62	27	19	20	1.87	3.23	2.58
28	40	42	44	22	19	11	1.83	3.39	4.04
21	32	35	34	14	9	8	2.26	3.74	4.36
16	26	27	27	8	9	7	3.27	8.70	3.35
12	20	21	21	5	5	2	4.33	6.62	13.24

Notes: N = 948 Whites and N = 249 Blacks who took the ACT. BR = 52% for Whites and BR = 26% for Blacks.

Since a preliminary analysis revealed systematic differences in the validities and base-rates for students who took the SAT and students who took the ACT, we analyzed both data sets separately.

The hit-rates and bias ratios of the SAT are presented in Table 8a for the full range of admission quotas. For any given quota, Q, the SAT hit-rates exceed the HSGPA hit-rates for Whites. The reverse is true for Blacks. The hit-rates for the composite predictor are intermediate.

This result is portrayed graphically in Figure 2. Figure 2a gives the SAT hit-rate curves separately for Whites (upper 3 curves) and Blacks (lower 3 curves). The distance between the two innermost curves is the hit-rate difference for HSGPA alone. For a 60% quota, this difference is 77%-42% = 35%, corresponding to a bias-ratio of 1.82. For the SAT, the values are 80%-28% = 52%, corresponding to a *HRB* of 2.87. Concretely, if only HSGPA is used to predict college graduation, then, for every qualified black athlete admitted, roughly two qualified white athletes are admitted. If the admission decision is based on the

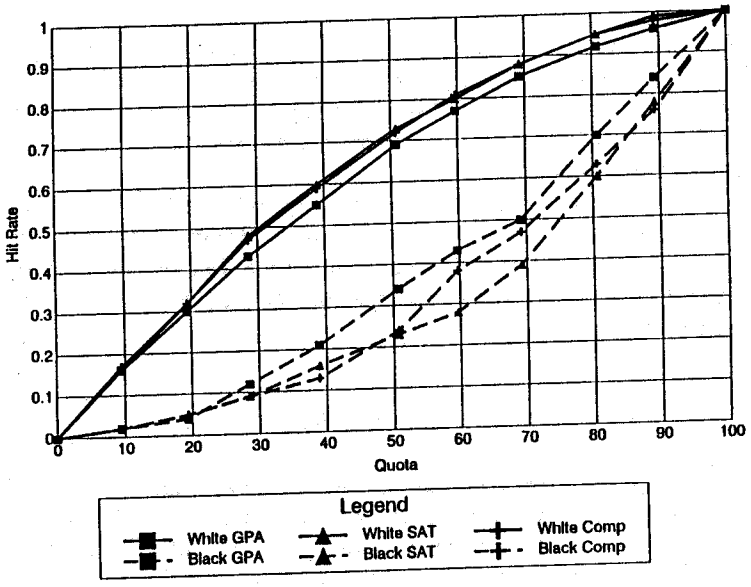

Figure 2a. Quota by hit rate for SAT.

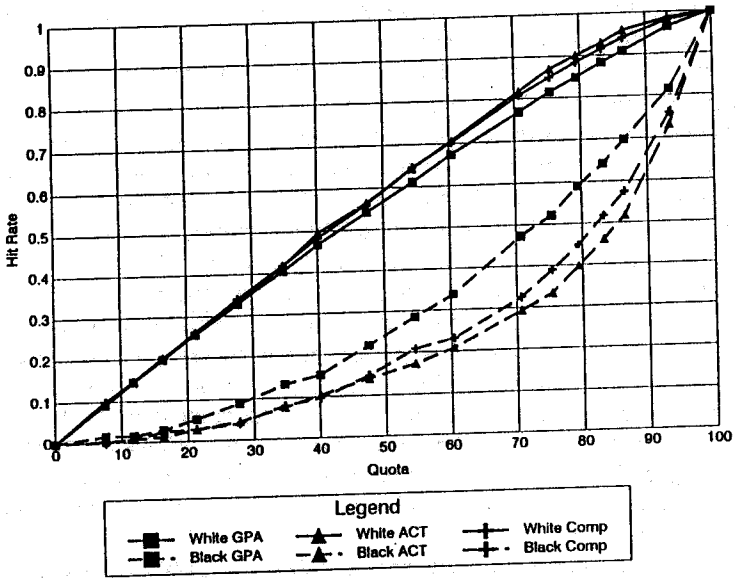

Figure 2b. Quota by hit rate for ACT.

SAT alone, then for every qualified Black admitted, three qualified Whites are admitted. The figures for the composite HSGPA+SAT are intermediate between these two extremes. The same general picture holds for the ACT (Table 8b, Figure 2b), except that the base-rates and validities are different.

Thus we find that the NCAA data independently corroborate Crouse and Trusheim's main conclusion that the addition of college entrance exams to high school record screens in unqualified Whites and screens out qualified Blacks, further aggravating already existing hit-rate bias against Blacks.

DISCUSSION

In this paper, we have attempted to contribute to an informed discussion of the bias problem in mental testing. Such discussions can only be fruitful after the term "bias" has been clearly defined, and something is known about its pervasiveness in the real world.

In the past, most discussions were deficient on both counts. We marvelled at the passions some contributors to a Special Issue on test bias in the *Journal of Educational Measurement* were able to generate in a virtual data vacuum. In contrast to Cronbach (1976, p. 31), we did not read this as a sign that "thinking about bias in selection has advanced rapidly in the past few years".

While we sympathize with Cole's definition of "test bias", neither she nor we ever believed that it is the only possible definition. If other scholars (e.g., Peterson & Novick, 1976) view other forms of test bias as more important, then they should adduce supporting data.

Only after the terms have been clearly defined and the relevant facts are known will it become possible to strike a rational balance between various imperfect policy alternatives in an attempt to reduce extant social injustices. Forty years ago, Meehl and Rosen (1955) already have pointed out the importance of judging the merit of a test not just in terms of single validity coefficients but in the context of quota and base rate. "... notably when the base rates of the criterion classification deviate greatly from a 50 per cent split, use of a test sign having slight or moderate validity will result in an *increase* of erroneous clinical decisions" (p. 210).

These early warnings went largely unheeded, partly perhaps, because at that time, illustrations of the interactions between classification rates and validities had remained eclectic and nonsystematic. Our explicit

expressions, equations 12 and 13, make it possible for the first time to assess the practical impact of these interactions in a systematic way.

ACKNOWLEDGEMENTS

We dedicate this paper to the memory of Allan Chase, author of *The Legacy of Malthus: The social costs of the new scientific racism*. We thank the McIntosh Foundation for financial support. We also gratefully acknowledge the assistance of Mr. Donovan Gay and Representative Cardiss Collins in obtaining the NCAA data discussed in this paper, and Professor Gordon Harrington for numerous constructive suggestions for improving the paper.

RÉSUMÉ

Nous présentons de nouveaux résultats sur une forme de biais dans un test qui semble avoir été largement ignoré antérieurement, à quelques exceptions près (notamment Cole, 1973 ; Hartigan et Wigdor, 1989). Nous l'appelons "biais du taux de succès" parce qu'il est défini comme la différence entre le taux de succès (c'est-à-dire la probabilité qu'un sujet de niveau requis réussisse le test) dans un groupe constitué de sujets à hautes performances et le taux de succès dans un groupe composé de sujets à basses performances. Typiquement, le test favorise le groupe à performances élevées. Contrairement à Cole (1973, nous nous sommes centrés sur des critères binaires, tels que l'obtention des diplômes. Dans une première partie théorique, nous présentons un théorème ("les limites du taux de succès") qui met en évidence le fait que l'augmentation des niveaux des prédicteurs n'est pas équivalente à l'augmentation des niveaux des critères, comme certains le croient. Au contraire, cela accroît spécifiquement le biais de taux de réussite. Ensuite, nous dérivons et nous présentons sous forme de tableau une simple approximation pour estimer les pourcentages de réussite en fonction de la validité, du taux de base et du quota d'admission. Dans la partie empirique de l'article, nous évaluons l'ampleur du biais du taux de succès par une nouvelle analyse d'ensembles de données utilisant le SAT, l'ACT et le GATB. Enfin, nous examinons comment l'addition des notes aux tests et des résultats au lycée affecte le biais du taux de succès pour la prédiction de réussite au diplôme et l'augmente.

REFERENCES

Cole, N. S. (1973) Bias in selection. *Journal of Educational Measurement, 10,* 237-255.

Coleman, J. S. (1990). *Equality and achievement in education.* Boulder, CO: Westview Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Harcourt Brace Jovanovich College Publishers.

Cronbach L. J. (1976). Equity in selection – Where psychometrics and political philosophy meet. *Journal of Educational Measurement, 10,* 31-42.

Crouse, J., & Trusheim, D. (1988). *The case against the SAT.* Chicago, IL: The University of Chicago Press.

Donlon, T. F. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: CEEB.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in Employment Testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *Journal of Educational Psychology, 59,* 375-380.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1983). The definition of intelligence and factor-score indeterminacy. *The Behavioral and Brain Sciences, 6,* 313-315.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216.

Novick, M. R., & Petersen, N. S. (1976). Towards equalizing educational and employment opportunity. *Journal of Educational Measurement, 13,* 77-88.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3-30.

Sawyer, R. L., Cole, N. S., & Cole, J. W. L. (1976). Utilities and the issue of fairness in a decision theoretic model for selection. *Journal of Educational Measurement, 13,* 59-76.

APPENDIX 1.

Theorem (Hit-Rate Bounds).

A hit-rate proportion never exceeds the ratio of admission quota over base-rate:

$$hr \leq q/br. \tag{15}$$

Proof: From the definitions 4, 1, 2 of Table 1:

$$\begin{aligned} \text{hr} &:= HR/100 = TP/(TP+FN) \leq (FP+TP)/(TP+FN) \\ &= [(FP+TP)/TOTAL]/[(TP+FN)/TOTAL] = q/br. \end{aligned} \tag{16}$$

APPENDIX 2.
Approximations of *HR* and *FR* based on linearized cumulative normal.

The approximations for *HR* and *FR* in terms of *BR*, *Q*, and *r*, given in the paper, rely on the fact that the cumulative normal distribution $N(z)$ is approximately linear around (0, .5). For the slope of a straight line through (0, .5) and (1, $N(1)$), one finds that $N(1) - .5 = .34$, which we approximate by 1/3. This gives the approximations

$$p = N(z) = .5 + z/3, \text{ and its inverse, } z = N^{-1}(p) = 3(p - .5), \quad (17)$$

presumed to be valid in the middle range. Now recall that

$$r = d' \sqrt{br(1-br)}, \text{ where} \quad (18)$$

$$d' := (\mu_q - \mu_u)/\sigma. \quad (19)$$

Since the definitions of *HR* and *FR* in Table 1 imply

$$1-hr = N[(c-\mu_q)/\sigma], \; 1-fr = N[(c-\mu_u)/\sigma], \quad (20)$$

one finds, on applying $N^{-1}(.)$ in (17), that

$$d' = N^{-1}(1-fr) - N^{-1}(1-hr) = 3(1-fr-.5) - 3(1-hr-.5) = 3(hr-fr) \quad (21)$$

which simplifies (18) to

$$r = 3(hr-fr)\sqrt{br(1-br)}. \quad (22)$$

To get rid of one of the two unknowns, say *fr*, we can use the definition of *Q* in Table 1:

$$q := fp+tp = fr*(1-br) + hr*br \quad (23)$$

which gives

$$fr = (q-hr*br)/(1-br). \quad (24)$$

Substituting this expression for *fr* in eq. (22) yields, after some simplification,

$$hr = q + r\sqrt{(1-br)/br}/3 \quad (25)$$

as an estimate of *hr* explicitly in terms *br*, *q*, *r*. Conversely, on solving *q* for *hr* in eq. (23) and substituting the result into eq. (22) for *r*, one obtains an approximation for *fr*:

$$fr = q - r\sqrt{br/(1-br)}/3. \quad (26)$$

In passing, note that the ratio,

$$(q-fr)/(hr-q) = (br)/(1-br) \quad (27)$$

is independent of the validity *r*. On the other hand, the product

$$(q-fr)(hr-q) = r^2/9 \quad (28)$$

can be used to estimate the validity as

$$r = 3\sqrt{(q-fr)(hr-q)} \quad (29)$$

without any need to consult normal tables. Computer simulations showed that, for the parameter range indicated above, these approximations for *hr* and *fr* produce the true values within .02 in more than 92% of all cases, and within .04 in 97%. For $r < .5$, only 1 out of 672 discrepancies exceeded .06.

Appendix 3.
Eliminating Hit-Rate Bias.

To eliminate Hit-Rate Bias, one has to choose two different cut-offs, c_A for the Advantaged Group and c_D for the Disadvantaged Group, such that

$$HR(\text{Advantaged}) = HR(\text{Disadvantaged}) \tag{30}$$

For binary criteria and normal predictors with common variance, the two cut-offs are given by the following:

$$c_D = c_A - \mu_{Aq} + \mu_{Dq} \tag{31}$$

where μ_{Dq} and μ_{Aq} are the conditional means for the qualified subjects in each group.
Proof: From

$$\begin{aligned} &1 - HR(\text{Advantaged} = N[(c_A - \mu_{Aq})/\sigma] = N[(c_D - \mu_{Dq})/\sigma] \\ &= 1 - HR(\text{Disadvantaged}) \end{aligned} \tag{32}$$

the result follows at once since $N(.)$ is everywhere invertible.

Received 30 May, 1995
Accepted 10 October, 1995