

## Utility Analysis: What Are the Black Boxes, and Do They Affect Decisions?

*John W. Boudreau, Michael C. Sturman and Timothy A. Judge*  
School of Industrial and Labor Relations, Cornell University

Center for Advanced Human Resource Studies,  
New York State School of Industrial and Labor Relations,  
Cornell University, Ithaca, New York 14853-3901, USA

Suppose we could calculate for a manager the return on investments in human resource management (HRM) programs, such as testing, training, compensation, and performance appraisal. Further, what if such calculations were based on accepted scientific research about the effects of such programs, and could be expressed in terms of financial and economic considerations (such as interest rates and taxes) to make them comparable to the calculations routinely carried out by analysts in such areas as marketing, finance, and manufacturing operations? Suppose further that the vast majority of evidence from studies using such methods had shown that the payoff from HRM programs was extremely high, often exceeding 100% even with very conservative assumptions. For example, one such study (Rich and Boudreau, 1987) calculated the return from using a computer programming ability test (the Programmer Aptitude Test, PAT) instead of traditional interview ratings to select replacements for a group of 201 computer programmers. The analysis suggested that using the PAT would produce an incremental return of US\$2.77 million over an 11-year period, for an additional investment of \$10 per applicant, or about US\$5811. A return-on-investment of over 400 times the initial investment. Moreover, the analysis suggests that there is less than a 2% chance that using the test will lose money, and the chance of seeing a net return of over one million US dollars is greater than 40%.

Such tools not only exist, they have been available to managers for over four decades, and the techniques to implement them have been the subject of intensive discussion and research during the last 15 years. In industrial and organizational psychology, these tools are called 'utility analysis'. Results of utility analysis suggest that in many situations it is possible to express the effects of HRM programs in monetary units and, moreover, the payoff from such programs seems extraordinary even under the most conservative of assumptions. As a result, one might expect to see HRM executives among the most respected within organizations, budgets for HRM programs expanding as organizations realize the value of such programs, and a general recognition of HRM programs as key components of organizational profitability. Yet, those who work within and consult with organizations know that such a renaissance

for HRM programs has not been forthcoming. Organizations persist with activities (such as the unstructured employment interview) that research has shown to produce low returns, and many HRM decisions continue to be made informally, without the use of financial analysis.

Utility analysis has been described as 'a family of theories and measures designed to describe, predict, and/or explain what determines the usefulness or desirability of decision options, and to examine how information affects decisions' (Boudreau, 1991, p. 621). Utility analysis (UA) evolved to provide tools for better describing and communicating the impact of HRM and industrial psychology interventions on organizational goals. Most UA research has been based on an implicit assumption that such communication would be aimed toward managers, who control the resources necessary to implement such interventions. Previous research has noted the need to move beyond developing new measures of utility parameters, and to focus on the role of UA information in managerial decision processes (Boudreau, 1989, 1991). Yet there remains a lack of research exploring these issues. In this chapter, we attempt to explicate some of these underlying assumptions and suggest how future UA research may fruitfully test them. We focus specifically on areas where the practice of human resource management seems to diverge from the implied behaviors of UA theory. These deviations provide potential clues about how to enhance the accuracy and usefulness of UA models, and suggest new directions for future UA research and practice.

The themes developed in this chapter relate to UA applications in training, compensation, performance assessment, and internal staffing. However, for simplicity and exposition, we will use the external selection model as our guiding framework. External selection is the focus of a significant proportion of UA research, so using this model will relate our themes to a large body of literature. Using the external selection model, this review will attempt to 'open up the black boxes' of four fundamental UA processes:

1. The relationship between predictors and criteria, represented by  $r$ , the correlation coefficient.
2. The nature of the criteria, represented by  $SD_y$ .
3. The nature of the selection process, represented by  $Z_x$ .
4. The nature of the implementation process, represented by  $C$ .

Figure 5.1 depicts the distinction between the traditional and proposed frameworks. The traditional framework focuses primarily on the new selection intervention and its ability to predict performance, assumes the predictor scores will be used in a system of top-down hiring (the highest-scoring applicants gets the first job offer, followed by the next-highest scorer, and so on), and focuses on

a one-dimensional measure of performance. The proposed model suggests that new interventions are usually combined with those currently being used, that predictor information is often used in ways that diverge from a top-down selection system, and that the notion of performance must include multiple dimensions or aspects.

### Selection Utility: An Example

An example will help to clarify the meaning and usage of selection utility parameters, and highlight some of the points to be addressed here. The example will illustrate the utility model application of Rich and Boudreau (1987), which incorporated many of the features that have been suggested by researchers, and provides a detailed report of its application. The present example will focus only on those parameters discussed in this paper.

Traditional UA Framework:



Proposed UA Framework:

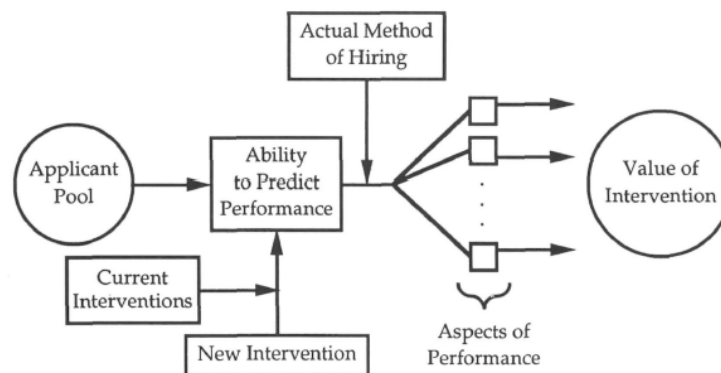


Figure 5.1 Traditional versus Proposed UA Framework

The example of Table 5.1 shows that the value of improved selection is determined by three main factors: (a) the quantity of employees and years affected by improved selection; (b) the value of the change in quality attributed to the program, and (c) the costs of implementing the program. Rich and Boudreau (1987) provide detailed summaries of the derivation of these factors.

Table 5.1 Example Application of Selection Utility Analysis

<i>Quantity of Person-Years Affected</i>	
11-year analysis period	
5-year application of the selection system	
201 existing employees	
LEVG (Projected leverage based on past separation patterns) = 963	
DF <sub>LEV</sub> (Discount factor, discount rate of 15%, 11-year analysis) = 0.519	
<i>Factors Affecting the Value of Quality Change in Job Performance per Person Year</i>	
Selection ratio (the proportion of applicants who are hired) = 0.398	
Z <sub>x</sub> (the average Z-score on the predictor, of those hired) = 0.97	
r <sub>INT</sub> (the expected validity of the interview) = 0.14	
r <sub>PAT</sub> (the expected validity of the Programmer Aptitude Test) = 0.73	
SD <sub>v</sub> (the value of a one-standard-deviation difference in job performance) = US\$15 888	
1-TAX (the proportion of increased profits retained after taxes) = 0.61	
<i>Factors Affecting the Costs of Implementing Selection Systems</i>	
N <sub>a</sub> (Quantity of future applicants tested or interviewed) = 765	
DF <sub>COST</sub> (Discount Factor, 15% discount rate over 5 years) = 0.761	
C <sub>INT</sub> (Cost of the interview, per person) = US\$634	
C <sub>PAT</sub> (Cost of PAT, per person) = US\$644	
1-TAX (the proportion of increased profits retained after taxes) = 0.61	
<i>Utility Calculation</i>	
Utility	= {[LEVG × DF <sub>LEV</sub> × Z <sub>x</sub> × SD <sub>v</sub> × r] - [C <sub>INT</sub> × N <sub>a</sub> × DF <sub>COST</sub> ]} × 1-TAX
US\$ 432 651 <sub>INT</sub>	= {[963 × 0.519 × 0.97 × US\$15 888 × 0.14] - [US\$634 × 765 × 0.761]} × 0.61
US\$3 201 248 <sub>PAT</sub>	= {[963 × 0.519 × 0.97 × US\$15 888 × 0.73] - [US\$644 × 765 × 0.761]} × 0.61
US\$2 768 597 <sub>DIF</sub>	= {[963 × 0.519 × 0.97 × US\$15 888 × 0.59] - [US\$10 × 765 × 0.761]} × 0.61

Note: Calculations based on Rich and Boudreau (1987)

For our purpose, the following points are most important. The quantity of employees and years affected is a function of anticipated hiring and separation patterns, and has a great deal to do with the projected value of improved selection. In this example, these hiring and separation patterns suggest that 963 person-years of performance are affected by selection in this group of 201 programmers. The quality of the performance change is a function of Z<sub>x</sub>, which reflects how choosy the organization can be when hiring. In this case, the organization acquires applicants in the top 39.8% of the distribution, producing an average standardized test score (Z<sub>x</sub>) of 0.97, or almost one standard-deviation higher than if only average applicants were acquired. Quality is also a function of the validity, or strength of the relationship between the predictor (x) and job behaviors, this relationship symbolized by the correlation coefficient, r. In this case, research suggests that the typical unstructured interview has a validity of 0.14, while the PAT has a validity of 0.73. The product of Z<sub>x</sub> and r predicts the average standardized level

of job behaviors we can expect, using the selection device, but this value is not in dollar terms. To translate standard scores into dollars, utility analysis typically adds a factor ( $SD_y$ ) reflecting the dollar value of a one-standard-deviation difference in job behaviors. In this case, a managerial survey was used to estimate the value of  $SD_y$  to be \$15,888 per person, per year (see Rich and Boudreau, 1987 for details of the estimation process). Finally, the cost of improved selection depends on the cost of selecting each individual, and on the number of applicants that must be examined to generate the selection information. In this case, the interview costs US\$633, and the PAT is assumed to add on additional US\$10 to that process (based on previous estimates by other researchers). After accounting for the organization's cost of capital (the discount rate) and tax obligations, Table 5.1 shows that the payoff to both selection devices is substantial, but the PAT produces an extremely high return on investment.

Understandably, such values have aroused great controversy among both researchers and practicing managers (see Boudreau, 1989, 1991 for reviews). There has been a suggestion that such large values must be unrealistic, and heated debate about the proper method of estimating the parameters of the model. Beyond questions of accuracy, however, are larger issues: Is the utility model an appropriate representation of the actual selection issues faced by managers? Is utility information useful in communicating the value of improved selection to managers interested in financial consequences? Why does it appear that utility analysis is used only by a very limited number of practicing managers? While this chapter cannot provide definitive answers to these questions, we propose to highlight some of the issues that bear on these questions, and to suggest directions for future research that should be useful for enhancing our understanding of the value of improved selection, as well as helping practicing managers better employ these tools in their work.

### **Multivariate Validity Over Time: The Black Box of $r$**

The validity or correlation coefficient forms the basis for all UA models. Indeed, frustration with the ability of the validity coefficient to adequately communicate the magnitude of organization consequences from employee selection was what originally led to the search for more complete utility models (Boudreau, 1983, 1991). As the example of Table 5.1 shows, even a predictor with relatively low validity and high cost (such as the interview with a validity of 0.14 and a cost of US\$634 per applicant) can produce handsome payoffs. In the 1950s, early utility-analysis researchers proposed that managers did not seem to understand the value of such predictors, or the value of increasing predictor validity, and speculated that dollar-valued utility calculations might communicate better.

Virtually all UA models have incorporated the validity coefficient, but it usually has been treated as a bivariate correlation between a single predictor and a single criterion (usually performance appraisals, as discussed later). However, evidence suggests that organizations construct their selection processes by using multiple predictors. Virtually every organization uses reference checks, some form of application blanks, and employment interviews (Gatewood and Feild, 1994; Milkovich and Boudreau, 1994), while a smaller proportion of organizations also use ability tests or biographical data. US and Japanese manufacturers are famous for implementing lengthy and detailed selection processes using multiple predictors, especially for team-oriented production processes (e.g. Bowen, Ledford and Nathan, 1991). Thus, when organizations consider implementing new predictors, such as cognitive ability tests or work samples, it seems likely that such methods will be added to the existing group of predictors, rather than being applied separately (Sturman and Judge, 1994). The information of greatest value to decision makers is the incremental validity new selection devices would produce if added to the existing group of predictors. The selection costs associated with the final selection procedure should reflect all of the costs involved in using all procedures, and the validity of the selection procedure should reflect the multivariate combined validity of both the previous and newly-added predictors. The need to consider multiple predictors has been noted before (Jones and Wright, 1992), but has not yet been explicitly incorporated into the UA framework. The resulting utility values under the reasonable assumption of multiple predictor usage may be quite different from those produced under the tenuous assumption that each predictor will be used individually.

For example, using data from Rich and Boudreau (1987), and applying the one-cohort selection utility model of Brodgen, Cronbach, and Gleser (B-C-G), the univariate validity of the interview was 0.14, and the univariate validity of the Programmer Aptitude Test (PAT) was 0.73. The single-device utility of the interview was \$431 744 and the single-device utility of the PAT was \$3 198 258. The implication of such a calculation is that the organization could achieve \$3 198 258 in increased value per cohort by abandoning the interview and using only the PAT. However, organizations seem unlikely to abandon the interview, so the most appropriate comparison is between the combination of the interview and the PAT, versus the interview alone. The formula for the multiple correlation of two predictors (Nunnally, 1978, pp. 176-177), suggest that if the two procedures are combined, the multiple correlation will differ depending on the intercorrelation between the two predictors. If we assume a range of correlations (0.00, 0.25, 0.50, or 0.75) between the PAT and interview, we obtain different multiple correlations for the predictor combination (0.74, 0.73, 0.77 and 0.95, respectively). This result, combined with a cost level that reflects the combined costs of the two predictors, suggests utilities for the predictor

combination of \$3,024,175 , \$2,972,715 , \$3,165,209 , and \$4,011,408 , respectively (Sturman and Judge, 1994). Thus, the incremental utility of combining the two predictors is somewhat lower (6-7%) than the utility of substituting the more-valid predictor for the interview, at moderate intercorrelation levels. Only when the two devices are highly correlated does it make sense to keep both. Using a model that acknowledges predictor intercorrelations provides a more accurate picture of the consequences of maintaining the interview as part of the selection process. Because it is rare for organizations to exclude interviews from their selection processes, such a model also provides a more accurate utility estimate.

Acknowledging the possibility of adding new predictors to old ones requires information about intercorrelations among predictors, which may be absent from validation and validity generalization data. Moreover, in actual selection situations, it appears that decision makers incorporate additional factors into their deliberations, beyond the typical selection dimensions of knowledge, skills, and ability. Individual dispositions and personality have been shown to relate to important employee behaviours (Judge, 1992). Actual selection strategies and decisions may incorporate such dispositions, as is evident from research on the characteristics recruiters look for in job applicants (Bretz, Rynes, and Gerhart, 1993). We need more research on the interrelationships between dispositions and other more traditional predictors, and on the reaction of organizational decision makers to using dispositional information and multiple predictors in their selection decisions. UA models that incorporate such information may be seen as more credible by decision makers familiar with the multiple-predictor reality of employee selection. To date, we have little evidence of how decision makers perceive the additive value of multiple predictors, or whether they can effectively use information about multiple-predictor relationships to make better selection strategy decisions.

Practicing managers might well consider whether they tend to simply add new predictors to old ones. As we have seen, such a practice may waste resources if a higher level of predictability could be attained by simple substitution of a new predictor for an old one.

#### *Temporal validity variations*

Validity may vary over time (Hulin, Henry and Noon, 1990), though evidence is mixed and sparse regarding predictive relationships between criteria and job-performance measures (Barrett, Alexander and Doverspike, 1992). The traditional UA model that uses one validity value for all future periods may misstate the actual value of a predictor whose validity changes. In the worst case, if validity declines over time, then a utility calculation based on constant validity values may significantly overstate the actual value of predictors. We do not know if managers consider this possibility when considering new

selection systems. Managers may have experience with temporal shifts in the relative performance of individuals over time (Hofmann, Jacobs and Gerras, 1992). For example, they may know of cases where the initial differences in performance are significantly reduced as employees gain experience, causing validity to decline as the lower-performing employees improve to the level of the initially-higher performers. Managers may consider such shifts in their decisions. Utility models and values that presume stable relative performance levels within each hired cohort, and thus stable validities, conflict with this experience. Future UA research could profitably explore the degree to which such temporal variations affect utility estimates, and the degree to which they affect decision makers' judgments about the value of selection strategies.

### **Multidimensional Criteria: The Nature of $SD_v$**

Both managers and researchers have focused on  $SD_v$  as a key factor in the credibility and usefulness of utility analysis. As the example of Table 5.1 indicates, the level of  $SD_v$  has profound effects on the dollar values produced, and for almost 15 years there has been a spirited debate about the best way to estimate this parameter, with a variety of alternative measures suggested. Boudreau (1991) reviewed the research investigating alternative methods of estimating the standard deviation of performance in dollars ( $SD_v$ ), and the question of whether it can be estimated accurately. To date, there is little consensus on the proper measurement method, though Boudreau suggested that the minimum value of  $SD_v$  necessary to communicate the value of improved selection is often much less controversial. In Table 5.1, the minimum, value of  $SD_v$  needed to justify using the PAT instead of the interview is US\$1059 per person, per year, which is approximately half of the lowest estimate provided by any manager in the Rich and Boudreau (1987) study. It is not our present purpose to add to the debate about the relative accuracy or appropriateness of various approaches of estimating  $SD_v$ . Rather, we would like to focus attention on a neglected aspect of the issue, the multidimensionality of employee value.

Boudreau (1991) pointed out that UA has its roots in multiattribute utility analysis (MAU), which offers tools and theories for making decisions with multiple and conflicting outcomes (Bazerman, 1990). Undoubtedly, criteria for employee selection are multidimensional. Virtually any introductory HRM textbook recognizes that performance ratings are not the only criterion for employee selection (e.g. Milkovich and Boudreau, 1994). Absenteeism, trainability, flexibility, likelihood of separation, and 'fit' between the individual's characteristics and those needed for long-term organizational success are also commonly valued outcomes of selection (Judge and Ferris, 1992). Yet, virtually all UA applications



presume a univariate criterion,  $y$ , which reflects the dollar value of individual employee performance. Some research has investigated the judgment processes underlying  $SD_y$ . Bobko, Karren and Parkington (1983) noted that sales managers reported using pay as an anchor for their estimates of 'overall worth'. Burke and Frederick (1984) found that supervisors of sales managers reported using five dimensions in judging employee value: (a) management of recruiting, training, and motivating personnel; (b) amount of dollar sales achieved; (c) management of sales coverage; (d) administration of performance appraisal; and (e) forecasting and analyzing sales trends. Burke (1985) found that supervisors of clerical workers followed job evaluation dimensions in their judgments, with salary-related factors most frequently used. Even these investigations focus primarily on job-specific performance behaviors, despite the fact that an individual's value to an organization realistically goes beyond their current performance in a particular job.

#### *Can Multivariate Criteria be represented by Single-Criterion Validity?*

A complete value criterion would include not only job performance, but also the level of withdrawal behaviors, speed with which new tasks are learned and applied, relationships with co-workers, ability to perform in a variety of future work situations, and 'fit' with the cultural and behavioral norms of the organization (Bowen, Ledford and Nathan, 1991; Judge and Ferris, 1992). The perceived value of an individual or group is likely to depend on the individual making the judgment, and these differences in perceived value may differ in ways that are not simply random, as suggested by research which compares the performance ratings of supervisors, co-workers, and subordinates (Harris and Schaubroeck, 1988). Though each may provide valid performance assessments, the results of each assessor may differ from the others. Moreover, meta-analytic studies have suggested that supervisory ratings explain less than 50% of the variance in objective results (Heneman, 1986), so both objective and subjective measures are needed to fully represent performance criteria. Typically,  $SD_y$  is estimated by having experts place a dollar amount on the 'overall value' of employees, which some might argue will automatically incorporate all the relevant performance dimensions. Previous efforts to detect the dimensions used by decision makers in making an overall  $SD_y$  estimate are helpful in testing this argument, but they fail to reflect the fact that different performance dimensions may have different relationships to each other, and may relate very differently to particular predictors.

Even when UA researchers estimate  $SD_y$  with methods that reflect multiple criterion dimensions, they typically measure validity ( $r$ ) as the correlation between a predictor and one criterion (such as performance ratings) to reflect a multidimensional estimate of employee value. Yet

performance ratings are used for many purposes, non-performance factors are often reflected in performance ratings as a way to achieve organizational goals, and managers frequently have little faith that performance ratings truly reflect differences in employee value (Murphy and Cleveland, 1991; Oppler, Campbell, Pulakos and Borman, 1992). There is also evidence that withdrawal behaviors may covary and may form a single common construct (Judge and Hulin, 1994). Maximum and typical performance levels may not be highly correlated (Sackett, Zedeck and Fogli, 1988). It would appear fruitful to consider alternatives to the tradition of using a validity coefficient based on performance ratings to represent the relationship between a selection strategy and the multiple dimensions of employee performance. It would also be fruitful to investigate whether UA estimates are more or less influential depending on how decision makers perceive the dimensionality of employee value. For managers, the question is whether one trusts a measure of  $r$  that reflects only job performance ratings as the criterion, or a measure of  $SD_y$  that may not reflect multiple performance perspectives, multiple performance dimensions, or the value of individuals beyond their current job assignment. Even considering the effects of these omissions, the utility values that have emerged are so large that one could argue they are sufficient to establish the fact that improved selection is worth the investment in almost all cases. Yet, organizations persist in using selection systems that do not reflect research, and utility analysis models remain rare in organizations. To date, we have little evidence on whether decision makers discount the credibility of utility values because they presume that computed values reflect a very limited set of criteria, and that effects of selection on other criteria may offset those included in the calculation.

#### *Using Employee Movement to Represent Multiple Criteria*

The fact that employees are unlikely to spend their entire career with the same organization, let alone in the same job, has been well publicized by the popular business press. UA concepts can reflect the effects of employee recruitment, separations, and movement between jobs (Boudreau, 1991). Thus, employee movement consequences of selection alternatives can be represented by the effects of movements on employee performance in other positions, the quality and quantity of recruits, and/or the quality and quantity of those leaving and staying in the job. For example, the traditional utility model presumes that all employees hired, regardless of their quality, will remain on the job for the same amount of time. Thus, selection mistakes due to lower-validity predictors are assumed to plague the organization for up to ten years in some analyses. A more realistic assumption might be that such

selection mistakes will be weeded out through attrition, thus increasing the value of the workforce even with less-valid selection, and reducing the difference between more and less-valid selection systems.

Much debate focuses on the notion of 'fit' as a long-term measure of the value of an individual to the organization, including multiple criteria of value and encompassing multiple possible future roles (Bowen, Ledford and Nathan, 1991; Judge and Ferris, 1992). It appears that recruiters and interviewers consider the broader concept of 'fit' in their deliberations (Bretz, Rynes and Gerhart, 1993). Other evidence also suggests that applicants choose jobs based on their perceived fit with the organization's values and other attributes (Judge and Bretz, 1992). Thus, utility models that encompass phenomena such as recruitment, separations, and internal movement may better reflect the long-term value of individual fit with the organization. To date, little research has explored the relative accuracy and value of such models, or their attractiveness to decision makers. While such models are undoubtedly more complex than simple single-job utility models (Hunter, Schmidt and Coggin, 1988), no evidence exists to suggest whether such models are viewed as more credible due to their complexity, or less credible due to their inability to incorporate more realistic phenomena.

#### *Multi-Attribute Utility Analysis to Define Criteria*

Extending UA models to reflect movement patterns beyond selection encompasses more of the performance domain, but such utility models still treat performance in any particular work role as a univariate phenomenon. To fully address the issue of multidimensional performance in a given work role requires that the diverse performance elements be combined. One might treat each element as a separate performance measure, calculating utility for each of several performance elements. Certain utility parameters would be the same for all elements, such as the number of applicants, the number of selected employees, the selection ratio, and the cost of the predictor(s). However, for each performance criterion, the validity of the predictor(s) may vary. Similarly, the  $SD_y$  value for each criterion may be different, depending on the amount of discretion afforded by the work role on that dimension, the value of that dimension to the organization, and the variability among applicants and employees in their capability to exhibit that performance dimension. Encompassing multiple performance dimensions in this way requires little modification of existing models. It merely combines several single-criterion models into an overall utility estimate by adding the individual utility estimates for each performance dimension.

However, such an additive approach is likely to produce misleading results depending on the correlations among the criteria. Performance elements may conflict (e.g. effort devoted to individual

selling may detract from helping co-workers and vice versa). If maximizing one performance dimension detracts from the other or, as in the case of typical and maximal performance, if two performance dimensions are uncorrelated, then the sum of the utilities based on them may not accurately reflect the true utility of their combination. What is needed is an approach to estimating  $SD_y$  that allows decision makers to consider multiple performance criteria, yet still obtains an overall utility value reflecting the combination of those criteria. Techniques for estimating overall utility functions from disparate criteria have been available for several decades, generally described as Multi-Attribute Utility (MAU) estimation techniques (Dawes, 1982; Huber, 1980). Such approaches attempt to determine the relative importance and trade-offs between multiple criteria in estimating the overall utility level. We know little about the relative weights decision makers place on different performance dimensions, how different performance dimensions relate to each other, and how these relationships affect the perceived relative value of different employees. Future research might fruitfully examine these questions. The answers will have implications not only for UA research, but also for research on job evaluation and performance assessment.

Moreover, such research is likely to shed light on the ways in which different constituents evaluate employee value. Many organizations are undergoing fundamental changes in the way they organize and assign responsibility for decisions, including flatter structures and decentralized decision power (see Sparrow in this volume, for instance). Such reorganization may mean that very different constituents will make decisions about the use of selection strategies or other human resource programs. The key constituency for UA information in organizations of the future may be teams, task forces, or subordinates, rather than supervisors or top managers. We have little data regarding how work groups define employee value, or whether such definitions differ significantly from those of more traditional supervisors. However, it seems likely, given the necessary interaction inherent in work groups, that interpersonal fit may have even greater value in team-based organizations (Bowen, Ledford and Nathan, 1991). Virtually all existing UA research reflects criteria from the perspective of the supervisor rather than the work group. Utility values calculated from the perspective of the supervisor may not reflect the criteria used by work groups in judging performance (Neuman, 1991). For managers, the key issue may not be whether human resource professionals understand the value of improved selection, but whether teams, task forces, and quality-improvement networks can understand it.

### *Combining Multiple Predictors and Criteria*

If the outcomes of selection strategies are affected by multiple criteria and multiple predictors, UA models should consider how to simultaneously encompass the effects of both. The 'black boxes' of both the correlation coefficient and  $SD_y$  contain complex relationships when multiple criteria and multiple predictors are involved. The utility framework proposed here suggests that actual organizations typically face selection strategy choices in which the consequences depend on multiple predictor-criterion relationships and differentially weighted criterion dimensions. Examining the utility consequences of such situations requires much more extensive information about both the predictor and the criterion space. We need information about predictor interrelationships and their relative ability to predict several different criteria (see Figure 5.1). We also need information on criterion interrelationships and their relative importance in determining overall employee value (Alliger and Janak, 1989). Such information may help explain why predictors that appear to offer high utility within the single-predictor, single-criterion model may be somewhat less attractive when viewed in combination with other predictors and criteria. If the interview is used to assess global traits that may contribute to many criteria, and if new predictors are likely to be combined with the interview in any practical selection system, a multivariate criterion and predictor utility model may offer an explanation for the resilience of the interview compared to such apparently superior predictors as work samples and cognitive ability tests.

### **The Black Box of $Z_x$**

A third utility 'black box' involves the average standardized predictor score of the selected group. As Table 5.1 shows, this parameter reflects the 'choosiness' with which applicants are selected based on their predictor scores. Typically, it is assumed to reflect a single predictor, and selection of applicants from the top down, in order of their predictor scores, until all vacancies are filled. Recognizing the existence of multiple predictors means that this parameter may best be considered a predictor composite, weighted to achieve maximum predictability. Existing utility models do not generally proceed from this perspective, but it is relatively simple to consider the variable  $x$  as a weighted composite of predictors with each individual's  $x$ -score representing the value of that composite. It seems plausible that managers conceive of the selection process in this way, with several predictors interacting to form a composite from which selection decisions are made.

A more significant limitation of existing utility models concerns the classic assumption that applicants are scored on the predictor (or composite), are ranked from the top down on those scores, job offers are made in order of predictor scores, and applicants accept all job offers. Such assumptions are convenient because they make the calculation of  $Z_x$  a simple transformation of the normal distribution, for which there are available tables of values (Naylor and Shine, 1965). Research has addressed the possibility that applicants reject job offers which, depending on the relationship between applicant qualifications and rejection patterns, reduces the effective value of  $Z_x$  (e.g. Murphy, 1986; Rich and Boudreau, 1987). Others have examined the utility effects of modifying a top-down selection strategy to enhance the racial and/or gender composition of the accepted group (Schmidt, Mack and Hunter, 1984).

Each of these modifications provides insights into the degree to which deviations from top-down hiring affect the traditional selection utility model. While some organizations undoubtedly do rank applicants and offer jobs according to those rankings, we believe that deviations from the top-down strategy are likely to be the rule, rather than the exception, and can occur for many reasons other than organizational attempts to reach diversity goals or to offset the effects of candidates rejecting offers. Decision makers seldom will be content simply to make selection decisions based on the top-down scoring of a particular selection device, especially a cognitive ability test. Selection strategies have many goals beyond obtaining high-quality employees. Interviews, for example, may be intended to communicate organizational goals, to present a certain image to applicants, or to involve certain organizational constituents in the selection process. Evidence from employee recruiting suggests that recruiter behaviors and characteristics are interpreted by applicants as indicators of organizational characteristics (Rynes, Bretz and Gerhart, 1991), so organizations may adopt recruiting, screening, and selection techniques based on these characteristics rather than on the ability to select the best candidates.

Even when decision makers attempt to use selection devices purely to identify the best quality candidates, there are likely to be deviations from the typical top-down strategy. Evidence from the decision-making literature suggests that people tend to ignore base rates, overweigh vivid information, and overestimate their ability to make accurate probabilistic judgments (Bazerman, 1990). These tendencies may make decision makers reluctant to adopt a simple top-down hiring strategy, even when such a strategy is statistically likely to yield the highest quality pool of new hires. For example, if a cognitive ability test is added to a selection system that previously involved only interviews, the highest utility strategy may be to ignore the interviews and select using a top-down ranking based on test

scores. Yet organizations persist in relying heavily on unstructured interviews, and we can assume that the results of interviews are used in combination with test scores. We have little evidence regarding how actual selection decision makers combine different types of selection information to make decisions. If the top-down approach is not adopted, perhaps decision makers mistakenly place too much weight on vivid interview information or fail to realize the fallibility of interview impressions.

Previous research has examined whether applicant characteristics influence selector willingness to offer jobs, but this research generally has focused on personal characteristics such as race, or objective qualifications such as academic credentials (e.g. Lin, Dobbins and Farh, 1992; Macan and Dipboye, 1990; Olian, Schwab and Haberfeld, 1988). Future research might profitably examine how decision makers combine information from different selection methods. When decision makers participate in one or more selection activities (such as interviews), their behaviours may not conform to the classic utility model assumption of top-down selection. If decision makers 'believe in' the validity of the interview and/or find the idea of 'testing' dehumanizing or otherwise inappropriate, they may choose to ignore test scores in favor of interview impressions. The classic utility model portrays the potential utility of tests when used rigorously, but it does not reflect the actual utility of tests interjected into situations where decision makers are unprepared to use them. The 'marginalist' approach (Jones and Wright, 1992), through which decision makers apply techniques with the lowest marginal cost first and progress to higher-cost techniques only as needed, may not describe actual decision behavior. Thus, the true effect of selection strategies is achieved only after the selection information is filtered through the cognitive behavior of decision makers. Even the most valid and least costly strategy may have low utility if its results are given too little weight when actual selection decisions are made.

Future utility research could fruitfully examine how decision makers react to the availability of different kinds of selection information. It would be interesting to know whether involving decision makers in the selection process increases their willingness to use selection information, or whether compelling evidence (such as utility analyses showing that top-down selection based on test scores is lucrative) can convince decision makers to alter their traditional practices. As noted earlier, much research on decision making has documented the biases exhibited when making decisions and using information. There is ample evidence that linear models outperform human decision makers across a variety of situations (Dawes, 1982), so we would expect utility models to be capable of outperforming human decision makers in selection (when accuracy in predicting job performance is the criterion). Yet there is also evidence that decision makers fail to adopt potentially lucrative human resource strategies

due to various cognitive limitations (e.g. Florin-Thuma and Boudreau, 1987). To date, there has been little integration between this research and the use of selection information.

### *Is Maximizing Organizational Performance the True Objective?*

One additional reason why decision maker behaviours may deviate from utility-analysis prescriptions is the possibility that utility analysis describes a model in which one strives to increase organizational performance. An alternative possibility is that organizational decisions are directed at very different objectives, such as individual political advancement, or that they occur through a random process called a 'garbage-can' model (Huber, 1982). For example, decision makers may persist in using unstructured interviews precisely because such techniques have low validity and thus maintain ambiguity in the selection process. Such ambiguity may provide a power base for those professionals who are recognized as successful in achieving good results in such a 'mysterious' process. If such results could be achieved simply by using test scores, the powerful selection 'experts' might well lose their organizational power. While admittedly a somewhat cynical view, this perspective seems plausible, at least in some situations. Thus, future research might examine the possibility that utility analysis is resisted not because the managers are irrational or do not understand the models, but because the managers' criteria are very different from those suggested by utility analysis.

### **Actual Barriers to Adopting New Interventions, the Full Costs, C**

The classic utility model reflects the incremental costs of new selection devices in the parameter C. These costs can be expressed in terms of the number of applicants, the number of hires, or the total program. The cost term in utility models virtually always reflects only the costs of actually administering and scoring selection procedures. This often ignores some very real costs which are likely to affect the value of improved selection strategies. Some organizations will be required to increase the pay and/or benefits of workers hired under a new selection strategy. Economic theory predicts that unless an organization adjusts wages to reflect increased qualifications, it will be difficult to attract the most qualified workers (Cascio and Morris, 1990; Jones and Wright, 1992). Becker (1989) has discussed several scenarios through which the apparent benefits of selection may be reduced due to increased remuneration obligations or because of decrements in quality caused by a failure to increase remuneration commensurate with increased qualifications. We have little research to suggest how applicant qualifications interact with organizational rewards, but it seems plausible that utility models which ignore the need to adjust rewards for qualifications may overstate utility levels. On the other



hand, substantial evidence suggests that variability in rewards for a given job is not large enough to reflect all variability in employee value (Bishop, 1987). Thus, the cost adjustment necessary to attract better-qualified employees may be less than the increase in value they afford the organization. Still, wages and benefits are not the only factors affecting organizational attractiveness. Other organizational attributes may require adjustment to attract desired applicants to the organization. The decision to select more rigorously may entail changes in the organizational features most likely to be valued by the more qualified applicants (Wanous, Poland, Premack and Davis, 1992; Williams and Dreher, 1992), and these changes likely entail costs. Furthermore, recent research suggests that use of some selection instruments (e.g. drug tests or stress tests) may be more costly to organizations because they are perceived as unfair by applicants, decreasing job seekers' willingness to accept a job offer from organizations using these instruments (Judge, Cable, Blancerio and Johnson, 1994; Smither, Reilly, Millsap, Pearlman and Stoffey, 1993).

The costs of implementing new selection strategies may be much higher than previously suggested because previous calculations did not include the resources necessary to convince decision makers to use selection information most efficiently. Organizations develop routines and customs, and adopting new HR programs may frequently challenge these routines and customs. A hidden cost may well be the effort and resources necessary to 'unfreeze' the organization, so that it is prepared even to consider new approaches. This sort of organizational inertia may be costly to oppose (Boudreau, 1984; Jones and Wright, 1992). It may be difficult to attain high potential utility because key decision makers will simply not use new methods. Those who do implement new HR programs are likely to encounter costs of convincing, training, and communicating with key constituents in order to gain acceptance. Typical costs might include training managers or employee teams on the benefits of new selection processes, employee communications explaining the need for new processes and their divergence from processes used before, and support systems (such as computer systems and data bases) to ease implementing and using new procedures. There also may be required changes in recruiter behaviours to attract high-quality applicants. To date, we have little information about the actual and perceptual barriers to implementing new selection systems and the costs of covering them. Both the subjective and objective costs of surmounting resistance to change are likely to affect actual utility applications, though they remain absent from most UA discussions.

## Making Progress through Practitioner- Researcher Collaboration

UA research provides models that identify some of the variables that affect selection utility, but the focus on model development and parameter measurement has left large gaps in our understanding of whether important assumptions are met, and whether information from such models is likely to be used by decision makers. If UA models are to achieve widespread use, future research must shed light on these issues to guide the further development of new UA models, and to provide assistance in presenting UA information to key stakeholders. We believe that UA models can improve communication and understanding of the value of selection, and can lead to better decisions about investments in improved selection, as well as other HR programs. However, the apparent lack of widespread adoption of UA, and the potentially damaging tendency for managers to fail to adopt valuable HR interventions, suggests a need for a more collaborative approach to future understanding. Both researchers and practitioners must play a role, if future UA model developments are to illuminate the questions posed here.

First, HR professionals should calculate UA values, even using simple traditional models. While published reports suggest that such values are high, we do not yet know if this is generally true. Numerous UA studies during the last 15 years, and the example of Table 5.1, suggest an extraordinary potential for lucrative returns to HR investments. However, we have little evidence that selectors are aware of such possibilities. If they begin to implement even simple UA models, such calculations will increase their understanding of the models. This will undoubtedly lead to questions, which will provide the kernels of future research. In this chapter, we have suggested that selection practitioners may wish to focus on the following questions:

1. Does the actual selection system use multiple predictors, instead of only single predictors? Do recruiters combine information from different predictors? If so, collaborative research to determine the relationships among Utility Analysis: 'Black Boxes' 93 predictors, and their relative ability to predict different criteria, may be useful.
2. Do recruiters consider only a single performance criterion, or does an employee's value to an organization depend on multiple factors, assessed from multiple perspectives? If HR practitioners believe that UA models are limited by a single-criterion focus, then collaborative research can help to identify the multiple criteria, multiple performance assessors, as well as the effects on UA values of considering

multiple predictor-criterion relationships. Also relevant here is the question of whether recruiters require that UA models acknowledge the movement of employees between positions. If so, researchers and practitioners could fruitfully collaborate to identify what movements are most critical to determining employee value, and how to incorporate those movements into the model in a way that is credible and useful to managers.

3. How do recruiters actually use selection information? Is a top-down hiring approach realistic, or do the practical demands of selection require that other approaches be used? Researchers and practitioners could work together to identify and describe how selection information is actually used, and then to determine if moving toward a top-down strategy is valuable. Such research could also provide selectors with useful information on how the information from multiple selection devices is combined and used by decision makers. Theories can suggest the sort of biases that are most likely to occur, and collaborative research could examine their occurrence and effects. If information is actually used in a non-optimal way, opportunities for training may be uncovered. There may also be potential here for 'expert systems' that can incorporate the technology of UA within automated systems that free the recruiter from the burdens of computation and technical understanding. Some UA research has used computer technology in this way (e.g. Rich and Boudreau, 1987), but to date we have little information about how such systems might work with practicing recruiters. Also relevant here is the question of whether the 'rational' approach of UA models is even realistic. Research is needed to clarify the true purposes served by selection systems, and the degree to which they confer personal, expert or political power on certain groups.
4. What are the true costs and resources necessary to change selection systems, and are they accurately reflected in the UA calculations? The example in Table 5.1 reflects the suggestion by many psychologists that improved testing will require only minimal investments (such as \$10 per test). Are such suggestions realistic? If not, what are the key implementation costs, and how can they better be reflected in future utility models?

In summary, we continue to believe that UA research has the potential to help recruiters make better HR decisions, and to communicate the impact of those decisions more effectively. However, to

achieve that potential requires research that focuses less on improvements to the traditional utility model, and more on examining whether that model fits the reality of HR decisions in actual organizations. Such research requires that HR managers and researchers work together. HR managers must strive to understand the UA framework, to help direct research that will make it as useful as possible. HR researchers must see practitioners not merely as subjects providing estimates of UA parameters, but as the recipient of UA information, and as the source of information to help improve UA models. Such collaborations have great potential to realize the potential for UA models in the future.

## References

- Alliger, G. M. and Janak, E. A. (1989). Kirpatrick's level of training criteria: Thirty years later. *Personnel Psychology*, 42, 331-342.
- Barrett, G. V., Alexander, R. A. and Doverspike, D. (1992). The implications for personnel selection of apparent declines in predictive validities over time: A critique of Hulin, Henry, and Noon. *Personnel Psychology*, 45, 601-617.
- Bazerman, M. (1990). *Judgment in Managerial Decision Making*, 2nd edn. New York: John Wiley.
- Becker, B. E. (1989). The influence of labor markets on human resources utility estimates. *Personnel Psychology*, 42, 531-546.
- Bishop, J. H. (1987). The recognition and reward of employee performance. *Journal of Labor Economics*, 5, S36-S56.
- Bobko, P., Karren, R. and Parkington, J. J. (1983). Estimation of standard deviations in utility analysis: An empirical test. *Journal of Applied Psychology*, 68, 170-176.
- Boudreau, J. W. (1983). Effects of employee flows on utility analysis of human resource productivity improvement programs. *Journal of Applied Psychology*, 68, 396-407.
- Boudreau, J. W. (1984). Decision theory contributions to HRM research and practice. *Industrial Relations*, 23, 198-217.
- Boudreau, J. W. (1989). Selection utility analysis: An agenda for future research. In M. Smith and I. Robertson (eds), *Advances in Personnel Selection and Assessment* (pp. 227-258). London: John Wiley.
- Boudreau, J. W. (1991). Utility analysis for decisions in human resource management. In M. D. Dunnette and L. M. Hough (eds), *Handbook of Industrial and Organizational Psychology*, 2nd edn (vol. 2, pp. 621-745). Palo Alto: Consulting Psychologists Press.

- Bowen, D. E., Ledford, G. E., Jr and Nathan, B. R. (1991). Hiring for the organization, not the job. *Academy of Management Executive*, 5, 35-51.
- Bretz, R. D., Rynes, S. and Gerhart, B. (1993). Recruiter perceptions of applicant fit: Implications for individual career planning and job search behavior. *Journal of Vocational Behavior*, 43, 310-327.
- Burke, M. J. (1985). An investigation of dimensions employed and percentile ordering effects in estimating performance standard deviations in dollars for clerical occupations. Unpublished manuscript.
- Burke, M. J. and Frederick, J. T. (1984). Two modified procedures for estimating standard deviations in utility analyses. *Journal of Applied Psychology*, 71, 334-339.
- Cascio, W. F. and Morris, J. R. (1990). A critical reanalysis of Hunter, Schmidt, and Coggin's (1988) 'problems and pitfalls in using capital budgeting and financial accounting techniques in assessing the utility of personnel programs'. *Journal of Applied Psychology*, 75, 410-417.
- Dawes, R. M. (1982). The robust beauty of improper linear models in decision making. In D. Kahneman, P. Slovic and A. Tversky (eds), *Judgment under Uncertainty: Heuristics and Biases* (pp. 391-407). Cambridge, UK: Cambridge University Press.
- Florin-Thuma, B. C. and Boudreau, J. W. (1987). Performance feedback utility in a small organization: Effects on organizational outcomes and managerial decision processes. *Personnel Psychology*, 40, 693-713.
- Gatewood, R. D. and Feild, H. S. (1994). *Human Resource Selection*, 3rd edn. Hinsdale, IL: Dryden Press.
- Harris, M. M. and Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43-62.
- Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology*, 39, 811-826.
- Hofmann, D. A., Jacobs, R. and Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology*, 77, 185-195.
- Huber, G. P. (1980). *Managerial Decision Making*. Glenview, IL: Scott, Foresman & Company.
- Huber, G. P. (1982). 'Decision support systems: Their present nature and future applications. In G. R. Ungson and D. N. Braunstein (eds), *Decision Making-an Interdisciplinary Inquiry*. Belmont, CA: Wadsworth Publishing Company.
- Hulin, C. L., Henry, R. A. and Noon, S. L. (1990). Adding a dimension: Time as a factor in predictive relationships. *Psychological Bulletin*, 107, 328-340.

- Hunter, J. E., Schmidt, F. L. and Coggin, T. D. (1988). Problems and pitfalls in using capital budgeting and financial accounting techniques in assessing the utility of personnel programs. *Journal of Applied Psychology*, 73, 522-528.
- Jones, G. R. and Wright, P.M. (1992). An economic approach to conceptualizing the utility of human resource management practices. In G. R. Ferris and K. M. Rowland (eds), *Research in Personnel and Human Resources Management* (vol. 10, pp. 271-279). Greenwich, CT: JAI Press.
- Judge, T. A. (1992). The dispositional perspective in human resources research. In G. R. Ferris and K. M. Rowland (eds), *Research in Personnel and Human Resources Management* (vol. 10, pp. 31-72). Greenwich, CT: JAI Press.
- Judge, T. A., Cable, D. M., Blancerio, D. and Johnson, D. E. (1994). Effect of selection systems on job choice decisions. Working paper, Cornell University.
- Judge, T. A. and Bretz, R. D., Jr. (1992). Effects of work values on job choice decisions. *Journal of Applied Psychology*, 77, 261-171.
- Judge, T. A. and Ferris, G. R. (1992). The elusive criterion of fit in human staffing decisions. *Human Resource Planning*, 15, 47-68.
- Judge, T. A. and Hulin, C. L. (1994). Job satisfaction and subjective well-being as influences on employee withdrawal. Working paper, *Center for Advanced Human Resource Studies*, Cornell University.
- Lin, T., Dobbins, G. H. and Farh, J. (1992). A field study of race and age similarity effects on interview ratings in conventional and situational interviews. *Journal of Applied Psychology*, 77, 363-371.
- Macan, T. H. and Dipboye, R. L. (1990). The relationship of interviewers' preinterview impressions to selection and recruitment outcomes. *Personnel Psychology*, 43, 745-768.
- Milkovich, G. T. and Boudreau, J. W. (1994). *Human Resource Management*, 7th edn. Homewood, IL: Richard D. Irwin.
- Murphy, K. M. (1986). When your top choice turns you down: The effect of rejected offers on the utility of selection tests. *Psychological Bulletin*, 99, 133-138.
- Murphy, K. R. and Cleveland, J. N. (1991). *Performance Appraisal: An Organizational Perspective*. Boston: Allyn & Bacon.
- Naylor, J. C. and Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology*, 3, 33-42.
- Neuman, G. A. (1991). Autonomous work group selection. *Journal of Business and Psychology*, 6, 283-291.
- Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

- Olian, J. D., Schwab, D. P. and Haberfield, Y. (1988). The impact of applicant gender compared to qualifications on hiring recommendations: A meta-analysis of experimental studies. *Organizational Behavior and Human Decision Processes*, 41, 180-195.
- Oppler, S. H., Campbell, J.P., Pulakos, E. D. and Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: review, results and conclusions. *Journal of Applied Psychology*, 77, 201-217.
- Rich, J. R. and Boudreau, J. W. (1987). Effects of variability and risk in utility analysis: An empirical test and simulation. *Personnel Psychology*, 40, 55-84.
- Rynes, S. L., Bretz, R. D. and Gerhart, B. (1991). The importance of recruitment in job choice: A different way of looking. *Personnel Psychology*, 44, 13-22.
- Sackett, P. R., Zedeck, S. and Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, 78, 482-486.
- Schmidt, F. L., Mack, M. J. and Hunter, J. E. (1984). Selection utility in the occupation of US park ranger for three modes of test use. *Journal of Applied Psychology*, 69, 490-497.
- Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K. and Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, 46, 49-76.
- Sturman, M. C. and Judge, T. J. (1994). Utility analysis for multiple selection devices and multiple outcomes. Working paper, Cornell University.
- Wanous, J. P., Poland, T. D., Premack, S. L. and Davis, K. S. (1992). The effects of met expectations on newcomer attitudes and behaviors: A review and meta-analysis. *Journal of Applied Psychology*, 77, 288-297.
- Williams, M. L. and Dreher, G. F. (1992). Compensation system attributes and applicant pool characteristics. *Academy of Management Journal*, 35, 571-595.