Psychological Assessment
2008, Vol. 20, No. 3, 195–205

1040-3590/08/$12.00 DOI: 10.1037/a0013175

Predictor Combination in Binary Decision-Making Situations

Robert E. McGrath
Fairleigh Dickinson University

Professional psychologists are often confronted with the task of making binary decisions about individuals, such as predictions about future behavior or employee selection. Test users familiar with linear models and Bayes's theorem are likely to assume that the accuracy of decisions is consistently improved by combination of outcomes across valid predictors. However, neither statistical method accurately estimates the increment in accuracy that results from use of additional predictors in the typical applied setting. It was demonstrated that the best single predictor often can perform better than do multiple predictors when the predictors are combined using methods common in applied settings. This conclusion is consistent with previous findings concerning G. Gigerenzer and D. Goldstein's (1996) "take the best" heuristic. Furthermore, the information needed to ensure an increment in fit over the best single predictor is rarely available.

Keywords: linear regression, Bayes's theorem, predictive power, clinical decision making, heuristics

Professional psychologists are often faced with the practical task of classifying people into one of at least two categories. Examples include whether to implement a treatment, whether to hire a person, or whether each of a series of diagnoses applies to an individual. The practical need to dichotomize cases often exists even when the variables used to make the decision are inherently dimensional, though the dichotomization of dimensional data is a problematic undertaking from a formal statistical perspective (e.g., Dwyer, 1996; MacCallum, Zhang, Preacher, & Rucker, 2002; but see Farrington & Loeber, 2000). This paradox highlights the importance of considering both practical considerations and formal statistical issues when one intends data to reveal something about real-world practices (McGrath, 2001).

For example, statistical methods familiar to applied psychologists tend to suggest that the accuracy of predictions is consistently improved by combination of multiple valid predictors. This article presents evidence that this is not necessarily the case, particularly when the pragmatics of predictor combination for purposes of classifying individuals in applied settings are taken into consideration. It demonstrates that under certain common circumstances, psychologists may be better served by basing their classification on the best single predictor and ignoring additional sources of information.

Statistical Methods Relevant to Predictor Combination

Two statistical methods commonly familiar to psychologists—linear regression and Bayes's theorem—can be taken to suggest that additional predictors will consistently improve prediction. This section provides a brief review of each model and discusses their limitations as a basis for such a conclusion in applied settings.

Linear Regression

Linear regression involves the identification of an optimal set of weights for generation of an additive composite of predictors. It is one of various computationally intensive methods that have been developed for combining data from multiple predictors. Other such methods are available—for example, classification trees, discriminant function analysis, neural networks, cluster analysis, boosting, and methods based on receiver operating characteristic curves (e.g., Friedman, Hastie, & Tibshirani, 2000; Hogarth & Karelaia, 2005; Swets, Dawes, & Monahan, 2000)—but linear regression is clearly the method most familiar to psychologists and so has the greatest influence on psychologists' beliefs about the advantages of multiple predictors.

The use of linear regression for combining information across predictors in applied settings is often discussed in terms of incremental validity (Hunsley & Meyer, 2003; Sechrest, 1963). Incremental validity may be defined as the extent to which additional predictors enhance the proportion of overlapping variance with the criterion. Some form of hierarchical regression is the standard statistical method for evaluation of the degree of incremental validity provided by additional predictors, and some variant of the correlation coefficient usually provides the corresponding effect-size index.

Because this article focuses on dichotomous decisions, subsequent discussion of linear regression focuses on logistic regression. Table 1 contains four examples of results from incremental validity studies that used logistic regression. Various statistics that provide an analogue to the correlation coefficient are available for logistic regression. SAS offers two, the Cox and Snell (1989) generalized coefficient of determination (R^2) and an adjusted version that corrects for possible range restriction in R^2 (max R^2; Nagelkerke, 1991). Each example provides the incremental validity of adding two predictors, B and C, over predictor A. Subsequent

I am deeply grateful to Robyn Dawes, Joshua Dwire, Lewis Goldberg, William Grove, and John Hunsley for their comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Robert McGrath, School of Psychology T-WH1-01, Fairleigh Dickinson University, Teaneck, NJ 07666. E-mail: mcgrath@fdu.edu

Table 1
Examples of Hierarchical Logistic Regression Incremental Validity Analyses

$\Delta R^2_{\text{Y.A}}$	max $\Delta R^2_{\text{Y.A}}$	$R^2_{\text{Y.ABC}}$	max $R^2_{\text{Y.ABC}}$	$\Delta R^2_{\text{Y.ABC}}$	Δmax $R^2_{\text{Y.ABC}}$
.07	.11	.13	.18	.05	.08
.17	.24	.28	.40	.12	.16
.18	.25	.37	.49	.18	.25
.32	.43	.35	.47	.03	.05

Note. R^2 is the generalized coefficient of determination (Cox & Snell, 1989). Max R^2 is an alternate version adjusted for possible range restriction (Nagelkerke, 1991). $\Delta R^2_{\text{Y.ABC}}$ is the difference between $R^2_{\text{Y.ABC}}$ and. $\Delta R^2_{\text{Y.A.}}$

discussion focuses on cases of three predictors but at times addresses issues of two predictors.

In the first example in Table 1, adding *B* and *C* to *A* increases the proportion of the variance of the criterion predicted by .06 according to the generalized coefficient and by .07 according to the adjusted coefficient. An important mathematical attribute of both statistics is that the coefficient cannot decrease as more predictors are added. That is, the multiple correlation for a set of predictors will always be at least equal to that of any subset of the predictors included in the set. Including additional valid predictors always enhances prediction (or at least does no harm). This attribute can foster the belief that, when the costs of additional testing are minimal and the results of linear regression can be considered reliable (i.e., shrinkage has already been accounted for), it is always desirable to increase the number of valid predictors.

Despite its familiarity, linear regression is rarely used as a combination method in applied settings, for practical reasons. Consider the conditions that must be met before linear regression can be used as the basis for binary decisions:

1. A sample that is sufficiently large relative to the number of predictors must be gathered to allow derivation of reliable weights and a cut score for the predicted scores.

2. Any changes in the set of predictors will require a new set of weights.

3. For optimal fit, future cases must reflect the same population as does the derivation sample.

4. In high-stakes decision-making situations, application of the combination method to the individual may need to be accomplished quickly.

It has been argued that in the case of multiple regression, the first condition can be avoided by the method called equal weighting or tallying (Dawes & Corrigan, 1974; Hogarth & Karelaia, 2005; Wainer, 1976). This method involves weighting those predictors positively correlated with the criterion by 1 and those predictors negatively correlated with the criterion by -1 after they have been standardized. Equal weighting can produce results superior to multiple regression under circumstances in which shrinkage is possible. However, application of this strategy to the case of dichotomous decisions would still require identification of an optimal cut score for the weighted combination of predictors and standardizing statistics; such a requirement reintroduces the need for a sizable derivation sample.

The second condition is unrealistic in applied settings in which the battery of predictors is tailored to the respondent on the basis of variations in the goals of the assessment, time constraints, respondent limitations, or issues of cost. The third condition is an untestable assumption in the individual case, and the fourth condition suggests that the application of linear models may be particularly unwieldy in precisely those settings in which accuracy in prediction is most important. Given the practical obstacles, test users almost always rely on less intensive methods of data combination in applied settings.

Bayes's Theorem

A more practical option for combining predictors in applied settings is referred to here as the vote-counting heuristic. If a predictor is not inherently dichotomous, it is first dichotomized with a cut score derived specifically for that predictor.[1] The decision is based on the majority outcome across predictors. This heuristic is mathematically equivalent to tallying but has two modifications that make implementation of the heuristic more practical:

1. Each predictor is dichotomized as $X- = 0$ and $X+ = 1$ prior to aggregation.

2. The cut score for the aggregate is pragmatic rather than optimized; it is based on the value that is half the maximum possible score.

The vote-counting heuristic should not be considered to be specific to psychological evaluation. It is generally applicable to settings in which decision making is based on standardized data-gathering procedures. For example, medical professionals often dichotomize outcomes on dimensional indicators (e.g., body temperature or white blood cell count) according to whether they fall within the normal or abnormal range and make a judgment based on the preponderance of evidence.

Bayes's theorem provides a second statistical method familiar to most psychologists that can be used in conjunction with the vote-counting heuristic to estimate the improvement in accuracy resulting from use of multiple predictors. As background to a discussion of the application of Bayes's theorem to vote counting, a classi-

[1] This practice violates a common recommendation for the use of local cut scores over global or standard cut scores (e.g., Meehl & Rosen, 1955). The recommendation is largely ignored in applied settings because it is often considered impractical to generate local cut scores, for the same reasons that impede applied use of linear regression. Also, local cut scores create the uncomfortable possibility that a person classified one way in one setting will merit reclassification in a subsequent setting. For example, an individual declared suicidal in an inpatient setting might not meet criteria for classification as suicidal at a later group-home placement because of a change in local cut score, even though the test outcome is the same. Such practices would be extremely problematic from a liability perspective. It is also worth noting that Hsu (1985) found that local cut scores are not necessarily superior to global cut scores, but the truth is that resistance to local cut scores is more practical than it is statistical.

		Population (Y)		
		Targeted ($Y+$)	Complement (Y-)	
Test Outcome ($X = A, B, C$)	Positive ($X+$)	$p(X+Y+)$ Sens: $p(X+\|Y+)$ PPP: $p(Y+\|X+)$	$p(X+Y\text{-})$ 1 – Spec: $p(X+\|Y\text{-})$ 1 – PPP : $p(Y\text{-}\|X+)$	$p(X+)$
	Negative (X-)	$p(X\text{-}Y+)$ 1 – Sens: $p(X\text{-}\|Y+)$ 1 – NPP: $p(Y+\|X\text{-})$	$p(X\text{-}Y\text{-})$ Spec: $p(X\text{-}\|Y\text{-})$ NPP: $p(Y\text{-}\|X\text{-})$	$p(X\text{-})$
		BR: $p(Y+)$	1 – BR: $p(Y\text{-})$	CF: $p(X+Y+) +$ $p(X\text{-}Y\text{-})$

Figure 1. The symbols displayed are used to represent various outcomes of a decision-making process and the associated probabilities. The upper right and lower left cells indicate various ways to present the probabilities associated with incorrect decisions; the upper left and lower right cells indicate the probabilities of correct decisions. Sens = sensitivity; PPP = positive predictive power; Spec = specificity; NPP = negative predictive power; BR = base rate; CF = correct fraction.

fication table (see Figure 1) is used to introduce some concepts from probability theory relevant to the case in which both a predictor and a criterion are dichotomous. The criterion variable Y is a dichotomous indicator of whether an individual falls in the targeted ($Y+$) or complement ($Y-$) population (e.g., whether the person meets or does not meet standards for employment). This criterion is predicted by dichotomized indicators $X = A$, B, and C, on which a respondent may produce a positive outcome ($X+$), predictive of membership in the targeted population, or a negative outcome ($X-$). The probability of belonging to the targeted population can be referred to as $p(Y+)$, though the more familiar term *base rate* (BR) is used here instead. The probability of being simultaneously a member of the targeted population and negative on predictor B, which would be a prediction error, is symbolized $p(B-Y+)$. The conditional probability of a positive outcome on B among members of the targeted population is symbolized $p(B+|Y+)$. The proportion of predictions that are correct is frequently referred to in the psychological literature as the hit rate, after Meehl and Rosen (1955), but this term has a different meaning in the general statistical literature, so the more contemporary term *correct fraction* (CF) is used instead.

Table 2 provides computational formulas for several statistics relevant to the analysis of prediction in 2 × 2 tables of this type, often referred to as diagnostic efficiency statistics. Sensitivity (Sens) is the probability of a positive test result given membership in the targeted population, or $p(X+|Y+)$. Specificity (Spec) is the probability of a negative result within the complement population, or $p(X-|Y-)$. These statistics reflect the probability of a correct decision within each population.

Positive predictive power (PPP), also referred to as the positive predictive value, is the probability the individual is a member of the targeted population, given a positive result, $p(Y+|X+)$, and negative predictive power (NPP) is the corresponding statistic concerning correct outcomes among individuals who are negative on the predictor, $p(Y-|X-)$. These statistics reflect the probability of a correct decision within each test outcome.

Sens and Spec have a statistical advantage over predictive power statistics in terms of sampling variability. Sens and Spec vary as a function of BR, at least under certain circumstances having to do with the causal model explaining the relationship between predictor and criterion (Choi, 1997). However, they tend to vary less than PPP and NPP, because predictive power is a direct function of the BR, Sens, and Spec. Consider the following restatement of the formulas for PPP and NPP for a given predictor, X:

$$PPP_X = \frac{BR \times Sens_X}{(BR \times Sens_X\} + \{(1 - BR) \times (1 - Spec_X)\}} \qquad (1)$$

$$NPP_X = \frac{(1 - BR) \times Spec_X}{\{(1 - BR) \times Spec_X\} + \{BR \times (1 - Sens_X)\}} \qquad (2)$$

These formulas indicate that, even if Sens and Spec remain constant, PPP will increase and NPP will decrease as the BR increases (see Meehl & Rosen, 1955). As a result, they demonstrate substantially greater sampling variability as a function of BR than do Sens or Spec (Brenner & Gefeller, 1997).

Even so, predictive power is often of greater interest than are Sens and Spec in applied settings, because the results are directly relevant to circumstances in which a conclusion must be drawn about the respondent's population membership (Y) on the basis of test results (X). As suggested in the preceding discussion of dichotomizing dimensional indicators, practical considerations are not always coincident with the optimal statistical approach.

Equations 1 and 2 are also interesting because they represent restatements of Bayes's theorem in terms of the symbols introduced here. From a Bayesian perspective, BR can be treated as the

Table 2
Computational Formulas for Diagnostic Efficiency Statistics

Statistic	Probability represented	Formula
Sens	$p(X+\|Y+)$	$\frac{p(+Y+)}{p(X+Y+) + p(X-Y+)} = \frac{p(X+Y+)}{p(Y+)}$
Spec	$p(X-\|Y-)$	$\frac{p(X-Y-)}{p(X-Y-) + p(X+Y-)} = \frac{p(X-Y-)}{p(Y-)}$
PPP	$p(Y+\|X+)$	$\frac{p(X+Y+)}{p(X+Y+) + p(X+Y-)} = \frac{p(X+Y+)}{p(X+)}$
NPP	$p(Y-\|X-)$	$\frac{p(X-Y-)}{p(X-Y-) + p(X-Y+)} = \frac{p(X-Y-)}{p(X-)}$
CF	$p(X+Y+$ or $X-Y-)$	$p(X+Y+) + p(X-Y-)$

Sens = sensitivity; Spec = specificity; PPP = positive predictive power; NPP = negative predictive power; CF = correct fraction.

(a)

A	Y +	Y -
+	$p(A+Y+) = .07$	$p(A+Y-) = .27$
-	$p(A-Y+) = .03$	$p(A-Y-) = .63$
	$p(Y+) = .10$	$p(Y-) = .90$

$p(Y+|A+) = .07/(.07 + .27) = .21$

B	Y +	Y -
+	$p(B+Y+) = .14$	$p(B+Y-) = .24$
-	$p(B-Y+) = .06$	$p(B-Y-) = .56$
	$p(Y+) = .21$	$p(Y-) = .79$

$p(Y+|A+B+) = .14/(.14 + .24) = .38$

C	Y +	Y -
+	$p(C+Y+) = .26$	$p(C+Y-) = .19$
-	$p(C-Y+) = .11$	$p(C-Y-) = .44$
	$p(Y+) = .38$	$p(Y-) = .62$

$p(Y+|A+B+C+) = .26/(.26 + .19) = .59$

(b)

A	Y +	Y -
+	$p(A+Y+) = .07$	$p(A+Y-) = .27$
-	$p(A-Y+) = .03$	$p(A-Y-) = .63$
	$p(Y+) = .10$	$p(Y-) = .90$

$p(Y+|A+) = .07/(.07 + .27) = .21$

B	Y +	Y -
+	$p(B+Y+) = .14$	$p(B+Y-) = .24$
-	$p(B-Y+) = .06$	$p(B-Y-) = .56$
	$p(Y+) = .21$	$p(Y-) = .79$

$p(Y+|A+B-) = .06/(.06 + .56) = .10$

C	Y +	Y -
+	$p(C+Y+) = .07$	$p(C+Y-) = .27$
-	$p(C-Y+) = .03$	$p(C-Y-) = .63$
	$p(Y+) = .10$	$p(Y-) = .90$

$p(Y+|A+B-C+) = .07/(.07 + .27) = .21$

Figure 2. Two examples of the iterative application of Bayes's theorem to the estimation of PPP. (a) Computing the probability of membership in the targeted population ($Y+$) if A, B, and C are all positive. A positive outcome on A increases the probability of $Y+$ from .10 to .21, of a positive outcome on B from .21 to .38, and of a positive outcome on C from .38 to .59. (b) Computing the probability of membership in the targeted population ($Y+$) if A and C are positive but B is negative. Notice that the results for A and B cancel each other.

prior probability of membership in the targeted population, that is, the probability of membership in the absence of additional information from the indicator. PPP represents the corresponding posterior probability (i.e., the probability of membership in the targeted population after a positive outcome has been found on a predictor). Similarly, $p(Y-)$ is the prior probability of membership in the complement population and NPP is the posterior probability, given a negative outcome on X.

One implication of Bayes's theorem is that, if X is a valid predictor of Y, a positive outcome on X will result in a posterior probability of membership in $Y+$ (PPP) that is greater than the prior probability (BR). In other words, a positive outcome should increase one's confidence that the individual is a member of the targeted population.[2] A reasonable extrapolation is that the iterative use of multiple predictors should incrementally improve PPP to the extent that the respondent produces positive results on each predictor (Waller, Yonce, Grove, Faust, & Lezenweger, 2006). Again, the statistic can be taken as implying that more is almost always better.

The application of Bayes's theorem to the case of multiple dichotomous predictors is demonstrated in Figure 2. In these examples, BR = .10 and $Sens_X = Spec_X = .70$ for all three predictors. The figure's left panel demonstrates the results for case $A+B+C+$, in which the respondent generated positive outcomes on all three predictors. A positive outcome on A suggests that the probability of membership in the targeted population is .21. When this value is used as the new prior probability of membership in the population, a positive outcome on B raises that value further to .38. A third positive outcome on C raises the posterior probability of membership in the targeted population to .59.

Three comments are worth making about the application of Bayes's theorem to vote counting. First, finding that all three outcomes were positive justifies assigning greater confidence to the assertion that the respondent is a member of the targeted population than does finding one positive outcome, but it would probably surprise many applied test users that there is still such a sizable probability (.41) that the respondent is not a member of that population. This finding reflects the low initial BR, so that even a substantial increase in the probability of membership in the targeted population does not approach certainty. The tendency to overestimate the confidence afforded by test results when the initial BR is ignored has been noted many times (e.g., Meehl & Rosen, 1955; Wiggins, 1972), but it continues to bedevil psychological (and medical) practice.

Second, it is worth noting that the enhancement of diagnostic efficiency resulting from the use of multiple predictors varies depending on the pattern of outcome across predictors. The right panel of Figure 2 represents the case in which predictor B is inconsistent with the other two predictors. As these three predictors are equivalent in validity, the divergent outcome for B offsets exactly the increment in PPP due to C, so the overall result is no better than that from A alone. When the predictors differ in their

[2] Following Meehl and Rosen (1955), expository writing on diagnostic efficiency for psychologists often notes that in cases where the BR is extremely low or high, use of the test may result in a lower CF than does "betting the base rate" (i.e., always predicting that the respondent is a member of the modal population; Hsu, 1985; Waller, Yonce, Grove, Faust, & Lezenweger, 2006). Although this is technically true, betting the BR is unacceptable in applied settings for very practical reasons. Consider the potential consequences for a clinical psychologist who refuses to predict that anyone is at risk of committing suicide because, given the very low BR for suicide, this practice results in the best CF.

Sens and Spec, each pattern of outcomes can be associated with a unique value for PPP or NPP.

Finally, the estimate of predictive power resulting from the application of Bayes's theorem to the vote-counting heuristic is very likely to be wrong. An extreme example demonstrates why this would be the case. Suppose that predictors *A*, *B,* and *C* all correlate perfectly. If so, the information about *Y* provided by each predictor is redundant and the posterior probability of membership in the targeted population is still only .21, even if all three test outcomes are positive. The iterative application of Bayes's theorem produces inaccurate results because it ignores dependencies among the predictors (Katsikopoulos & Martignon 2006; see also Waller et al., 2006).[3]

Direct Computation of Probabilities

A more accurate method of determining the effectiveness of the vote-counting heuristic involves computation of the overall probability of a correct decision, given a positive outcome (PPP) or a negative outcome (NPP), on the basis of the majority of test outcomes. For example, the overall PPP in the three-predictor case can be generated by determining the proportion of cases in which at least two of the predictors are positive that involve members of the targeted population.

This approach was first considered in the context of the two-predictor case. This case immediately presents a problem for the vote-counting heuristic. If both predictors are positive or both predictors are negative, the majority decision is clear. The decision becomes uncertain when *A* is positive and *B* is negative or vice versa. One reasonable heuristic for breaking the tie suggests a bias in favor of the predictor with the higher level of criterion-related validity; this option has been described, for example, by Ganellen (1996, pp. 72–73). That is, if $r_{YA} > r_{YB}$, the decision is positive if both *A* and *B* are positive or *A* alone is positive and is negative if both *A* and *B* are negative or *A* alone is negative.[4] Although this rule seems to be intuitively reasonable and may well reflect what test users do in applied settings, an analysis of the implications of this strategy for PPP and NPP produces a surprising result. Assume that *A* is the more valid predictor. If so, the heuristic suggests that if *A* is positive the decision based on both predictors will always be positive, whereas if *A* is negative the two-predictor decision will always be negative. In other words, the diagnostic efficiency of combining *A* and *B* is no different than is the diagnostic efficiency of using *A* alone. This suggestion may seem counterintuitive, because the PPP for the case in which both *A* and *B* are positive should be greater than the PPP for either *A* or *B* alone, if it is assumed that *A* and *B* do not correlate so highly that they are essentially redundant. However, this gain is offset by the lower PPP for the case in which *A* is positive but *B* is negative or vice versa. The same pattern holds for NPP and HR. The point is demonstrated mathematically in the Appendix.

If the decision is the same regardless of the outcome on *B*, then *B* adds nothing but psychological comfort to the overall predictive power of the assessment. What seemed to be a reasonable, relatively complete, and practically useful heuristic for the integration of results from two predictors offers no incremental validity over the predictor that is awarded dominance for purposes of tie breaking. This conclusion holds even if the second predictor demonstrates incremental validity according to hierarchical regression.

Application of the vote-counting heuristic is more straightforward in the three-predictor case. One reasonable option would be to declare the individual positive when at least two out of three predictors are positive and negative when at least two of the three predictors are negative. An important variant of this heuristic is commonly used in medical diagnostics, when two tests are administered (or the same test is administered twice) and a third is administered as a tiebreaker if they disagree.

The analytic development of this heuristic is provided in the Appendix, and the results are equally unintuitive. If predictor *A* is more valid than predictors *B* and *C*, the analysis demonstrates it would not be unreasonable to find that the PPP, NPP, and HR for *A* alone are greater than are the corresponding values based on combining all three predictors. More specifically, if $p(A{+}B{-}C{-}Y{+}) > p(A{-}B{+}C{+}Y{+})$, or $p(A{-}B{+}C{+}Y{-}) > p(A{+}B{-}C{-}Y{-})$, or both are true, then *A* by itself will outperform the vote-counting heuristic on the basis of all three predictors. This finding suggests an alternative heuristic for applied decision making, which is referred to as the best single predictor (BSP).

The conclusion that the BSP can outperform multiple predictors echoes similar conclusions drawn concerning Gigerenzer and Goldstein's (1996) "take the best" (TTB) heuristic, which they presented as one method people use for comparison of pairs of objects or options when time and/or information is limited. Because TTB was developed as a model of naturalistic decision making and BSP is proposed as a model for decision making in more formal testing situations, TTB differs from BSP in several important ways. TTB specifically describes a method for predicting ordinal placement within pairs of objects, whereas BSP is a method for predicting dichotomous placements of objects one at a time. Gigerenzer and Goldstein proposed that the first cue used in TTB is always recognition of the options and that the tendency is to reject those options unfamiliar to the decision maker. Presumably, all tests used will be familiar to the test user. Finally, TTB incorporates the possibility that in a particular comparison, the best single environmental cue may not provide a clear preference for one object over the other, in which case the decision maker is expected to proceed through additional cues until a decision is possible. In contrast, BSP relies on a dichotomous predictor, so placement on the basis of a single predictor is always possible.

Despite the differences, there are enough similarities that evidence concerning the accuracy of TTB should provide some support for the potential of BSP. In circumstances in which information is limited, TTB is often as effective as or more effective than methods based on linear regression and Bayesian methods as a basis for decision making (Gigerenzer, Czerlinski, & Martignon,

[3] In recent years, the study of Bayesian networks has allowed researchers to consider dependencies among predictors when they estimate posterior probabilities (e.g., Almond, DiBello, Moulder, & Zapata-Rivera, 2007). However, this method is even more computer intensive than is linear regression. Furthermore, as the focus here is on those statistical models that contribute to the presumption among psychologists that more predictors is always better, Bayesian networks will not be considered further.

[4] This heuristic is still technically incomplete, as it ignores the case in which $r_{YA} = r_{YB}$. This case is probably rare enough that it deserves to be relegated to a footnote, but it could still be addressed by, for example, randomly awarding precedence to *A* or *B*. So long as one predictor is treated as dominant, the conclusions drawn in the text remain valid.

2002; Hogarth & Karelaia, 2005; Martignon & Laskey, 1999). To evaluate whether the same was true for BSP, the researcher created a series of data simulations to compare the various approaches that have been reviewed.

Generating Simulations

The simulations were created with an algorithm intended to sample from the universe of combinations of dichotomous predictors and criteria that could reasonably occur in well-designed applied settings. Each simulation was based on a set of 16 probabilities drawn from two $2 \times 2 \times 2$ contingency tables; each table represented one of the two criterion populations. The first cell of the first table represented the probability that all three predictors were positive in the targeted population, or $p(A+B+C+|Y+)$. The other seven cells in the table reflected conditional probabilities for the other possible combinations of predictor outcomes, given membership in the targeted population. The second table reflected conditional probabilities for the complement population.

The probability for each cell was iteratively increased from 0 to .80 by .10. The BR was similarly set to $p(Y) = .02$, .10, .30, and .50. BR values $> .50$ were omitted, as they would have simply mirrored the results for smaller BRs, with PPP and NPP switched. When the BR and the 16 conditional probabilities were used, it was possible to compute the diagnostic efficiency statistics for each predictor, the correlation between each predictor and the criterion, and the correlations between the predictors. Simulations were eliminated if they failed to meet any of the following criteria:

1. The sum of the eight probabilities within each of the two tables equaled 1.0.
2. The sum of the probabilities that determined the Sens for each of the three predictors fell within the interval $.50 \leq Sens_X \leq .90$.
3. For each of the three predictors, $.50 \leq Spec_X \leq .90$.
4. For each predictor, either $Sens_X$ or $Spec_X$ was $> .50$.
5. Correlations with the criterion fell in the interval $.10 \leq r_{YX} \leq .70$.
6. Correlations between predictors fell in the interval $0 \leq r_{XX} \leq .70$.
7. $r_{YA} \geq r_{YB}$ and $r_{YB} \geq r_{YC}$.

The first criterion restricted the simulations so they were consistent with the mathematical requirements for conditional probability tables. Criteria 2–6 were used to limit the simulations to the types of outcomes likely to occur in well-designed applied settings. The last criterion assured that predictors were ordered from most to least correlated with the criterion. This process generated 186,301 unique simulations.

For each simulation, logistic regression was computed for predictor *A* and for all three predictors with SAS Version 9.1. In addition to the correlational statistics described earlier, the researcher generated a classification table that assumed a prior probability equal to BR. Results from this table were used to generate estimates of PPP, NPP, and CF for the case in which three predictors are combined.

The Bayesian estimate of CF for three predictors was computed with procedures described by Waller et al. (2006). To generate an overall Bayesian estimate of PPP, the researcher used the same procedures to compute the PPP for each combination of test outcomes that would lead to a positive prediction according to the vote-counting heuristic (at least two of three predictors positive). These PPPs were weighted by the probability of that combination occurring and were averaged. The same process was used to generate the Bayesian estimate of NPP.

Equations A7, A9, and A11 (see the Appendix) were used to directly compute diagnostic efficiency statistics for the three-predictor case. Finally, to evaluate the BSP heuristic, the researcher determined the diagnostic efficiency of predictor *A* alone.

Results

Descriptive statistics for the simulations may be found in Table 3. Results are presented for regression-based correlational statistics. The table also contains diagnostic efficiency statistics generated with four methods: the BSP heuristic and logistic regression, as well as the application of Bayes's theorem and direct computation to the vote-counting heuristic. For logistic regression, the mean value for the two correlational statistics is provided for predictor *A* and for all three predictors combined.

The findings were generally consistent with expectation. The addition of *B* and *C* increased the mean proportion of variance predicted for both the generalized coefficient and the adjusted version. The proportion of variance accounted for increased as the BR approached .50 (see McGrath & Meyer, 2006). For the BSP heuristic, correlations between CF and correlational statistics were higher when the latter was based on one predictor rather than three, whereas the reverse was true for the other classification methods based on three predictors.

One unexpected finding was the relatively low correlations between CF based on logistic regression and the correlational statistics. These correlations were substantially lower than were those for CF estimates based on the BSP heuristic or Bayes's theorem. This finding seemed to be a function of the differential effects of BR on the correlational statistics versus CF derived via logistic regression. After BR had been partialed, the correlations between CF and the correlational statistics increased to a level consistent with those for direct computation. The finding suggests that expectations about the value of additional predictors derived from literature on incremental validity may not generalize even to regression-based diagnostic efficiency statistics.

In the remainder of Table 3, comparisons are organized by diagnostic efficiency statistic.[5] As expected, mean statistics derived with logistic regression and Bayes's theorem were consistently higher than were those based on direct computation or the BSP heuristic. However, their improvement over BSP was not substantial, and mean values were consistently larger for BSP than

[5] In 5,523 simulations with low BR, logistic regression did not identify any cases as positive. Statistics presented in Tables 3 and 4 were computed twice, once setting the PPP in these cases to 0 and once setting it to missing. The exclusion of those simulations had no effect on the interpretation of the results, so the tables present results based on all simulations.

Table 3
Descriptive Statistics

Method and predictor	*M*	*SD*	Correlations BR	BSP	LR	Bayes	DC
LR[a]							
$R^2_{\text{Y.A}}$	.21	.09	.17	.69	.22	.58	.48
$R^2_{\text{Y.ABC}}$	.30	.13	.12	.63	.30	.67	.56
max $R^2_{\text{Y.A}}$	.29	.12	−.01	.81	.38	.68	.52
Δmax $R^2_{\text{Y.ABC}}$	.43	.18	−.09	.73	.49	.77	.59
BSP	.68	.16	.81				
LR	.79	.20	.00	.21			
Bayes	.69	.15	.78	.91	.23		
DC	.62	.20	.68	.78	.23	.87	
NPP							
BSP	.79	.11	−.80				
LR	.80	.11	−.61	.74			
Bayes	.80	.10	−.81	.91	.71		
DC	.76	.13	−.75	.79	.69	.87	
CF							
BSP	.75	.07	−.37				
LR	.80	.10	−.60	.59			
Bayes	.77	.07	−.37	.80	.56		
DC	.70	.10	−.04	.54	.34	.71	

Note. BR = base rate; BSP = best single predictor (predictor *A* alone); LR = logistic regression; DC = direct computation; PPP = positive predictive power; NPP = negative predictive power; CF = correct fraction.
[a] Correlations for logistic regression are with correct fraction.

for direct computation. In other words, the BSP on average generated diagnostic efficiency statistics almost as good as those based on combinations of three predictors and better than the actual diagnostic efficiency associated with the popular vote-counting heuristic. BSP diagnostic efficiency statistics also tended to correlate well with those statistics resulting from the combination of three predictors.

The correlation matrices provided in the table highlight the important role of BR in diagnostic efficiency. Excluding logistic regression, BR alone accounted for 46%–66% of variability in PPP and NPP. Because these effects are in opposite directions, the finding that their combined effect on the HR was attenuated is not surprising.

Table 4 contains the results of direct comparisons with BSP. The top panel is based on all simulations. For reference purposes, the first two columns of statistics include information about the incremental validity of three predictors when compared with one predictor according to logistic regression. The mean increments in the correlational statistics are restatements of the information in Table 3. Across simulations, none were associated with a decrement in effect size when the set of predictors was increased from one to three, and only .02% remained the same. It is worth noting that these results ignore the potential for shrinkage. Under that condition, linear regression correlational statistics consistently suggest an improvement in fit over the BSP.

The results are very different when diagnostic efficiency is considered. Though logistic regression and the application of Bayes's theorem to the vote-counting heuristic both were associated with a mean increase in all three diagnostic efficiency statistics, the use of three predictors was associated with a decline in diagnostic efficiency in 14%–43% of comparisons with BSP. The results were substantially poorer for the direct computation of diagnostic efficiency. Across the three statistics examined, BSP did at least as well as three predictors in 70% or more of the simulations.

To demonstrate the conclusions drawn earlier about the circumstances under which three predictors would prove better than one, the researcher repeated the analyses using only those simulations in which $p(A-B+C+Y+) \geq p(A+B-C-Y+)$ and $p(A+B-C-Y-) \geq p(A-B+C+Y-)$. The results are given in the lower panel of Table 4. This restriction generally enhanced the increment in fit resulting from use of three predictors. As expected, this enhancement was particularly evident for direct computation. This finding eliminated simulations in which the use of three predictors reduced diagnostic efficiency.

Unfortunately, the joint probabilities one needs to determine whether multiple predictors combined via vote counting will improve over a single predictor are not available to test users. To offer some guidance on circumstances in which the vote-counting heuristic can potentially offer some benefit over the BSP heuristic, the study next addressed the question of whether it is possible to identify circumstances in which additional predictors are likely to increase diagnostic efficiency by using commonly available statistics. For this purpose, it was assumed that the following statistics would be available to a test user or at least estimable: BR, the correlation of each predictor with the criterion, and the correlations between the predictors. Simulations were dichotomized according to whether ΔPPP for direct computation was > 0 versus ≤ 0. The same was done for NPP and HR. Point–biserial correlations were then computed with the statistics assumed to be available, as well as various combinations of those statistics based on similar analyses by Hogarth and Karelaia (2005). The largest point–biserial correlations were associated with the criterion-related validity coefficient for the least valid predictor, r_{YC}, varying between .26 and .34. The best cut scores proved to be .393 for PPP, .40 for NPP, and .41 for HR.

On the basis of these results, it would be reasonable to suggest adding predictors if the validity coefficient for the least valid

Table 4
Improvement Over Best Single Predictor (BSP)

Correlational statistic	LR		ΔPPP			ΔNPP			ΔCF		
	ΔR^2	$\Delta\max R^2$	LR	Bayes	DC	LR	Bayes	DC	LR	Bayes	DC
M	0.09	0.14	0.11	0.01	−0.05	0.01	0.13	−0.03	0.05	0.02	−0.05
SD	0.07	0.10	0.22	0.07	0.12	0.08	0.23	0.08	0.08	0.04	0.09
< 0 (%)	0.00	0.00	30.23	38.54	67.31	43.48	35.73	60.48	13.66	33.86	63.88
$= 0$ (%)	0.02	0.02	12.99	1.62	9.40	11.52	0.53	10.38	33.10	2.99	15.90
> 0 (%)	99.98	99.98	56.78	59.84	23.28	45.00	63.75	29.14	53.24	63.15	20.23
$p(A{-}B{+}C{+}Y{+}) \geq p\,(A{+}B{-}C{-}Y{+})$ and $p\,(A{+}B{-}C{-}Y{-}) \geq p\,(A{-}B{+}C{+}Y{-})$[a]											
M	0.12	0.17	0.12	0.04	0.07	0.03	0.09	0.03	0.06	0.04	0.05
SD	0.08	0.11	0.19	0.06	0.10	0.07	0.18	0.04	0.08	0.03	0.05
< 0 (%)	0.00	0.00	23.33	21.19	0.00	23.01	37.84	0.00	8.08	13.41	0.00
$= 0$ (%)	0.02	0.02	10.34	1.63	37.65	8.45	0.58	37.67	26.94	2.28	37.65
> 0 (%)	99.98	99.98	66.33	77.18	62.35	68.54	61.59	62.33	64.98	84.31	62.35

Note. In each case, results based on three predictors are compared with results from the BSP. LR = logistic regression; PPP = positive predictive power; NPP = negative predictive power, CF = correct fraction; DC = direct computation.
[a]N = 44,123 simulations.

predictor is .40 or higher. This validity coefficient may strike the reader as improbably high for the least valid of three predictors in psychological settings. It should also be noted that the CFs based on this heuristic varied between .70 and .78. In particular, 57% or more of simulations in which additional predictors were useful were misclassified when the cut score of .40 was used. Finally, the test user must consider the costs of collecting additional tests for this minimal payoff. It would seem then that additional predictors are only desirable when they demonstrate relatively high validity and relatively low cost.

Discussion

When reading the literature on applied assessment, one can occasionally find warnings that more information is not necessarily better than less (e.g., Faust, 1989). Even so, when charged with making decisions that have potentially life-altering consequences, psychologists and other users of standardized testing procedures cannot be faulted for associating a greater sense of subjective comfort with larger amounts of information. This association is particularly apt given familiar statistical methods (e.g., linear regression and Bayes's theorem) that reinforce this belief. In fact, linear regression and Bayesian methods in general provide an optimal approach to quantitative prediction under optimal circumstances. What is one to do, though, when conditions are suboptimal, in particular, when information is incomplete about whether each individual is in fact a member of the population used to generate the statistical model? The results of these analyses suggest that the BSP according to zero-order correlation with the criterion can be a better option than is the vote-counting heuristic in cases of three predictors.

An important question to consider is how often in practice the conditions are met under which BSP would trump vote counting. Unfortunately, there is no way to answer this question, because the information about the relative size of certain key joint probabilities is never available in practice. The simulations that served as the basis for the outcomes in Tables 3 and 4 sampled from the array of possible real-world scenarios but were not weighted according to the probability of those scenarios. Furthermore, because they permit the correlations between variables to be as large as .70, the rules used to identify acceptable simulations can be faulted for including too many cases in which the relationships between variables are unusually high. This is particularly true in the case of dichotomous variables, because dichotomization tends to attenuate the size of correlations (MacCallum et al., 2002). The fact that two thirds of the simulations demonstrated a decrement in diagnostic efficiency when vote counting was used instead of BSP does not imply that the same would be true in two thirds of applied testing situations. It should raise serious concerns about the possibility of a decrement in any testing situation, however.

As noted previously, the statistics that are likely to be available to the test user are not particularly helpful for determining whether additional predictors will contribute to diagnostic efficiency. An example of this point is provided in Table 5. Three simulations are presented that are equivalent on BR and the six zero-order correlations. In the first case, the proportion of cases in which A and Y are positive but B and C are negative is larger than is the proportion of cases in which all predictors but A are positive. The result is a decline in all three diagnostic efficiency statistics when B and C are added to A. In the second case, the difference between the first pair of joint probabilities is offset by the difference between the second pair. In this case, PPP is reduced but NPP increases and HR is stable. In the third case, the probability of B, C, and Y being negative is higher than is the probability that only A and Y will be negative; this results in an increase in all three diagnostic efficiency statistics when B and C are considered. The correlations are the product of the 16 joint probabilities and BR; thus, the correlations are equivalent across simulations, because the differences between the joint probabilities listed in the table are offset by differences in other joint probabilities.

This article does not really address utility issues as they apply to practical decision making, but the findings raise serious questions about the "more is better" philosophy, at least in the case of two to three predictors. The analysis offered for the three-predictor case in the Appendix does suggest that as the number of predictors is

Table 5
Sample Simulations

				BSP			DC		
$p(A{+}B{-}C{-}Y{+})$	$p(A{-}\ B{+}C{+}Y{+})$	$p(A{-}B{+}C{+}Y{-})$	$p(A{+}B{-}C{-}Y{-})$	PPP	NPP	CF	PPP	NPP	CF
0.20	0.10	0.00	0.00	0.67	0.75	0.70	0.60	0.60	0.60
0.00	0.15	0.15	0.00	0.75	0.67	0.70	0.64	0.83	0.70
0.00	0.00	0.10	0.20	0.67	0.75	0.70	0.80	0.80	0.80

Note. In each simulation, base rate = .50, r_{YA} = .408, r_{YB} = .314, r_{YC} = .302, r_{AB} = .171, r_{AC} = .123, and r_{BC} = .390. BSP = best single predictor (predictor *A* alone); DC = direct computation; PPP = positive predictive power; NPP = negative predictive power, CF = correct fraction.

expanded further, to four to five predictors per criterion, the probability that a single predictor will prove equal or superior to the vote-counting method declines substantially. However, one must consider the issue of cost when using so many predictors for a single criterion.

It should be noted that heuristic test aggregation in applied settings can take more complicated forms than vote counting. One common alternative modifies the interpretation on the basis of unique characteristics of each predictor. For example, a positive outcome on a valid performance-based measure of thought disorder combined with a negative outcome on a self-report measure of the same construct might be interpreted as evidence of a lower level disorder than full-blown psychosis or of a lack of insight into the oddity of one's thinking. Such an approach could potentially provide more accurate information than could the purely statistical methods discussed here. It also offers some insight into why practitioners often prefer broadband scales that are sensitive to multiple related psychological constructs (Cronbach & Gleser, 1957). In employee development or clinical settings, inconsistencies in outcomes on such measures can be perceived as the starting point for a more fine-grained analysis of the respondent.

Although this approach to aggregating inconsistent test findings can produce intriguing conclusions, it demonstrates a troubling similarity with ad hoc approaches to explaining inconsistent results in significance testing. For example, Schmidt (1996) hypothesized that the use of post hoc explanations based on moderator variables to understand inconsistent outcomes across significance tests, rather than treatment of those inconsistencies as a logical outcome of insufficient power, tends to result in overly complex interpretations of findings. This unnecessary complexity in turn interferes with the accumulation of knowledge in psychology. Similarly, the ad hoc approach that modifies the interpretation of the tests when results seem inconsistent overlooks the possibility that such disparities are due to random variation in indicator outcomes. The result can lead to overly complex and incorrect person descriptions. This analysis is not intended to suggest that the ad hoc approach to integration of inconsistent outcomes is necessarily invalid, just as Schmidt could not be accused of claiming that differences in outcomes across significance tests never occur because of moderators. It does suggest that test users may be insufficiently skeptical about the modified interpretation of tests as a means of explaining inconsistencies in outcomes across multiple measures. This problem is particularly salient when test outcomes are dichotomized, so that small differences in scores can translate into substantial differences in the interpretation of a test.

The findings also raise questions about the appropriate statistical standard for an adequate predictor. The usual standard is a history of significant correlations with the criterion. However, this evidence alone is insufficient to assure that a test enhances prediction. The simulations were limited to cases in which Sens and Spec were each at least .50, because lower values for either would mean that the test could actually result in more errors in one population than could random placement. The truth, however, is that a test with very high Spec but very low Sens can easily produce significant correlations in a sample of reasonable size. Inspection of diagnostic efficiency statistics provides a more reasonable basis for judgments about the use of tests for decision-making purposes. If an estimate of BR is available, direct estimation of PPP and NPP can be particularly useful as a means of avoiding excessive confidence in the implications of a particular test outcome.

The final point to be raised here is the broad applicability of the terms *test* and *predictor,* as used in this article. They are not restricted to formal procedures, such as standardized instruments, but can include interviews, discrete or global clinical impressions, biographical data, and information gathered from significant others. The psychologist who assumes that the issues raised in this article are relevant only to standardized data-gathering procedures is sadly mistaken. Informal procedures demonstrate the same statistical properties as do formal procedures, though there is the added complication that those statistical properties are unknown. For example, when the best psychometric predictor of a construct demonstrates greater criterion-related validity than an interview, if one's goal in gathering data is to make judgments about the test taker and if the outcomes will be combined via vote counting, one must question from a cost–benefit perspective whether there is any practical benefit to interviewing at all. On the other hand, there are often legal and personal expectations about interviewing that might mandate its continued use, even though it may actually reduce accuracy. The combination of predictors as an alternative to the BSP always warrants justification, no matter what the nature of those predictors.

References

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44,* 341–359.

Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine, 16,* 981–991.

Choi, B. C. K. (1997). Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. *Epidemiology, 8,* 80–86.

Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). London: Chapman and Hall.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.
Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.
Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment, 8,* 360–362.
Farrington, D. P., & Loeber, R. (2000). Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health, 10,* 100–122.
Faust, D. (1989). Data integration in legal evaluations: Can clinicians deliver on their premises? *Behavioral Sciences & the Law, 7,* 469–483.
Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics, 28,* 337–374.
Ganellen, R. J. (1996). *Integrating the Rorschach and the MMPI–2 in personality assessment.* Mahwah, NJ: Erlbaum.
Gigerenzer, G., Czerlinski, J., & Martignon, L. (2002). How good are fast and frugal heuristics? In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 559–581). Cambridge, England: Cambridge University Press.
Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.
Hogarth, R. M., & Karelaia, N. (2005). Ignoring information in binary choice with continuous variables: When is less "more"? *Journal of Mathematical Psychology, 49,* 115–124.
Hsu, L. M. (1985). Efficiency of local versus standard MMPI norms: A comment. *Journal of Personality Assessment, 49,* 178–180.
Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15,* 446–455.
Katsikopoulos, K. V., & Martignon, L. (2006). Naïve heuristics for paired comparisons: Some results on their relative accuracy. *Journal of Mathematical Psychology, 50,* 488–494.
MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40.
Martignon, L., & Laskey, K. B. (1999). Bayesian benchmarks for fast and frugal heuristics. In G. Gigerenzer, P. M. Todd, and the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 169–188). Oxford, England: Oxford University Press.
McGrath, R. E. (2001). Toward more clinically relevant assessment research. *Journal of Personality Assessment, 77,* 307–322.
McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d. Psychological Methods, 11,* 386–401.
Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.
Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78,* 691–692.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.
Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23,* 153–158.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.
Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 312–317.
Waller, N. G., Yonce, L. J., Grove, W. M., Faust, D., & Lenzenweger, M. F. (2006). Problem sets and solutions for Bayes's theorem, base rates, and prediction. In N. G. Waller, L. J. Yonce, W. M. Grove, D. Faust, & M. F. Lenzenweger (Eds.), *A Paul Meehl reader: Essays on the practice of scientific psychology* (pp. 237–247). Mahwah, NJ: Erlbaum.
Wiggins, J. S. (1972). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Appendix

Analytic Approach to the Two-Predictor Case

The PPP for the two-predictor case can be restated as follows, assuming that A has been awarded dominance over B:

$$\mathrm{PPP}_2 = \frac{p(A+B+Y+)+p(A+B-Y+)}{p(A+B+Y+)+p(A+B-Y+)+p(A+B+Y-)+p(A+B-Y-)}. \tag{A1}$$

Compare this to the formula for the PPP of A alone when the formula for the PPP of a single predictor (see Table 2) is expanded in consideration of there being a second predictor that is ignored:

$$\mathrm{PPP}_A = \frac{p(A+B+Y+)+p(A+B-Y+)}{p(A+B+Y+)+p(A+B-Y+)+p(A+B+Y-)+p(A+B-Y-)}. \tag{A2}$$

That is, the formulas are exactly the same, and the addition of a second predictor B offers no improvement in the overall PPP. The same relationship holds for NPP_2 versus NPP_A,

$$\mathrm{NPP}_2 = \frac{p(A-B-Y-)+p(A-B+Y-)}{p(A-B-Y-)+p(A-B+Y-)+p(A-B-Y+)+p(A-B+Y+)} \tag{A3}$$

$$\mathrm{NPP}_A = \frac{p(A-B-Y-)+p(A-B+Y-)}{p(A-B-Y-)+p(A-B+Y-)+p(A-B-Y+)+p(A-B+Y+)}, \tag{A4}$$

and for the CF:

$$\mathrm{CF}_2 = p(A+B+Y+)+p(A+B-Y+)+p(A-B+Y-)+p(A-B-Y-) \tag{A5}$$

$$\mathrm{CF}_A = p(A+B+Y+)+p(A+B-Y+)+p(A-B+Y-)+p(A-B-Y-). \tag{A6}$$

Analytic Approach to the Three-Predictor Case

The formula for PPP with three predictors is

$\mathrm{PPP}_3 =$

$$\frac{\begin{array}{l} p(A+B+C+Y+) + p(A+B+C-Y+) \\ \quad + p(A+B-C+Y+) + \underline{p(A-B+C+Y+)} \end{array}}{\left\{\begin{array}{r} p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) \\ + \underline{p(A-B+C+Y+)} + p(A+B+C+Y-) + p(A+B+C-Y-) \\ + p(A+B-C+Y-) + \underline{p(A-B+C+Y-)} \end{array}\right\}} \tag{A7}$$

That is, the numerator represents the probability of at least two predictors being positive and the individual being a member of the targeted population. The denominator represents the probability of at least two predictors being positive.

In contrast, the formula for the PPP of A alone when there are three predictors expands to

$\mathrm{PPP_A} =$

$$\frac{\begin{array}{l} p(A+B+C+Y+) + p(A+B+C-Y+) \\ \quad + p(A+B-C+Y+) + \underline{p(A+B-C-Y+)} \end{array}}{\left\{\begin{array}{r} p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) \\ + \underline{p(A+B-C-Y+)} + p(A+B+C+Y-) + p(A+B+C-Y-) \\ + p(A+B-C+Y-) + \underline{p(A+B-C-Y-)} \end{array}\right\}} \tag{A8}$$

That is, the numerator represents the probability of at least A being positive and the individual being a member of the targeted population, and the denominator represents the probability of at least A being positive.

The two formulas are surprisingly similar. Only the underlined terms differ. Comparison of the formulas suggests the following conclusion: If $p(A+B-C-Y+) > p(A-B+C+Y+)$, or especially if $p(A-B+C+Y-) > p(A+B-C-Y-)$, then $\mathrm{PPP_A} > \mathrm{PPP_3}$. These conditions are particularly likely if A is the best single predictor of population.

Similar comparisons can be offered for NPP and HR, as indicated by the following equations:

$\mathrm{NPP_3} =$

$$\frac{\begin{array}{l} p(A-B-C-Y-) + p(A-B-C+Y-) \\ \quad + p(A-B+C-Y-) + \underline{p(A+B-C-Y-)} \end{array}}{\left\{\begin{array}{r} p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) \\ + \underline{p(A+B-C-Y-)} + p(A-B-C-Y+) + p(A-B-C+Y+) \\ + p(A-B+C-Y+) + \underline{p(A+B-C-Y+)} \end{array}\right\}} \tag{A9}$$

$\mathrm{NPP_A} =$

$$\frac{\begin{array}{l} p(A-B-C-Y-) + p(A-B-C+Y-) \\ \quad + p(A-B+C-Y-) + \underline{p(A-B+C+Y-)} \end{array}}{\left\{\begin{array}{r} p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) \\ + \underline{p(A-B+C+Y-)} + p(A-B-C-Y+) + p(A-B-C+Y+) \\ + p(A-B+C-Y+) + \underline{p(A-B+C+Y+)} \end{array}\right\}} \tag{A10}$$

$\mathrm{CF_3} =$

$$\begin{aligned} & p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) \\ & + \underline{p(A-B+C+Y+)} + p(A-B-C-Y-) + p(A-B-C+Y-) \\ & + p(A-B+C-Y-) + \underline{p(A+B-C-Y-)} \end{aligned} \tag{A11}$$

$$\begin{aligned} \mathrm{CF_A} = {} & p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) \\ & + \underline{p(A+B-C-Y+)} + p(A-B-C-Y-) + p(A-B-C+Y-) \\ & + p(A-B+C-Y-) + \underline{p(A-B+C+Y-)} \end{aligned} \tag{A12}$$

In all three cases, the same sets of joint probabilities determine whether three predictors offer an improvement over one. Specifically, if $p(A+B-C-Y+) > p(A-B+C+Y+)$ and/or $p(A-B+C+Y-) > p(A+B-C-Y-)$, the diagnostic efficiency of the first indicator exceeds that of all three predictors. The only difference across the three statistics is the relative influence of the two comparisons. For PPP, the comparison between $p(A-B+C+Y-)$ and $p(A+B-C-Y-)$ is the more salient to the size of the difference. For NPP, it is the comparison between $p(A+B-C-Y+)$ and $p(A-B+C+Y+)$ that matters most, and the two are equipotent for the CF.

Received June 27, 2007
Revision received June 12, 2008
Accepted June 12, 2008 ■