EDUCATIONAL RESEARCH, METHODOLOGY, AND MEASUREMENT: AN INTERNATIONAL HANDBOOK
SECOND EDITION

Edited by
JOHN P. KEEVES
Flinders University of South Australia, Adelaide, Australia

PERGAMON
Decision Theory in Educational Testing

W. J. van der Linden

In most educational situations, tests are used for decision-making rather than measurement purposes. The ultimate purpose in these cases is to use test scores not as quantitative ability estimates but merely as data on which qualitative decisions can be based. Examples of such decisions are admissions to training programs, pass/fail decisions, certification, treatment assignment in individualized instructional systems, and the identification of optimal vocational alternatives in guidance situations. In all of these examples, decisions are ordinarily based on cutting scores carefully selected in order to optimize the actions to be taken.

In spite of the fact that tests are used mostly for decision-making, much psychometric research has been aimed at improving the use of educational and psychological tests as means for estimating ability scores from test performances. The first to recognize this paradox were Cronbach and Gleser (1965) in their classical monograph Psychological Tests and Personnel Decisions. At first their plea for a more decision-theoretical approach to testing had more impact on (personnel) psychologists than on educators. During
the 1970s, however, the situation changed dramatically, and in the mid 1990s the use of decision theory in educational testing is one of the main research topics. The major impetus for this has come from the introduction of novel testing procedures in individualized instructional systems, and from such politically controversial issues as culture-fair selection for schools.

This entry describes the various decision problems that can be met in the practice of educational testing and shows how (Bayesian) decision theory can be applied to optimize the use of tests for decision-making.

1. A Typology of Test-based Decision-making

Test-based decisions can be classified in many ways. A simple typology is the following, which is based on the use of flowcharts to define different types of decision-making. In each decision problem, three common elements can be identified: (a) the test that provides the information the decision is based on, (b) the treatment with respect to which the decision is made, and (c) the criterion by which the success of the treatment is measured. "Treatment" is a generic term here, standing for any manipulation aimed at improving the condition of individuals. Examples include training programs, the use of special instructional materials, therapeutic measures, and the like. The criterion may be any type of success measure, but is often a test itself. With the aid of these elements, four basic flowcharts can be formed, each defining a different type of decision-making.

1.1 Selection Decisions

In selection problems, the decision is whether or not to accept individuals for a treatment. The test is administered before the treatment takes place and only individuals promising satisfactory results on the criterion are accepted. Selection decisions may imply that individuals who are rejected are not admitted to the institute providing the treatment, or have to leave the institute if they were already in it. Figure 1 shows the flowchart of a selection problem. Examples of selection decisions are admission examinations to schools, hiring of personnel in industry, or the intake of students for a remedial program.

The selection problem is the oldest decision problem in the history of educational testing. Traditionally, the problem has been approached as a prediction problem in which regression lines or expectancy tables are employed to predict whether the criterion scores of individuals exceed a certain value. Only individuals with criterion scores above the threshold are accepted. Selection decisions with quota restrictions (see below) have long been evaluated with the aid of Taylor-Russell (1939) tables, which give success ratios for a number of parameters characterizing the selection situation (see Regression of Quantified Data).

1.2 Mastery Decisions

Mastery decisions are made for individuals who have already undergone a treatment. Unlike selection decisions, the question is not whether individuals are qualified for admittance to a treatment, but whether they have profited enough from a treatment to be dismissed. Figure 2 shows the flowchart of a mastery decision. For this type of decision problem the test and the criterion coincide. More particularly, the test is an unreliable representation of the criterion, or, equivalently, the criterion can be considered the true score underlying the test. As the test is unreliable, erroneous decisions are possible and a mastery decision problem exists.

Mastery decisions usually imply that individuals may leave the institute providing the treatment, or proceed with another treatment. Examples of mastery decisions are pass-fail decisions, certification decisions, and decisions with respect to therapeutic success.

Interest in mastery decisions has grown because of the introduction of such modern instructional systems as individualized instruction, mastery learning, and computer-aided instruction. In the past the main concern has been for issues related to standard setting procedures; that is, to procedures for selecting threshold values on the (true-score) criterion separating "masters" from "nonmasters." The influence of measurement error on decision-making was simply ignored. That this may lead to serious decision errors was clearly demonstrated by Hambleton and Novick (1973).

1.3 Placement Decisions

This type of decision problem differs from the preceding two in that alternative treatments are available.
decision theory in educational testing

The success of each treatment is measured by the same criterion. All individuals are administered the same test, and the task is to assign them to the most promising treatment on the basis of their test scores. Unlike the selection problem, each individual is assigned to a treatment. The case of placement decisions with two treatments is represented in Fig. 3. Examples of placement decisions can be found in individualized instruction, where students typically are assigned to different routes through an instructional unit or are offered alternative instructional materials.

Interest in placement decisions has emanated from aptitude-treatment interaction (ATI) research, which was motivated by the finding that individuals may react differentially to treatments and that above average treatments may be worse in individual cases. The placement decision problem has mostly been approached as a prediction problem to be tackled using linear-regression techniques. For each treatment there is a regression line of the criterion on the test score, and individuals are assigned to the treatment with the largest predicted criterion score. The methodology needed for detecting ATIs is reviewed in Cronbach and Snow (1977).

### 1.4 Classification Decisions

In classification decisions, the problem also consists of a choice among a number of different treatments. As opposed to placement decisions, however, each treatment has a different criterion. The situation is as shown in Fig. 4. In order to be able to compare criterion performances across different treatments, it is necessary to transform each criterion to a common utility scale. Examples of classification decisions can be found in vocational guidance situations in which most promising schools or careers must be identified or in testing for military service.

The most popular approach to classification decisions has been the use of linear-regression techniques again. For each treatment, the regression line of its utility (which equals transformed criterion) scores on the test scores is estimated and individuals are assigned to the treatment with the largest predicted utility score. Usually, as more than one criterion is present, there is not a single test, but a battery of tests covering the various aspects of all of the criteria. If so, the use of multiple- rather than bivariate-regression techniques has been the traditional choice.

### 1.5 Combinations of Elementary Decisions

In practice, the decisions in the above typology are mostly met in combination. A simple example is a decision problem in which more than one treatment is available, but not all individuals are accepted for a treatment. These features create a combination of a selection and a placement problem. Another example is a selection decision, where after the treatment a mastery test is administered to assess its success. However, all such combinations of decisions can be mapped on flowcharts built up of Figs. 1–4 as elements.

A new development is to simultaneously optimize decision rules for systems of combinations of decisions rather than to optimize the individual decisions one at a time (Vos 1990, 1991, Vos and van der Linden 1987). The advantage of a simultaneous approach is that more realistic utility structures can be adopted; for example, structures in which the utilities involved the separate decisions are modeled as a function of an individual’s position on the last criterion in the system. Also, test scores on one test automatically provide collateral information for decisions based on scores on other tests. A practical application of simultaneous optimization of decision rules is individualized instruction, for example, as implemented in CAl systems. In such systems, at various points of time small sets of items are used to make a selection, mastery, or placement decision.

### 1.6 Possible Constraints and Extensions

For each type of decision-making, one or more of the following constraints or extensions may be in force:

(a) Quota restrictions. In some cases, the number of vacancies per treatment is constrained by a quota. Such quotas usually simplify the derivation of optimal rules for selection decisions (accept individuals with the highest test scores until the quota is filled), but complicate placement or classification decisions.

---

**Figure 3**
Flowchart of a placement decision (case of two treatments)

**Figure 4**
Flowchart of a classification decision (case of two treatments)
Decision Theory in Educational Testing

(b) Multiple criteria. More than one criterion can be necessary to measure success of treatments. In decision theory such cases are known as multiple-objective or multiattribute decision-making. Methodology to analyze cases with multiple criteria is given in Louviere (1988).

(c) Multiple tests. The presence of multiple criteria may also be a reason to replace a single test by a full battery of tests. As a consequence, the decision must be made on the basis of multivariate information, which complicates the decision rule.

(d) Multiple populations. In some applications, populations varying on a socially relevant attribute can be distinguished. If the test is biased against one of the populations, the problem of fair decision-making arises (Gross & Su 1975; Novick and Petersen 1976).

(e) Adaptive testing. Tests may be administered adaptively, selecting consecutive items to match the current estimate of the individual’s ability. In this case, decision rules have to be modified accordingly; for example, using the framework of sequential Bayesian decision-making (Lindgren 1976).

2. Optimizing Decisions Based on Tests

Decision theory is a branch of statistics addressing the use of data as an aid in decision-making. More specifically, it is concerned with how random data on “true” or future states of nature and utilities associated with the possible outcomes of the decision process can be combined to design optimal decision rules. In the above decision problems, the data are provided by a test. Since a test may not produce reliable data, it has to be considered a random indicator of an individual’s performance. For a population of individuals, test and criterion scores relate to each other in a way that is fully specified by their joint probability distribution. In some decision problems, a psychometric model is needed to specify this distribution.

Several approaches to optimizing decisions can be taken, one of which is Bayesian decision theory. It is indicated below how (empirical) Bayesian theory tackles the four decision problems and provides optimal decision rules (called Bayes rules). First, the Bayesian solution to a classification problem is described. This problem, in a formal sense, is the most complicated decision problem. Then, it will be indicated how solutions to the other decision problems can be obtained by imposing certain restrictions and modifications on the classification model. To enhance understanding, however, in treating these topics some mathematical precision will be sacrificed (see Bayesian Statistics in the Analysis of Data).

The classification problem is formalized as follows. Suppose a series of individuals, who can be considered as being randomly drawn from a population, must be classified into $t$ treatments indexed by $j = 0, \ldots, t$. The observed test scores are denoted by a random variable $X$ with discrete values $x = 0, \ldots, n$. Each treatment leads to a certain performance on its corresponding criterion, which is denoted by a continuous random variable $Y_j$ with range $R_j$. It is assumed that the joint distributions of $X$ and $Y_j$ are given by probability functions $p(x, y)$.

$$\lambda_j(x) = \lambda(x)$$

for all values of $j$.

Generally, a decision rule is a function that indicates for each possible observation which of the possible actions is to be taken. In the present problem the observations take the form $X = x$, and the possible actions are the assignments to one of the treatments $0, \ldots, t$. It is assumed that the optimal rule takes a monotone form; that is, it can be defined using a series of cutting scores $0 = c_0 \leq c_1 \leq \ldots \leq c_t = n (t \leq n)$, where treatment $j$ is assigned to individuals whose scores satisfy $c_{j-1} < X \leq c_j$. For an optimal rule to be monotone, some conditions must be met (Ferguson 1967), which are not unrealistic for the present problems.

Suppose that for individuals assigned to treatment $j$, the decision-maker is able to express his or her preferences for the outcomes $Y_j = y_j$ on a numerical scale. Technically, such an evaluation is known as a utility function. If utility functions can be established for all treatment–criterion combinations, all possible outcomes of the decision have been made comparable on a common scale. To express its dependency on both the criterion and the chosen treatment, utility functions will be denoted as $u = u(y_j)$. Figure 5 shows some examples of utility functions that have received some interest in the literature. The threshold function represents the case where a critical value on the criterion discriminates between successful and unsuccessful performances. The other two functions increase more gradually with the criterion performance. The choice of a utility function may be facilitated by varying its form and studying the robustness of the optimal decision rule to these variations (e.g., Wijm and Molenar 1981).

For each possible series of cutting scores $(c_1, \ldots, c_t)$ the expected utility of the decision procedure can be calculated as:

$$B(c_1, \ldots, c_t) = \sum_{j=0}^{t} \sum_{i=1}^{n} \eta_j(y) \eta_j(x, y)dy$$

The set of optimal cutting scores in the Bayesian sense is the choice of values for $(c_1, \ldots, c_t)$ maximizing the expected utility. A simple procedure to find these values is as follows. Using Eqn. (1) the expected utility can be written as:
the relation of test scores to criterion scores. Hence, the subscript $j$ in $\omega_j(y \mid x)$ can be dropped. Usually two true states are defined—the mastery ($\tau \geq \tau_c$) and the nonmastery state ($\tau < \tau_c$), where $\tau_c$ is a standard chosen for instructional reasons. Different utility functions for the mastery and the nonmastery decision are distinguished. Generally, the former increases in the criterion $\tau$, because the utility of a mastery decision tends to be larger for students with higher true scores. Analogously, the utility of a nonmastery decision decreases in the true score. The mastery decision model follows if, in addition to Eqsns. (4–6), the following restrictions are imposed:

$$y_1 = \tau_1 = \text{EX}_i$$  

$$\omega_j(y \mid x) = \omega(y \mid x)$$  

with $u_0(y)$ decreasing in $y$, $u_1(y)$ increasing in $y$, and where $j = 0, 1$ denotes the nonmastery and mastery decision respectively. Cutting scores maximizing the expected utility under these restrictions have been examined for threshold (Huynh 1976), linear (van der Linden and Mellenbergh 1977), and normal ogive utility functions (Novick and Lindley 1978, van der Linden 1980).

As in mastery decisions, selection decisions involve only one treatment for which individuals are either accepted or rejected. When an individual is rejected, he or she is of no value to the institute making the selection decisions. This is formalized by putting the utility for the rejection decision equal to zero. Further, the criterion is always an external variable and not a true score underlying the test. In summary, optimal selection decisions follow from maximizing the expected utility in Eqn. (3), if in addition to the restrictions in Eqsns. (4–6) the following restrictions are imposed:

$$\omega_j(y \mid x) = \omega(y \mid x)$$  

$$u_0(y) = 0$$  

where $j = 0, 1$ denotes the rejection and acceptance decision, respectively.

3. Miscellaneous Results

As indicated above, in practice the four types of decisions discussed in this entry are not always met in their pure form. Further constraints and extensions may apply. Optimal rules for selection and placement decisions under quota restrictions are given in Chuang et al. (1981) and van der Linden (1990) respectively. If multiple populations must be distinguished, as is the case, for example, in culture-fair selection, each population may entail its own utility and probability functions, and cutting scores must be selected for each population separately. Optimal rules for culture-
fair selection are given in Gross and Su (1975), Mellenbergh and van der Linden (1981), and Peterson (1976). Reckase (1983) discusses the use of Wald’s probability ratio test with mastery decisions based on adaptive testing procedures. If the decision rule has to be based on multiple tests, the posterior distributions in Eqn. (3) are conditional on multivariate data. The derivation of optimal rules is still possible, but numerically more involved than in the case of univariate data.

References

Gross A L, Su W H 1977 Defining a “fair” or “unbiased” selection model: A question of utilities. J. Appl. Psychol. 60(3): 345–51

van der Linden W J 1981 Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores. Psychometrika. 46(3): 257–74