

Running head: BINARY DECISION-MAKING

Predictor Combination in Binary Decision-Making Situations

Robert E. McGrath

Fairleigh Dickinson University

Author Note

Robert McGrath, School of Psychology, Fairleigh Dickinson University.

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

Address correspondence to Robert McGrath, School of Psychology T-WH1-01, Fairleigh Dickinson University, Teaneck NJ 07666. Email: mcgrath@fdu.edu.

Abstract

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 multiple regression and Bayesian methods can be left with the impression that decisions are consistently improved by combining outcomes across valid predictors. However, neither method accurately reflects the combination process in many applied settings. It was demonstrated that the single best predictor can often perform better than multiple predictors when the predictors are combined using methods common in applied settings, a conclusion consistent with Gigerenzer and Goldstein's (1996) *take the best* approach. Circumstances under which the combination of multiple predictors can improve outcomes are discussed.

Keywords: incremental validity, multiple regression, Bayesian analysis, predictive power, clinical decision-making

Predictor Combination in Binary Decision-Making Situations

Professional psychologists are often faced with the practical task of classifying people, usually into one of two categories. Examples include whether to implement a treatment, whether to hire a person, or whether each of a series of diagnoses applies. The practical need to dichotomize cases is often evident 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 demonstrates the potential for statistical and practical considerations to lead to different conclusions about the most appropriate statistical methods (McGrath, 2001).

It is a common belief among psychologists that the accuracy of predictions is consistently improved by combining multiple valid predictors. The purpose of this article is to review the justification for this belief, and to evaluate the degree to which it is accurate given the procedures commonly used by psychologists in professional settings to combine outcomes across multiple predictors. It will be demonstrated that, depending on circumstances, the psychologist may be better served by considering the single most valid predictor and ignoring additional sources of information.

Statistical Methods for Combining Predictors

Incremental Validity

The justification for combining multiple valid predictors as a means of improving prediction rests on two distinct literatures on statistical methodology. The first, and the one that is probably more familiar to psychologists, has to do with multiple regression and incremental validity (Hunsley & Meyer, 2003; Sechrest, 1963). Incremental validity may be defined as the extent to

which additional indicators enhance prediction over a simpler prediction strategy. Hierarchical regression is the standard statistical method used to evaluate the degree of incremental validity provided by additional predictors, with some form of the correlation coefficient providing the corresponding effect-size index. Though the linear combination of predictors resulting from multiple regression is dimensional, the best estimate can be used to make a dichotomous decision by identifying an optimal cut score. Regression is one example of what can be referred to as a *formal* combination method, where the optimal combination is derived statistically. A number of other formal methods have been identified in the literature, including logistic regression, classification trees, simple summing of predictors (tallying), 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). Since linear regression is the most familiar and thoroughly discussed of the alternatives, at least in the applied psychological literature, subsequent discussion will use it as the exemplar of formal methods.

Table 1 provides four examples of using hierarchical regression analysis in an incremental validity study. Each example provides the incremental validity of predictor *B* over predictor *A*, and the combined incremental validity of predictors *B* and *C*. Subsequent discussion will focus primarily on cases of three predictors but will at times address issues of two predictors.

Based on the six zero-order correlations involving the predictors and criterion, it is possible to compute the multiple correlation for *A* and *B* as combined predictors of criterion *Y*, and for three predictors *A*, *B*, and *C* used simultaneously. Two statistics are often generated using these multiple correlations to indicate the increment in fit afforded by the additional predictors. ΔR^2 , or the semipartial correlation, is the simple difference between the squared multiple correlation for

the full and initial sets of predictors (see Table 2). The semipartial correlation indicates the increment in the proportion of criterion variance predicted by adding the set of predictors S_2 to the set of predictors S_1 . In the first example in Table 1, adding B to A increases the proportion of the variance of the criterion predicted by .016, while adding both B and C to A enhances the proportion predicted by .018.

The second incremental validity index in the table is Cohen's (1988) f^2 , which involves dividing the semipartial correlation (ΔR^2) by the proportion of variance not predicted by the complete set of predictors. Though less commonly reported than ΔR^2 , Cohen offered benchmarks for the interpretation of f^2 . A value of .02 is considered indicative of a small effect, .15 of a medium-sized effect, and .35 of a large effect. On the other hand, the fact that ΔR^2 has an upper bound of 1.0 while f^2 does not can make the former seem more comprehensible.

Both ΔR^2 and f^2 increase as a function of the extent to which S_2 is correlated with Y and S_2 is uncorrelated with S_1 . An important feature of linear regression is that the multiple R 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, and ΔR^2 and f^2 will always be ≥ 0 . This characteristic fosters the belief that when the costs of additional testing are minimal, and when the results of the multiple regression are considered reliable (shrinkage has already been accounted for), it is always desirable to increase the number of valid predictors.

Bayesian Analysis with Multiple Predictors

The second literature has to do with the Bayesian analysis of multiple predictors. Figure 1 illustrates the case when both a predictor and criterion are dichotomous. The criterion variable Y is a dichotomous indicator of whether an individual falls in the targeted ($Y+$) or complement ($Y-$)

population, for example, whether the person meets or does not meet criteria for employment.

This criterion is to be predicted by dichotomized tests $X = A, B,$ 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 target population can be referred to as $p(Y+)$, though the more familiar term base rate (BR) will generally be used instead. The probability of being simultaneously a member of the targeted population and negative on predictor $B,$ which would be a prediction error, will be symbolized $p(B-Y+)$. The conditional probability of a positive outcome on B among members of the target population will be symbolized $p(B+/Y+)$.

Table 2 provides computational formulas for several statistics relevant to the analysis of prediction in 2x2 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 untargeted population, or $p(X-/Y-)$. These statistics reflect the probability of a correct outcome 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, $p(Y+/X+)$, while negative predictive power (NPP) is the corresponding statistic concerning correct outcomes among individuals negative on the predictor, $p(Y-/X-)$. Finally, the overall rate of correct decisions will be referred to as the hit rate (HR), reflecting the probability of a correct prediction.

Sens and Spec have statistical advantages over predictive power statistics. The former are considered characteristics of the respective populations. As a result, they are insensitive to

variations in the BR so long as it is reasonable to assume samples are drawn from the same population. In contrast, PPP and NPP are greatly affected by BR, and therefore should show greater sampling variability. Consider the following restatements of the formulas for PPP and NPP for a given predictor X :

$$(1) \quad PPP_X = \frac{BR * Sens_X}{\{BR * Sens_X\} + \{(1 - BR) * (1 - Spec_X)\}}$$

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

These formulas indicate that, even if Sens and Spec remain constant, PPP will increase and NPP decrease as the BR increases (see Meehl & Rosen, 1955). Even so, predictive power is often of greater interest in applied settings than Sens and Spec because the results are directly relevant to circumstances in which a conclusion must be drawn about the respondent based on test results. There is a direct parallel here with the earlier discussion suggesting that while formal statistical considerations argue against dichotomizing dimensional indicators, practical considerations often mandate it.

Equations 1 and 2 are also interesting because they represent restatements of Bayes' theorem in terms of the symbols introduced here. From a Bayesian perspective, BR can be treated as the prior probability of membership in the targeted population, the probability of membership in the absence of additional information from the indicator. PPP represents the corresponding posterior probability, the probability of membership in the targeted population after finding a positive outcome on a predictor. Similarly, $p(Y^-)$ is the prior probability of membership in the complement population, while NPP is the posterior probability given a negative outcome on X .

One implication of Bayes' theorem is that, if X is a valid predictor of Y , a positive outcome on X will result in a posterior probability (PPP) that is greater than the prior probability (BR). In

other words, a positive outcome should increase one's confidence the individual is a member of the targeted population.¹ A reasonable extrapolation is that the iterative use of multiple predictors should incrementally improve PPP to the extent the respondent produces positive results on each predictor (Waller et al., 2006).

Figure 2 demonstrates this Bayesian method for estimating the benefit of multiple predictors. In this example, $BR = .10$ and $Sens_X = Spec_X = .70$ for all three predictors. The left panel demonstrates the results for case $A+B+C+$, the case where the respondent generated positive outcomes on all three predictors. The right panel provides an example for a case where the results are not completely consistent, $A+B-C+$. Reviewing the left panel, a positive outcome on predictor A suggests the probability of membership in the target population is .21. Using this value as the new prior probability of membership in the population, a positive outcome on B raises that value further to .38. Still a third positive outcome on C raises the posterior probability of membership in the target population to .59. Compared to multiple regression, the Bayesian method is *informal* in that the predictors can be combined on an ad hoc basis.

Finding a positive outcome on all three predictors allows a more confident assertion that the respondent is a member of the target population than does one positive outcome, but it would probably surprise many professional psychologists that there is still a sizeable probability (.41) 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 target 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 before (e.g., Meehl & Rosen, 1955; Wiggins, 1972), but continues to bedevil psychological (and medical) practice.

Limitations of the Multiple Regression and Bayesian Methods

Both the incremental validity literature and Bayes' theorem suggest the advantage of multiple valid predictors over one predictor. It is not surprising that psychologists familiar with one or both of these methods would tend to assume additional predictors always enhance the validity of their predictions. However, both methods demonstrate limitations as a basis for this conclusion. Consider the conditions that must be met before multiple regression can be used as the basis for binary decision-making in real-world cases:

- 1) A sufficiently large sample must be gathered to allow derivation of reliable weights for the desired predictors and a reliable estimate of effect size.
- 2) An optimal cut score must be derived for optimally distinguishing between populations.
- 3) Any changes in the set of predictors used would require deriving a new set of weights and cut score.

The first condition can be circumvented by equally weighting predictor scores after they have been standardized (Dawes & Corrigan, 1974; Hogarth & Karelaia, 2005; Wainer, 1976). Equal weighting, or tallying, can actually produce superior results when compared with multiple regression under circumstances where shrinkage is possible, though the latter will be superior if the results of the multiple regression are reliable.

The second condition cannot be fulfilled without extensive sampling. The third condition is unrealistic in applied settings where the battery of predictors is tailored to the respondent based on variations in the goals of the assessment, time constraints, respondent limitations, or issues of cost. Accordingly, it is often impractical in field settings to base decisions on linear regression. Similar restrictions apply to other formal strategies for combining predictors as well.

More typically, applied practice involves examining the pattern of dichotomous outcomes across predictors. In cases of dimensional predictors, scores are first dichotomized using cut scores that were established for each scale as part of the scale development process.² Psychologists are not alone in this practice: medical professionals often dichotomize outcomes on dimensional indicators such as 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 pattern of results.

This discussion would imply the informal Bayesian method offers a more practical approach to the applied task. The problem is that the Bayesian method almost invariably misestimates the posterior probabilities based on multiple outcomes. In the example described in the left panel of Figure 2, the test user concluded there is a .59 posterior probability of membership in the targeted population if a person produces three positive outcomes. In fact, this estimate is very likely to be wrong. Using an extreme example to demonstrate why this would be the case, suppose predictors *A*, *B*, and *C* all correlate perfectly. If so, then the information about *Y* provided by each predictor is completely redundant, and the posterior probability of membership in the targeted population is still only .21. The iterative Bayesian method represents what Katsikopoulos and Martignon (2006; see also Waller et al., 2006) called a *naïve* heuristic, because it ignores dependencies among the predictors. In real-world circumstances, where predictors tend to correlate strongly but not perfectly, the true posterior probability in the left panel of Figure 2 would fall at an indeterminate value somewhere between .21 and .59 (Cooper, 1990).

The Vote-Counting Method

Since neither multiple regression nor the iterative Bayesian method provided an accurate representation of real-world decision-making processes, a third method was developed that reflected an informal approach to combining outcomes across correlated predictors. This method will be referred to as vote-counting, as it is based on the assumption that the test user's decision will reflect the majority outcome across indicators. This method was first considered in the context of the two-predictor case. This case immediately presents a problem for the vote-counting method. 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 would suggest a bias in favor of the predictor with the higher level of criterion-related validity. This solution in the two-predictor case is discussed for example by Ganellen (1996, pp. 72-73). That is, if $r_{YA} > r_{YB}$, then the decision is positive if both A and B are positive or if A alone is positive, negative if both A and B are negative or A alone is negative.³ While this would seem to be an intuitively reasonable rule, and may well reflect what test users in applied settings do, an analysis of the implications of this strategy for PPP and NPP produces a surprising result. Assume A is the more valid predictor. If so, then the heuristic suggests that if A is positive the decision based on both predictors is always positive, while if A is negative the two-predictor decision is always negative. In other words, the diagnostic efficiency of combining A and B is no different than using A alone. This 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, assuming 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 where

either A is positive but B is negative. The same pattern holds for NPP and HR. The point is made 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 ecologically valid heuristic for the integration of results from two predictors offers no incremental validity over the predictor that is awarded precedence for purposes of tie-breaking. This conclusion holds even if the second predictor demonstrates incremental validity according to hierarchical regression.

Decision-making according to the vote-counting method is more straightforward in the three-predictor case: the decision is positive when at least two out of three predictors are positive and negative when at least two of 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 if they disagree a third is administered as a tie-breaker.

The analytic development of this heuristic is again provided in the Appendix, and the results are equally unexpected based on intuitive grounds. 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 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 combination of A , B , and C by majority vote.

This finding is consistent with Gigerenzer and Goldstein's (1996) *take the best* (TTB) heuristic. TTB bases the decision between two options on the best single predictor. If the best predictor is neutral concerning the two choices, the decision-maker inspects additional predictors in decreasing order of validity until a predictor is reached that suggests a preference for one

choice over the other. Gigerenzer and Goldstein claimed people frequently use TTB to make decisions, especially in circumstances where there is limited time and/or information. They also claimed it is a remarkably effective heuristic. Research on this hypothesis has often found TTB is accurate even though it involves ignoring most of available information. In fact, in circumstances where the results of multiple regression are subject to shrinkage, TTB has often been found superior to multiple regression at identifying the best option (Gigerenzer, Czerlinski, & Martignon, 2002; Hogarth & Karelaia, 2005). There is also evidence to suggest that even when provided with the results of a formal combination strategy, practitioners will simplify the information and use a form of the TTB (Green & Mehr, 1997). The prior discussion suggests TTB will often prove superior to the vote-counting method that is commonly used in applied settings.

Generating Simulations

To explore the alternative methods further, a series of data simulations was created. Each simulation was evaluated using multiple regression, Bayesian analysis for multiple predictors, vote-counting, and the TTB method.

The simulations were created using an algorithm intended to sample from the universe of combinations of dichotomous predictors and criteria likely to be found in applied settings. Each simulation was based on a set of 16 probabilities drawn from two 2x2x2 contingency tables, with one table representing each 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, which can be symbolized $p(A+B+C+/Y+)$. The other seven cells in the table reflected conditional probabilities for the other possible combinations of predictor outcomes given

membership in the target 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 then iteratively set to $p(T) = .02, .10, .30, \text{ and } .50$. BR values $> .50$ were omitted as they would have simply mirrored the results for smaller base rates with PPP and NPP switched. Using the BR and the 16 conditional probabilities 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 $.50 \leq \text{Sens}_x \leq .90$.
3. For each of the three predictors, $.50 \leq \text{Spec}_x \leq .90$.
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 bounds for conditional probability tables. Criteria 2-6 were used to limit the simulations to the types of outcomes likely to occur in applied practice. The last criterion was added to reflect the TTb method as a default, since combining predictors would be unjustified unless doing so proves superior to the best single predictor. This process generated 186,301 unique simulations.

Given the dichotomous nature of Y , these simulations do not reflect the typical context for the use of multiple regression, but the multiple regression analyses were mainly intended to provide a comparison with the more practical methods. The Bayesian estimate of HR (HR_B) was computed using procedures described by Waller et al. (2006). PPP_B was computed by averaging the PPPs for each combination of test outcomes that would lead to a prediction of membership in the target population (at least two of three tests positive), weighted by the probability of that combination occurring. A similar process was used to generate NPP_B . Using equations A7, A9, and A11 provided in the Appendix, PPP_3 , NPP_3 , and HR_3 were computed for the vote-counting method.

Results

Preliminary descriptive statistics may be found in Table 3. For multiple regression, the mean squared correlation for predictor A and for all three predictors combined are provided. The addition of B and C increased the mean proportion of variance predicted from .24 to .33.

The remainder of the table organizes comparisons by diagnostic efficiency statistic. As expected, results derived using the Bayesian method consistently overestimated the benefit resulting from the combination of multiple predictors when compared with the vote-counting method. Even for the Bayesian method, though, improvements over the best single predictor were not substantial, and never exceeded .02 on average. The means for the vote-counting approach were smaller than those for the best single predictor.

The correlation matrices provided in the table highlight the important role of BR in diagnostic efficiency. BR alone accounted for at least 46-66% of variability in PPP and NPP across all three methods examined. Since these effects are in opposite directions, it is not surprising to find their combined effect on the HR was much smaller. Results from the Bayesian

and TTB methods tended to covary very strongly with each other and with results from multiple regression. The correlations between vote-counting and the other methods were smaller, but still consistently large.

Table 4 provides the results of direct comparisons with the best single predictor. If the possibility of sample shrinkage can be ignored, multiple regression consistently offers an improvement in fit over the best predictor. The results are very different when an informal method of combination is used instead. Though on average the Bayesian method suggests an improvement in diagnostic accuracy when three predictors are used instead of one, more than 1/3 of simulations were associated with a decline in accuracy. The results are substantially worse for the vote-counting method. Across the three statistics examined, TTB did at least as well as vote-counting in 70% or more of the simulations.

To demonstrate the conclusions drawn earlier about the circumstances under which three predictors will surpass one, the analyses were repeated using only those simulations where $p(A-B+C+Y+) \geq p(A+B-C-Y+)$ and $p(A+B-C-Y-) \geq p(A-B+C+Y-)$. The results may be found in the lower panel of Table 4. As predicted, these restrictions eliminated simulations where the vote-counting approach suggested three predictors reduced diagnostic efficiency. The improvement was attenuated for the Bayesian method but still evident. However, it is important to keep in mind that statistics for the Bayesian method are inaccurate because of the failure to take dependency between predictors into consideration.

Unfortunately, the joint probabilities needed to determine whether multiple predictors combined via vote-counting will improve over a single predictor are unavailable in the literature. To offer some guidance on circumstances where vote-counting can potentially offer some benefit, the next question addressed was whether it is possible to develop a simple heuristic for

identifying circumstances where additional predictors are likely to increase diagnostic efficiency. 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 vote-counting 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 best single predictor turned out to be the criterion-related validity coefficient for the least valid predictor, r_{YC} . The best cut scores proved to be .393 for PPP, .40 for NPP, and .41 for HR. From these results, it would be reasonable to suggest adding predictors if the validity coefficient for the least valid predictor is .40 or higher. This may strike the reader as an improbably high validity coefficient for the least valid predictor in psychological settings. It is also noteworthy that the hit rates based on this heuristic varied between .70 and .78. In particular, 57% or more of simulations in which additional predictors were useful were misclassified using the cut score of .40. Finally, the test user must consider the costs of collecting additional tests when considering the use of multiple predictors. It would seem then that additional predictors are only desirable when they demonstrate relatively high validity and relatively low cost.

Discussion

To summarize, a variety of approaches to combining outcomes from multiple predictors for purposes of applied prediction were considered. The results substantiate prior discussions in the statistical literature suggesting that multiple regression provides an optimal approach to prediction if one can assume a reliable linear model has been established (e.g., MacCallum et al., 2002). However, what is one to do given the practical obstacles to the implementation of formal

aggregative methods? Test users who are familiar with the standard multiple regression approach to incremental validity, or who understand what has been referred to here as the Bayesian method, cannot be blamed for assuming that additional predictors categorically improve the power of their predictions. The issue ignored by this assumption is that neither method provides an accurate representation of the informal vote-counting method commonly used to combine results across predictors.

The results of these analyses suggest that when a cross-validated formal combination of predictors is not available, the best option is often the best single predictor, the predictor with the highest zero-order correlation with the criterion. When one considers the practical obstacles to the use of formal combinations—the cost of developing formal methods, the impact of test revisions, the individualization of test batteries, the cost of administering multiple tests, and the potential for practitioners to simplify the model if the decision is not computed for them mechanically—the case is strong that the best single predictor is generally the best choice in applied settings.

The results also highlight the importance of distinguishing between evaluations of validity and utility (McGrath, 2001). When relationships between variables are examined for purposes of evaluating construct validity (or theory corroboration, for that matter), the issues at stake are conceptual. They do not point to one set of statistical analyses over another, and so the choice is best made on the basis of formal statistical considerations. When relationships between variables are examined for purposes of evaluating the clinical utility of administering a certain indicator, the issues at stake are pragmatic. The optimal practical strategy need not match the optimal strategy from a formal statistical perspective. Unfortunately, psychologists often ignore this important distinction when deciding how to model decision-making in applied settings.

This article does not offer a full consideration of utility issues as they apply to practical decision-making. To do so would require information about the value of different decisions, information that varies as a function of the decision-making situation (Swets et al., 2000; Wiggins, 1972). Several comments can still be made about the general impact of cost on the issues discussed in this article, though. If informal strategies for predictor aggregation are often inferior to the best single predictor, and it is difficult to predict whether the former will be superior to the latter, one must wonder under what circumstances it is worth administering more than the best single predictor of any criterion. Also, the analysis offered for the three-predictor case in the Appendix suggests that as the number of predictors is expanded further, to 4-5 predictors of a single 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.

A third implication to be drawn from these findings is that test users probably tend to overestimate the increase in subjective probabilities associated with consistent outcomes across predictors (Meehl & Rosen, 1955). This error stems largely from the failure to consider the “tyranny of the base rate” as a factor in determining subjective probability. BR proved to be a dominant force in determining how well an indicator can predict a dichotomous criterion. Even if multiple predictors can enhance the prediction of a criterion, if the BR for the targeted population is low, PPP can still be very far from certainty unless one is willing to use a very large set of independent predictors. Test users might learn a valuable lesson from computing PPP and NPP for themselves when using tests to estimate the probability of membership in some population. Sens and Spec information for a test is often available. The local BR may not be, but even a

reasonable guess at the BR could be used in conjunction with Equations 1 and 2 to provide a rough estimate of PPP and NPP.

It should also be recognized that the informal process of test aggregation in applied settings can take more complicated forms than are modeled here. One common alternative modifies the interpretation based on 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 a lower-level disorder than full-blown psychosis, or as a lack of insight into the oddity of one's thinking. Such an approach could potentially provide more accurate information than the more 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 the outcomes on such measures can be perceived as the starting point for a more fine-grained analysis of the respondent. While this approach to aggregating inconsistent test findings sometimes produces fascinating 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 treating these 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 interpretive approach to integrating inconsistent outcomes is necessarily invalid, just as Schmidt could not be accused of claiming 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 in the context of the dichotomization of test outcomes, where differences of a few points can translate into substantial differences in the interpretation of a test.

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 also include interviews, discrete or global clinical impressions, biographical data, and information gathered from significant others. The psychologist who assumes the issues raised in this article are only relevant to standardized data-gathering procedures is sadly mistaken. Informal procedures demonstrate the same statistical properties as formal procedures, though there is the added complication that those statistical properties are unknown. For example, if the best psychometric predictor of a construct demonstrates greater criterion-related validity than an interview, if the goal of data gathering is to make judgments about the test taker, and if the outcomes will be combined informally, 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 in the absence of any reason to believe it will enhance prediction. The combination of predictors as an alternative to the best single predictor always warrants justification, no matter what the nature of those predictors.

Footnotes

¹Following Meehl and Rosen (1955), expository writing on diagnostic efficiency by psychologists often notes that in cases of extreme base rates, using the test may actually result in a lower hit rate than “betting the base rate,” that is, always predicting the respondent is a member of the modal population (e.g., Hsu, 1985; Waller, Yonce, Grove, Faust, & Lenzenweger, 2006). While technically true, betting the base rate is unacceptable in applied settings for very practical reasons. Consider for example the consequences for a clinical psychologist who refuses to predict anyone is at risk of committing suicide because, given the very low BR for suicide, this practice results in the best HR.

²This practice is in contrast to recommendations 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 impeding the applied use of multiple 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 categorized as 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 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 statistical.

³This heuristic is still technically incomplete, since it ignores the case where $r_{YA} = r_{YB}$. This case is probably rare enough that it deserves being relegated to a footnote, but could still be

addressed by, for example, randomly awarding precedence to *A* or *B*. So long as one predictor is considered dominant, the conclusions drawn in the text remain valid.

References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42, 393-405.
- Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana, IL: 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.
- Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The 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: Cambridge University Press.
- Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650-669.

- Green, L. A., & Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *The Journal of Family Practice*, *45*, 219-226.
- 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.
- McGrath, R. E. (2001). Toward more clinically relevant assessment research. *Journal of Personality Assessment*, *77*, 307-322.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194-216.
- Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, *23*, 153-158.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.
- 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' 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.

Table 1

Examples of Hierarchical Regression Incremental Validity Analyses

r_{YA}	r_{YB}	r_{YC}	r_{AB}	r_{AC}	r_{BC}	$R^2_{Y.AB}$	$R^2_{Y.ABC}$	$\Delta R^2_{YB.A}$	$\Delta R^2_{YBC.A}$	$f^2_{YB.A}$	$f^2_{YBC.A}$
0.114	0.138	0.104	0.116	0.110	0.385	0.029	0.031	0.016	0.018	0.016	0.019
0.314	0.524	0.436	0.099	0.252	0.663	0.344	0.347	0.245	0.248	0.374	0.380
0.280	0.184	0.393	0.361	0.005	0.012	0.086	0.239	0.008	0.161	0.009	0.211
0.503	0.302	0.524	0.616	0.032	0.390	0.253	0.591	0.000	0.338	0.000	0.826

Table 2

Computational Formulas for Incremental Validity and Diagnostic Efficiency Statistics.

Statistic	Probability Represented	Formula
Incremental Validity Statistics		
$\Delta R_{YS_2.S_1}^2$		$R_{Y.S_1S_2}^2 - R_{YS_1}^2$
$f_{YS_2.S_1}^2$		$\frac{R_{Y.S_1S_2}^2 - R_{YS_1}^2}{1 - R_{Y.S_1S_2}^2} = \frac{\Delta R_{YS_2.S_1}^2}{1 - R_{Y.S_1S_2}^2}$
Diagnostic Efficiency Statistics		
Sensitivity (Sens)	$p(X+/Y+)$	$\frac{p(X + Y+)}{p(X + Y+) + p(X - Y+)} = \frac{p(X + Y+)}{p(Y+)}$
Specificity (Spec)	$p(X-/Y-)$	$\frac{p(X - Y-)}{p(X - Y-) + p(X + Y-)} = \frac{p(X - Y-)}{p(Y-)}$
Positive Predictive Power (PPP)	$p(Y+/X+)$	$\frac{p(X + Y+)}{p(X + Y+) + p(X + Y-)} = \frac{p(X + Y+)}{p(X+)}$
Negative Predictive Power (NPP)	$p(Y-/X-)$	$\frac{p(X - Y-)}{p(X - Y-) + p(X - Y+)} = \frac{p(X - Y-)}{p(X-)}$
Hit Rate (HR)	$p(X+Y+ \text{ or } X-Y-)$	$p(X + Y+) + p(X - Y-)$

Table 3

Descriptive Statistics for the Four Methods.

	<i>M</i>	<i>SD</i>	Correlations ^a			
			BR	TTB	Bayes	Vote-Counting
MR						
$R^2_{Y,A}$.24	.11	.07	.81	.67	.54
$R^2_{Y,ABC}$.33	.15	.05	.74	.76	.62
PPP						
TTB	.68	.16	.81			
Bayes	.69	.15	.78	.91		
Vote-Counting	.62	.20	.68	.78	.87	
NPP						
TTB	.79	.11	-.80			
Bayes	.80	.10	-.81	.91		
Vote-Counting	.76	.13	-.75	.79	.87	
HR						
TTB	.75	.07	-.37			
Bayes	.77	.07	-.37	.80		
Vote-Counting	.70	.10	-.04	.54	.71	

^aCorrelations for MR are with HR.

Note. MR = multiple regression; PPP = positive predictive power; NPP = negative predictive power, HR = hit rate; BR = base rate; TTB = take the best (predictor A alone).

Table 4
Improvement over Take the Best.

	MR		ΔPPP		ΔNPP		ΔHR	
	ΔR^2	\bar{r}^2	Bayes	Vote-Counting	Bayes	Vote-Counting	Bayes	Vote-Counting
<i>M</i>	.09	0.19	.01	-.05	.13	-.03	.02	-.05
<i>SD</i>	.07	2.39	.07	.12	.23	.08	.04	.09
< 0 (%)	0.00	0.00	38.54	67.31	35.73	60.48	33.86	63.88
= 0 (%)	0.00	0.00	1.62	9.40	0.53	10.38	2.99	15.90
> 0 (%)	100.00	100.00	59.84	23.28	63.75	29.14	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								
<i>M</i>			.04	.07	.09	.03	.04	.05
<i>SD</i>			.06	.10	.18	.04	.03	.05
< 0 (%)			21.19	0.00	37.84	0.00	13.41	0.00
= 0 (%)			1.63	37.65	0.58	37.67	2.28	37.65
> 0 (%)			77.18	62.35	61.59	62.33	84.31	62.35

^a*N* = 44,123 simulations.

Note. In each case, results based on three predictors are compared with results from the best single predictor. MR = multiple regression; PPP = positive predictive power; NPP = negative predictive power, HR = hit rate.

		Population (Y)		
		Targeted (Y+)	Complement (Y-)	
Test Result (X = A, B, C)	Positive (X+)	$p(X+Y+)$ Sens: $p(X+ Y+)$ PPP: $p(Y+ X+)$	$p(X+Y-)$ 1 - Spec: $p(X+ Y-)$ 1 - PPP: $p(Y- X+)$	$p(X+)$
	Negative (X-)	$p(X-Y+)$ 1 - Sens: $p(X- Y+)$ 1 - NPP: $p(Y+ X-)$	$p(X-Y-)$ Spec: $p(X- Y-)$ NPP: $p(Y- X-)$	$p(X-)$
		$BR: p(Y+)$	$1 - BR: p(Y-)$	$HR: p(X+Y+) + p(X-Y-)$

Figure 1. Symbols used as the basis for possible outcomes of decision analyses. The upper right and lower left cells indicate various ways to present the probability of incorrect decisions, the upper left and lower right cells the probability of correct decisions.

		Y				Y	
		+	-			+	-
A	+	$p(A+Y+) = .07$	$p(A+Y-) = .27$	A	+	$p(A+Y+) = .07$	$p(A+Y-) = .27$
	-	$p(A-Y+) = .03$	$p(A-Y-) = .63$		-	$p(A-Y+) = .03$	$p(A-Y-) = .63$
		$p(Y+) = .10$ $p(Y-) = .90$				$p(Y+) = .10$ $p(Y-) = .90$	
		$p(Y+ A+) = .07/(.07 + .27) = .21$				$p(Y+ A+) = .07/(.07 + .27) = .21$	
		Y				Y	
		+	-			+	-
B	+	$p(B+Y+) = .14$	$p(B+Y-) = .24$	B	+	$p(B+Y+) = .14$	$p(B+Y-) = .24$
	-	$p(B-Y+) = .06$	$p(B-Y-) = .56$		-	$p(B-Y+) = .06$	$p(B-Y-) = .56$
		$p(Y+) = .21$ $p(Y-) = .79$				$p(Y+) = .21$ $p(Y-) = .79$	
		$p(Y+ A+B+) = .14/(.14 + .24) = .38$				$p(Y+ A+B-) = .06/(.06 + .56) = .10$	
		Y				Y	
		+	-			+	-
C	+	$p(C+Y+) = .26$	$p(C+Y-) = .19$	C	+	$p(C+Y+) = .07$	$p(C+Y-) = .27$
	-	$p(C-Y+) = .11$	$p(C-Y-) = .44$		-	$p(C-Y+) = .03$	$p(C-Y-) = .63$
		$p(Y+) = .38$ $p(Y-) = .62$				$p(Y+) = .10$ $p(Y-) = .90$	
		$p(T A+B+) = .38$				$p(T) = .10$	
		$p(Y+ A+B+C+) = .26/(.26 + .19) = .59$				$p(Y+ A+B-C+) = .07/(.07 + .27) = .21$	
		(a)				(b)	

Figure 2. Two examples of the iterative Bayesian approach 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, a positive outcome on B from .21 to .38, and 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 the results for A and B cancel each other out.

Appendix

Analytic Approach to the Two-Predictor Case

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

$$(A1) \quad PPP_2 = \frac{p(A+B+Y+) + p(A+B-Y+)}{p(A+B+Y+) + p(A+B-Y+) + p(A+B+Y-) + p(A+B-Y-)}.$$

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 two predictors:

$$(A2) \quad PPP_A = \frac{p(A+B+Y+) + p(A+B-Y+)}{p(A+B+Y+) + p(A+B-Y+) + p(A+B+Y-) + p(A+B-Y-)}.$$

That is, the formulas turn out to be 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 ,

$$(A3) \quad NPP_2 = \frac{p(A-B-Y-) + p(A-B+Y-)}{p(A-B-Y-) + p(A-B+Y-) + p(A-B-Y+) + p(A-B+Y+)}$$

$$(A4) \quad NPP_A = \frac{p(A-B-Y-) + p(A-B+Y-)}{p(A-B-Y-) + p(A-B+Y-) + p(A-B-Y+) + p(A-B+Y+)},$$

and for the HR:

$$(A5) \quad HR_2 = p(A+B+Y+) + p(A+B-Y+) + p(A-B+Y-) + p(A-B-Y-)$$

$$(A6) \quad HR_A = p(A+B+Y+) + p(A+B-Y+) + p(A-B+Y-) + p(A-B-Y-).$$

Analytic Approach to the Three-Predictor Case

Using the heuristic of based the decision based on a majority of outcomes across predictors, the formula for PPP when three predictors are used becomes:

$$(A7) \quad PPP_3 = \frac{p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) + p(A-B+C+Y+)}{\left\{ \begin{array}{l} p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) + p(A-B+C+Y+) \\ p(A+B+C+Y-) + p(A+B+C-Y-) + p(A+B-C+Y-) + p(A-B+C+Y-) \end{array} \right\}}$$

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

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

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 $PPP_A > PPP_3$. These conditions can be met if A is a better predictor of population than B and C .

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

$$(A9) \quad NPP_3 = \frac{p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) + \underline{p(A+B-C-Y-)}}{\left\{ \begin{array}{l} p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) + \underline{p(A+B-C-Y-)} \\ p(A-B-C-Y+) + p(A-B-C+Y+) + p(A-B+C-Y+) + \underline{p(A+B-C-Y+)} \end{array} \right\}}$$

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

$$(A11) \quad HR_3 = \frac{p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) + \underline{p(A-B+C+Y+)} + p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) + \underline{p(A+B-C-Y-)}}{p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) + \underline{p(A+B-C-Y-)}}$$

$$(A12) \quad HR_A = \frac{p(A+B+C+Y+) + p(A+B+C-Y+) + p(A+B-C+Y+) + \underline{p(A+B-C-Y+)} + p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) + \underline{p(A-B+C+Y-)}}{p(A-B-C-Y-) + p(A-B-C+Y-) + p(A-B+C-Y-) + \underline{p(A-B+C+Y-)}}$$

In all three cases, it is the same sets of joint probabilities that determine whether three predictors offer an improvement over one. Specifically, if $p(A+B-C-Y+) > p(A-B+C+Y+)$ and/or $p(A-B+C+Y-) > p(A+B-C-Y-)$, then 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 most

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 HR.