Personnel selection constitutes a well-known example of the application of measurement theory and methods to practical problems. For a long time, "applied measurement" was synonymous with personnel selection. Even today there is a profitable cross-fertilization between psychometrics and personnel selection. In the past, personnel selection has benefited several times — and with encouraging results — from various psychometric concepts and procedures, such as the development of paper-and-pencil tests and everything associated, reliability and validity concepts, application of decision theory, and operations research. Another recent example is validity generalization (see chapter 6 in this book). It can be noted, too, that several subjects have drawn the attention of psychometricians after they were discovered in the application field of personnel selection. Examples such as restriction of range, stability of regression parameters, utility analysis, and the definition of criteria are just a few of these subjects. The relationship between personnel selection and measurement theory has been fruitful and will continue to be so, as we will hope to show in this chapter.

Some recent developments in personnel selection methodology will be considered here. This discussion will not focus on measurement devices
as such, but rather on their significance for the design of selection procedures. This reflects our opinion that important developments in the field of personnel selection during the last decade do not lie at the level of test construction, measurement models, or validity estimation. It seems to us that here the greatest profit has already been obtained, and that the outcomes, shortcomings included, have stabilized. Schmitt and associates (1984) reported mean validity coefficients of general mental ability and special aptitudes of (approximately) .25. For personality tests, a mean coefficient of .15 was reported. Work samples and assessment centers can be considered as two of the better predictive instruments with validities of .38 and .41, respectively. For other nonpsychological predictors, analogous results are in order. The selection interview has a mean validity of .15; for reference checks .17 has been found; for academic achievement .17, and for expert recommendation .21. (These latter values are corrected for several artifacts such as differences in predictor and criterion reliability; see Hunter & Hunter, 1984, p. 83.) Most of the aforementioned tests and other predictors can be qualified as more or less well-developed instruments. Reliability and validity cannot be much improved through modifications. Concepts and methods that relate to the integration of various elements into overall selection procedures seem to be more promising. They have to do with the following issues:

1. What predictors can be used for specific questions?
2. Which types of predictors (tests, work samples, references, diplomas, interviews, etc.) are suitable for which type of question?
3. How can several instruments be combined in order to maximize the utility-cost ratio?
4. How does one arrive at a decision? What utility functions can be applied? Which decision principle, cutoffs?
5. How can test results be communicated to executives and managers?
6. How can acceptability of predictors be taken into account?

While concentrating on selection procedures rather than predictive tests, special attention will be given to the notion of "design." This is because we feel that systematic design of selection procedures can lead to considerable improvement over the prevailing approach in which selection procedures are little more than a clinical scenario for administering a battery of tests to a group of applicants who have applied for some type of job.

Recruitment and selection are important activities that can have far-reaching consequences for the selecting organization, the applicant, and
society. From the viewpoint of the organization, these activities offer the possibility to attract and maintain an adequate work force. The presence of the proper personnel not only offers the possibility to work effectively and efficiently but also supplies the necessary creativity and energy for survival. For the individual, employment chances are at stake: selection decisions affect career perspectives, social status, the benefits of work (material as well as immaterial), and so on. At the societal level, recruitment and selection are the mechanisms by which labor is distributed. They determine largely who is going to be affected by discrimination and unemployment, and in what way. There is an important economic function of recruitment and selection as well. Selection according to appropriate procedures may lead to considerable productivity increases. The overall benefit of good selection practice may range in the order of billions of dollars each year (e.g., Schmidt et al., 1979). From these three perspectives, different requirements are being put forward for selection procedures. Employers require a quick fulfillment of vacancies by well-qualified employees at minimal costs. Applicants demand adequate procedures and fair decisions. In fact, their demands have become much stronger in recent years. A primary reason for this lies in the risk of discrimination that is implied in selection, especially from interviews and tests (Schmitt & Noe, 1986). In the United States, this has led to legislation and additional rulings (e.g., the Uniform Guidelines on Employee Selection Procedures, 1978) for nondiscriminatory selection. The discrimination issue has become more urgent because of the long-lasting condition of labor surplus. Such a condition, which implies small selection ratios, makes the opportunity for fair selection less favorable. A qualitative implication is that minority and majority groups, as well as educational groups, are competing with each other for available slots.

Other factors that influence demands regarding personnel selection result from changes in work and working conditions, caused by technological innovations and organizational restructuring. In addition, new strategies of personnel management lead to great changes in labor conditions, such as reduction of working hours, part-time work or remuneration, and so on. Finally, there is increasing pressure toward higher efficiency, effectiveness, quality, and flexibility, not only in general but also with regard to personnel management. Methods are needed to decrease labor costs, and personnel selection seems to offer attractive possibilities because it can help to create quantitative and qualitative work force changes within a relatively short period of time. The trend toward more efficiency can be found within the domain of selection itself, in the increasing orientation toward cost-benefit aspects.
It will be clear that, against this background (new technologies, changing working conditions, discrimination, displacement effects on the labor market, optimization of human resources), effective selection is not an easy objective. In the following sections, some available methods of attaining this objective will be considered.

A Technological View of Personnel Selection

Looking back over the 70-year history of personnel selection, it is striking that, while a great number of tests and other predictive instruments have been invented and many procedures for psychometric evaluation have been developed, little attention has been devoted to the actual use of tests under practical conditions, and little developmental effort has been directed at the construction of integral selection procedures that are suited for such conditions.

In fact, there seems to be only one basic recipe for developing procedures, sometimes referred to as the "classical model" (see, e.g., Cascio, 1987). This method includes such steps as: job analysis, choice of criteria, choice of predictors, validation of predictors, and revision or establishment of the predictor set on the basis of a validity study. Revisions of this model have added steps concerning the combination of predictor scores (e.g., multiple regression, multiple cutoff, multiple hurdle) and the evaluation of predictor sets in terms other than validity (decision-making accuracy and utility). All of these latter variants share the idea that a selection procedure is some optimally weighted battery of tests (or other instruments) that serves for predicting future performance. They imply that there is essentially one best solution to the selection problem, basically determined by psychometric criteria.

The approach fails to consider a number of important issues:

1. As the focus lies on prediction, issues concerning the decision-making process, the composition of the procedure, and the communication between employer, candidate, and selection consultant are given insufficient attention.
2. With regard to prediction, few options tend to be considered. The main choice is usually considered to be between a single criterion-classical regression system and a multiple criterion-clinical prediction system, while other possibilities are ignored.
3. Problems concerning the actual use of selection procedures under
practical conditions, e.g., time pressures, specific information needs of employers, withdrawals of candidates, and so on, tend to be neglected. Such aspects play a limited role in the evaluation and improvement of procedures.

We feel that personnel selection should be placed in a different perspective, one that helps to get a better view of relevant issues, facilitates the construction of procedures, and enables critical evaluation and further development. This perspective can be characterized as “technological” (Roe, 1987, 1989). It considers theoretical know-how on personnel selection to be a technology that can be drawn from when trying to find solutions for practical problems of personnel procurement within particular companies. The task for the selection specialist is to find a solution for these problems by developing suitable procedures. Thus, the central concept is “selection procedure,” i.e., a series of steps to obtain relevant information from employers and applicants, and to transform this information into a valuable employment decision or advice, which has been made up in such a way that the needs of the company, the client, law, and so on, can be satisfied while all kinds of practical conditions are taken into account.

This “technological approach” defines the framework for the discussion of developments in selection testing in this chapter. The main question to be addressed is: what do we know about the development of selection procedures and what has research added to our knowledge in recent years? We will also discuss some issues for which further developmental work is needed.

Although there definitely is a tradition of designing procedures within the field of personnel selection, there are only a limited number of concepts and methods, and much of the know-how has remained implicit. We think that applied psychology and personnel selection in particular can find a proper methodology in the so-called “design methodology” from the engineering sciences. This design methodology is basically a set of rules for defining, making, and maintaining a technical product, in this case a personnel selection procedure. An important methodological concept is the “design cycle” (Eekels, 1983), which specifies a number of steps that have to be taken in the course of an iterative process. These steps are:

1. **Definition** of the functions of the selection procedure within a given context: basically, these functions are to collect relevant information, make a prediction of performance, make a decision, and report it to the employer and candidate.
2. **Analysis** of the functions in terms of specific requirements and constraints regarding the input data, the decision, and the transformation...
process, and so on; here the employer’s criteria, ethical standards, costs, etc., are to be specified.

3. Synthesis, production of a preliminary selection procedure, making use of knowledge about people and their behaviors, specific tools and procedural options, and knowledge of designing itself (choosing or constructing elements, assembling parts, etc.).

4. Simulation, testing the operational, predictive, and economical properties of the selection procedure; this means: establishing validity, effectiveness, utility, as well as duration, capacity, costs, and so on, either empirically or by the use of models.

5. Evaluation, assessing the value of the selection procedure against the requirements and constraints found in step 2; ascertaining whether the procedure as a whole is satisfactory.

6. Decision making, either accepting the selection procedure for operational use, or rejecting it, followed by a return to step 2 or 3.

The principle of iteration is typical for the design process: the proper solution is only found after a number of efforts. This follows from the nature of designing as a reductive rather than a deductive process. Logically speaking, several solutions are feasible; it is impossible to determine the best solution in a single round. Of course, the “design cycle” only represents the basic structure of the design process (Eekels, 1983). In real practice this process is far more complex, as there are several consecutive processes starting with a global design and proceeding in a step-by-step manner toward a completely detailed design. During specific phases of the whole process, a number of parallel design processes may take place, directed at the creation of separate components. It will be evident that the techniques and methods of personnel selection can be fitted within this framework. Most of them relate to components made in steps 3 through 5 of this cycle.

The main functions that selection procedures should fulfill can be described as follows (Roe, 1989):

1. Information gathering: obtaining information about job openings, job content, job requirements, and on physical, behavioral, and biographical characteristics of applicants;

2. Prediction: transforming information on (past or present) applicant characteristics into predictions about their future behavior, and the resulting contributions to organizational goals;

3. Decision: transforming predictive information on applicants into a preferred action (eventually also execution of this action);

4. Information characteristics, predicting and communicating it to the employee.

These four functions below, there are sectors of the options of sample-based prediction functions, that is of modeling. It will be clear these options largely. Many selection pi prediction and/or d variables and relations prediction and decision modeling, is a comp by applying the de following steps (Roe, 1989):

1. Definition of conditions of use;
2. Identification on these goals and design criteria;
3. Choice of the elements;
4. Choice of the between elements;
5. Choice of the variables types and relational;
6. Evaluation of the formation of performance and decision on the data of the development.

What has been presented here as well: models of iterations is necessary. In the following I discuss development procedures. The inf
4. Information supply: producing information on applicant characteristics, predicted behaviors, plans for actions (decisions), and communicating it to managers.

These four functions can be realized in several ways. As we will discuss below, there are some basic design options for each of them. For example, one of the options regarding the prediction function is that of sign versus sample-based prediction. Another option, which applies to each of the functions, is that of informal (clinical) versus formal information processing. It will be clear that the choices made by the designer with regard to these options largely determine what form the procedure will finally take.

Many selection procedures, especially the more advanced ones, contain prediction and/or decision models. Such models, which specify relevant variables and relationships between them, serve to operationalize the prediction and decisions functions. Developing such models, also called modeling, is a complex activity, which can be approached systematically by applying the design methodology. Such an approach implies the following steps (Roe, 1984):

1. Definition of the problem, resulting in specification of goals and conditions of use;
2. Identification of requirements that the model should satisfy, based on these goals and conditions; also specification of these requirements into design criteria;
3. Choice of the model’s contents, i.e., the constituent elements;
4. Choice of the model’s structure, i.e., the set of relationships between elements;
5. Choice of the model’s format, i.e., type of representation of elements and relationships;
6. Choice of the model’s parameters, i.e., the values that specify variables types and degrees of relationships;
7. Evaluation of the model against the design criteria;
8. Decision on the acceptability of the model or a need for revision.

What has been put forward for the design process in general pertains here as well: models can usually not be found at once. Instead, a number of iterations is necessary in order to arrive at an acceptable result.

In the following sections, we will discuss methodological developments regarding the prediction and the decision function. Consequently, we will discuss developments with regard to the overall composition of selection procedures. The information-gathering and supply functions will be left
out of this review. For these latter subjects, the interested reader is referred to Roe (1989).

**Prediction**

**Major Design Considerations**

Personnel selection implies making decisions with regard to people’s future behavior. Therefore, some type of prediction has to take place.

We have described the prediction function of the selection procedure as transforming information on (past or present) applicant characteristics into predictions about their future behavior, and the resulting contributions to organizational goals. With respect to this function, there seem to be two fundamental design questions:

1. Whether to predict on the basis of the deductive-nomological (or sign) approach, or the domain-sampling (or sample) approach; and
2. Whether to use formal or informal (clinical) methods for arriving at a prognostic statement on performance.

In combination, there are four basic forms of prediction, from which one or more may be chosen:

1. Prediction on the basis of a nomological model. The model contains a formalized specification of the hypothetical relationship(s) between one or more predictor variables and one or more criterion measures.
2. Prediction on the basis of work samples. “Content-oriented devices” are used to measure past or present performance; scores are generalized in a formal way (i.e., statistically) to future performance estimates, e.g., using confidence intervals.
3. Clinical prediction based on predictor comparison. The scores of applicants on predictor variables are compared in order to find the one with the best overall profile; it is assumed that this person’s performance on the job will be best.
4. Clinical prediction based on criterion analogies. The work performance of applications in similar situations is analyzed in order to draw analogies; thus an idea of future performance is derived from past performance.

In this chapter, we will concentrate on prediction with the help of prediction models, as they have received most attention and are at the center of methodological dev...

Designing Prediction: a special case of perfor mance variables - exogenous variables on these criteria or t...
methodological developments. No further consideration will be given to clinical methods. Apart from the reconciliatory reviews of the clinical-statistical controversy by Sarbin (1986), Mehl (1986), and Holt (1986), and Einhorn's (1986) proposition that clinicians and statisticians differ mainly in their conception of prediction error, there is little new to report.

**Designing Prediction Models.** Prediction models can be considered as a special case of performance models. In a performance model, the performance variables (criteria) under study are explicitly related to a set of exogenous variables. The latter variables either serve to predict outcomes on these criteria or to facilitate the understanding of performance behaviors.

Performance modeling has received considerable attention in recent years (e.g., Campbell, 1983; Naylor, 1983; Vance et al., 1989). It should be noted that several of the performance models are of a generic type; they list a large number of variables that influence performance. As an example of such a general model, the reader is referred to the well-known management performance model of Campbell and colleagues (1970). In this model, the performance of managers is depicted as a function of various individual characteristics (intelligence, aptitudes, knowledge, temperament, preferences, expectations) and environmental variables (i.e., climate and cultural conditions). Although such models are informative in the sense that the available literature and research are nicely summarized, they cannot be considered as working models for guiding interventions. In personnel selection, the model should only contain specific variables that (1) are (presumably) relevant for the problem at hand, (2) can be assessed at the moment of application, and (3) are stable enough to allow predictions over a longer time period. In addition, the number of variables should be minimized in order to get an acceptable utility-cost ratio.

In the context of personnel selection, a prediction model can be defined as a model that transforms information on applicant characteristics or behaviors into a diagnostic statement about future behavior (or behavioral outcomes). Technically speaking, the model is said to transform information on predictors into information on criteria. Applying the aforementioned design methodology to prediction models leads to the following description of the design process.

1. Defining the problem and specifying requirements: in what way is the information on applicant characteristics transformed into predictions of future work performance?
2. Defining the model's content: what are criteria and predictors?
3. Choosing the model's structure: how are predictor variables inter-
correlated? And how do criterion variables interrelate? How are both sets related to each other?

4. Choosing the model’s format: normally, an algebraic function of the type \( Y = f(X) \) is chosen.

5. Estimating model parameters: of special importance are instrumental parameters like test type and length, number of raters, etc., on the one hand, and relational parameters like regression weights and constants on the other hand.

6. Evaluating the model: part of this evaluation includes an estimation of the validity to be obtained upon actual application, but the evaluation may extend to the model’s utility as well.

**Sign Versus Sample-Based Prediction.** In designing prediction models, two basic prediction principles can be followed. Wernimont and Campbell (1968) have called these approaches “sign” versus “sample.”

The sign approach has its basis in some law (or nomos) on human behavior. Generally speaking, a law states that for a given set of people a relationship exists between characteristic A and a certain type of (future) behavior E. If the law is given, it suffices to know that the characteristic A is present in order to predict that behavior E will occur. This type of deductive reasoning applies not only to deterministic laws but also to probabilistic laws (see Stegmüller, 1974; Roe, 1983). The sample approach rests on the principle of generalization. The basic idea is that when it is known how a person will behave on a sample of occasions, one may generalize to behavior on other occasions belonging to the same universe. In this way the trait concept is avoided. The line of reasoning is as follows: a definition is given the domain of work, or performance, by using some type of task analysis; from this domain of work or universe of tasks, a sample is drawn which is presented to the applicant, or judged on the basis of his or her former performance. The applicant’s performance in the sample is generalized to his or her future performance on the job. The assumption underlying the generalization process is that of “behavioral consistency”: comparable tasks and conditions will produce similar behavior results for the same person. Because of the content similarity between predictor and criterion, this type of instrument is referred to as a “content based selection device.”

In spite of the differences between the two methods of prediction, there are some common design problems. In the sign approach, performance is modeled according to the formula \( Y = f(X) \) (cf., Naylor, 1983). The question is how to define \( Y \), how to identify suitable \( X \)’s, and how to choose an appropriate function \( f \). For the sample approach, the basic formula is

\[ Y = f(Y') \]

The identifier domain as \( Y \). On the the assumed stability possible changes in w and so on.

There is little devi cies go; both are we as a result of validity come more popular a in the United States.

**Defining Prediction M**

**Choice of Criteria.**

The theme for selection Kendall, 1955; Ronan in criterion theory has like to pinpoint three Thorndike (1949) criteria. The ultimate in a particular type of job activities; it can be susceptible to an easy to accomplish the alteration, one proceeds criteria of success. This indicates the relative introduces the notion of index of success as an activity. “Relevancy” tion coefficient betw A shortcoming of not recognize the fact performance behavior study, the degree to the organization’s by managers. The sa ferent managers, or l on how they define has suggested leaving
\[ Y = f(Y') \]. The identification of \( Y' \) is easier because it stems from the same domain as \( Y \). On the other hand, the designer should answer questions on the assumed stability or consistency of work behavior, taking into account possible changes in working conditions, leadership, learning opportunity, and so on.

There is little development as far as the application of these two principles go; both are well accepted. The sign approach has remained popular as a result of validity generalization studies; the sample approach has become more popular as a result of anti-discrimination litigation, especially in the United States.

**Defining Prediction Models, Contents**

**Choice of Criteria.** The choice and definition of criteria is a traditional theme for selection psychologists (see, for example, Thorndike, 1949; Kendall, 1955; Ronan & Prien, 1971; Smith, 1976). The greatest progress in criterion theory has been accomplished in the past. Below, we would like to pinpoint three major developments.

Thorndike (1949) introduced the notion of ultimate versus substitute criteria. The ultimate criterion is the complete final goal of performance in a particular type of job. It represents the "true" order of success in the job activities; it can only be stated in very broad terms that are often not susceptible to an easy practical quantitative evaluation. Instead of trying to accomplish the almost impossible task of predicting the ultimate criterion, one proceeds in practice by predicting more concrete, "substitute" criteria of success. The distinction between ultimate and substitute criteria indicates the relative value of a diversity of performance measures and introduces the notion of criterion "relevance"; the extent to which an index of success as applied is related to the true order of success in a given activity. "Relevancy" can be conceptualized as the hypothetical correlation coefficient between the criterion used and the ultimate criterion.

A shortcoming of Thorndike's way of conceptualizing is that it does not recognize the fact that criteria are mixtures of facts and values. While performance behaviors and outcomes are facts that are open to empirical study, the degree to which the performance is considered to contribute to the organization's goals is a matter of judgment or utility assignment by managers. The same performance may be judged differently by different managers, or by the same managers at different times, depending on how they define organizational goals. For this reason, Roe (1983) has suggested leaving organizational goal attainment out of the criterion
notion, thus limiting it to performance and performance outcomes. Moreover, he has suggested defining criteria both at the level of theoretical notions and at the level of measures that operationalize these notions. In this chapter, we will follow his suggestion and speak of job goals, conceptual criteria, and operational criteria.

A second issue pertains to the multidimensionality of performance behaviors. An overwhelming majority of studies involving statistical analyses of sets of criterion measures (i.e., factor analysis) finds that these analyses rarely yield one single factor. Job performance tends to be complex and “multidimensional,” i.e., reflects various independent aspects (Ronan & Prien, 1971). This multidimensionality gives rise to another classical problem in the history of criterion development: how many criteria should be utilized? Is it advisable to use a single overall measure or should one operationalize multiple criteria? This issue has been shown by Schmidt and Kaplan (1971) to be a pseudo-controversy. Both approaches can be of value: multiple criteria when predicting behavior and single overall criteria when making decisions about applicants.

A third development refers to the classification of criteria. Smith (1976) has developed a classification scheme for criteria. Three dimensions seem to cover most criteria:

1. The time span covered: criterion measures can be obtained either very soon after actual on-the-job behavior has occurred or many years afterwards;
2. The specificity desired: some criteria refer to specific instances of behavior, while others give rise to a global assessment;
3. Degree of closeness to organizational goals: criteria range from the description of actual behavior through the evaluation of immediate results to estimates of payoff for the organization.

Since Smith’s (1976) integrative article, with few exceptions, further developments in criterion theory and criterion development have been modest, to say the least. A review article of Barrett and associates (1985), has sparked off some discussion about the concept of “dynamic criterion” (Austin et al., 1989; Barrett & Alexander, 1989).

Floshman’s work culminated in a worthwhile contribution to criterion development. In the book Taxonomies of Human Performance, by Flabishman and Quaintance (1984), several bases for the classification of job performance are discussed. In the “behavior description approach,” categories of task activities are formulated based on observations and descriptions of what job incumbents actually do while performing a task.
emphasis is placed upon a description of overt behavior as manifested. The "behavior requirement approach" relies on the cataloguing of behaviors that should be emitted or are assumed to be required in order to achieve desired criterion levels of performance. In the "ability requirements approach," tasks are described, contrasted, and compared in terms of the abilities that are required for job performance. Finally, human performance can be classified on the basis of the "task characteristic approach." This approach is predicated on a definition of work performance that treats the task as a set of conditions that elicit and stimulate performance: tasks can be described in terms of these "triggering" conditions, placing emphasis on aspects such as task instructions, procedures, motivational contingencies, and so on.

In almost every instance, conceptual criteria are chosen on the basis of job analysis methods, especially those job analysis techniques that can be attributed to the behavior description approach mentioned above. This can be useful in stipulating conceptual criteria. In this category, a further subdivision into job-oriented (work-oriented, task-oriented) and worker-oriented methods can be made. Job-oriented methods result in a description of specific job activities, also taking into account special instruments, materials, and equipment. A recent example is the Occupation Analysis Inventory (OAI) (Cunningham et al., 1983). This instrument contains 622 individual items that fall into five generalized categories: information received, mental activities, observable behaviors, work content, and work context. The OAI attempts to achieve as much specificity in occupational descriptions as possible while maintaining its applicability to the entire occupational spectrum. Worker-oriented methods lead to a more general description of job activities. The specific job context is not considered in detail; job activities are described by means of verbs indicating relevant performance elements: for example, making decisions, reading, manipulating objects, using far-reaching tools, and so on. The traditional example is the Position Analysis Questionnaire (PAQ) (McCormick, 1972).

Which job analysis method is preferable in a specific instance depends on the nature of the prediction model (Algera & Greuter, 1989). A prerequisite for applying the sign approach is that work performance is conceptualized in an abstract manner, thereby incorporating performance dimensions that can be connected, at least in principle, to capacities, personality traits, and other stable characteristics. For this purpose, worker-oriented job analysis techniques such as the PAQ are recommended. For developing content-based selection devices following the sample approach, very detailed information on the job and the job context is necessary in order to arrive at a complete description of the job domain,
and to simulate relevant parts by means of work samples. Job-oriented techniques are to be considered here.

The critical incident technique (Flanagan, 1954) might serve a dual purpose: (1) incidents of superior, average, and inferior performance do lend themselves to direct simulation in a work sample; and (2) incidents can be categorized in more abstract behavioral dimensions that form the basis for a sign-type prediction.

Studies using Schmidt and Hunter's (1977) validity generalization approach (to be discussed later) have cast some doubt on the use of job analysis. The general finding has been that more refined job analysis techniques do not lead to higher validity estimates than global techniques. It has even been suggested that the job title would give sufficient information for identifying predictors and their validity within a database. Pearlman and associates (1980) concluded, on the basis of an extensive study on the relative merits of job analytic methods within the context of validity generalization, that refined ("molecular") analytical techniques are not needed. Or, in the words of Schmidt and colleagues (1985; p. 724), "validity generalization studies do not require fine-grained, task-based job analysis; job analysis does not have to be detailed." It should be noted that this type of research is based on validity coefficients that relate to overall performance ratings in very broad job categories. It is evident that, within this context, fine-grained job analysis methods are of little use. When the prediction model has to produce information on different aspects of performance rather than on overall suitability of applicants, and/or when it should be applied within a narrow range of jobs, fine-grained methods are, of course, indispensable.

For the purpose of evaluating prediction models it is necessary to dispose of operational criteria. In contrast to the limited attention for criterion development at a conceptual and theoretical level, there is a considerable amount of research aimed at the question of how criteria can be operationalized (see, for example, Landy et al., 1983; Landy & Farr, 1983; Latham, 1986). Interest continues in the effect of the rating format, although several review articles have made clear that further refinements of rating scales do have zero effect on rating accuracy and rating error tendencies as halo. "A moratorium on rating scale development is in order" (Landy et al., 1983, p. 6). Improving performance rating by training is another traditional theme. Bernardin and Buckley (1981) suggested that training does succeed in reducing response tendencies such as halo and leniency; on the other hand, rating accuracy is not improved because training only replaces one response tendency with another. A more optimistic view results from a review by Smith (1986), who evaluated 24 studies on training effects. See rating errors, how to do it, how to come to a clear definition, etc.) are relevant as contributing process work after all, although not every kind of most widely used training process is appropriate for improving to apply standardized procedures.

A final point must be made: processes that may affect, for example, Feldman (1982). This kind of understanding of performance, has not led to any in practice. Almost all in laboratory settings, these contrived appraisals in real work settings. Choice of appraisals, such as outcome expectancies, laboratory (Igen & Fa

Choice of Predictors.

be taken to develop a section of conceptual pred (2) operationalization is commercially or otherwise.

The first (and very relevant to future success) predictors with high requirements may be task-oriented. Examine duration of administration, and so on. The prediction process determines what type of oriented ("content-based"") predictors to choose from four
on training effects. Several training methods (learning how to suppress rating errors, how to develop stable standards that are uniformly applied, how to come to a clearer understanding of the meaning of performance dimensions, etc.) are reported as effective in reducing leniency and halo or as contributing positively to rating accuracy. Training does seem to work after all, although the available research makes it abundantly clear that not every kind of training works for all purposes: for example, the most widely used training approach, called “rating error training,” is inappropriate for improving rating accuracy, but other methods (e.g., learning to apply standardized performance standards) do have positive effects.

A final point must be made regarding the increasing interest in cognitive processes that may affect the measurement of performance criteria (see, for example, Feldman, 1981; DeNisi et al., 1984; Carroll & Schneier, 1982). This kind of study has enough potential for a better theoretical understanding of performance behaviors and their determinants but, until now, has not led to any viable suggestions for improving the rating process in practice. Almost all of the cognitive research has been implemented in laboratory settings, and there is not much evidence that the results of these contrived appraisal processes do generalize to performance appraisal in real work settings. On the contrary, critical aspects of real life performance appraisal, such as rater motivation, rating purposes, rater experience, outcome expectancies, rater resentment, and so on, are ignored in the laboratory (Ilgen & Favero, 1985; Banks & Murphy, 1985).

**Choice of Predictors.** Once the criteria are established, two steps must be taken to develop a set of suitable predictors (Roe, 1983): (1) identification of conceptual predictors, i.e., required traits or critical behaviors; and (2) operationalization into operational predictors by selecting tests that are commercially or otherwise available, or by constructing new instruments.

The first (and very obvious) requirement is that predictors should be relevant to future success criteria. Assessors should look for conceptual predictors with high validity and low intercorrelations. Several other requirements may be of importance as well, especially in finding operational predictors. Examples of such requirements are intrusion of privacy, duration of administration, costs, personnel implications, and practical aspects (paper-and-pencil).

The prediction principle that has been preferred (signs or samples) determines what type of predictors will be chosen: trait-oriented, behavior-oriented ("content-based selection devices"), or both. The strategy for identifying predictors at the conceptual level is the same, however. One can choose from four options: (1) meta-analysis of published validity
(validity generalization); (2) theoretical analysis; (3) job analysis; and (4) empirical try-out (exploratory validation).

Published validity data can be of help in identifying and selecting relevant predictor variables. Older reviews of validity data are those by Ghiselli (1966, 1973), Lent and associates (1971), Lawshe and Balma (1966), Guion (1965), Bemis (1968), and Asher and Sciarriano (1974, work samples). More recent reviews are given by Arvey and Campion (1982; the interview), Reilly and Chao (1982; various predictors), Campion (1983; selection for physically demanding jobs), Sackett and Harris (1984; tests for honesty), Hunter and Hunter (1984; various predictors), Schmitt and associates (1984; various predictors), Gaugler and associates (1987; Assessment Centers), Reilly and Israeli (1988), and Robertson and Downs (1989; trainability tests).

Traditionally, published validity data were analyzed in an informal, narrative manner. More recently, reviews are characterized by the use of meta-analytical techniques (see chapter 6).

Within the field of industrial and organizational psychology, the validity generalization method (Schmidt & Hunter, 1977, 1981) has become the prevailing method of analysis. The method has been applied to personnel selection as well as to several other subjects.

The validity generalization method has not remained undisputed, however (see Algina et al., 1984; Burke, 1984; James et al., 1986; Jensen et al., 1986; Kemery et al., 1987, 1989; Osburn et al., 1983; Paese & Switzer, 1988; Rasmussen & Loher, 1988; Roe, 1984; Sackett, et al., 1985, 1986; Spector & Levine, 1987; and Thomas, 1988). Some major points of criticism are: (1) the lack of classification rules that can prevent compilation and analysis of validities with widely differing referents (types of tests, criteria, and jobs), (2) the lack of power of the procedure used for analysis and validation of validities with widely differing referents (types of tests, criteria, and jobs), (2) the lack of power of the procedure used for testing situational specificity; and (3) inadequacies in the procedure for correcting for artifacts and estimating true validity.

A second method for identifying predictor constructs is by using theoretical analysis; however, this is the least developed approach. The principle is that one uses some type of job analysis for describing the tasks and the behavioral or trait requirements implied in them. In the sample approach, this idea is more or less straightforward. One needs some refined method for analyzing tasks and behavioral requirements, such as the Systems Task Vocabulary (Miller, 1973). A less elaborate method, which aims at defining a roster of task elements that can be helpful in defining critical behaviors, has been published by Guion (1978).
With the sign approach, the derivation of predictors is more cumbersome, since knowledge about the trait prerequisites for behavior patterns and behavior outcomes is far from complete. The only published systematic work in this domain is that of Fleishman on the Task Abilities Scales (TAS) (Fleishman & Quaintance, 1984). Here, the job analyst should indicate to which degree a fixed set of ability factors is supposed to play a role in the job. The scales are anchored with examples of tangible work behaviors, which were found by empirical research and expert studies. For some of the abilities, decision trees have been developed; this implies that the analyst can confine himself to a series of dichotomous (yes/no) decisions in order to determine whether a given ability factor is relevant.

The third approach is more or less related to the second in the sense that it tries to determine the traits required while performing the job analysis. The difference is that the theoretical phase is lacking; the inference of required traits is made directly by the job analyst, the job incumbent, or some other “job expert.” An example of this approach can be found in the Minnesota Job Requirements Questionnaire (MJRO) (Desmond & Weiss, 1973, 1975), which asks job incumbents and superiors for the relevance of problems that are used to define aptitudes. Another example is the Threshold Trait Analysis (TTA) (Lopez et al., 1981). Here it is the job analyst who should interpret which traits are required for adequate job performance.

In addition to judging trait relevance directly, one can also opt for an indirect procedure. The job content is described first, but it is followed by a mechanical, empirically based identification of required personal characteristics. The traditional example is that of the Position Analysis Questionnaire (PAQ) (McCormick et al., 1972). The factor structure of the PAQ has been established in two ways: (1) correlations between job elements (job data) and (2) correlations between personal characteristics (attribute data). According to subjective standards, both factor structures appear to be similar, which implies that it is possible to place an empirically based relationship between criterion and predictor variables. In addition, the relationships between job dimension scores and GATB test scores have been examined in two studies (McCormick et al., 1972, 1979). These results are promising as well, but they do not supply rules for the choice of predictors in a specific case. Also, it should be noted that considerable criticism has been leveled at the PAQ. Points of criticism are: the PAQ is hard to read (Ash & Edgell, 1975), and it would only measure common sense knowledge about jobs (Smith & Hakel, 1979; Cornelius et al., 1984). Recently, Harvey and Theodore (1986)
have shown, on the basis of Monte Carlo studies, that the reliability values that have traditionally been found can be expected on the basis of mere chance. This is a result of the great number of "does not apply" answers; values of .50 are obtained when 15 percent to 20 percent of the items are answered with "does not apply" and the rest of them answered at random. A study by DeNisi and associates (1987) shows that large numbers of "does not apply" answers can also artificially inflate the agreement between expert and naive raters, thus incorrectly giving rise to the proposition that expertise is not an important factor contributing to accurate job descriptions.

A more recently developed method is the Occupational Analysis Inventory (OAI) (Cunningham et al., 1983); the developmental research is of the same nature as the PAQ. Ratings of 1,414 jobs have been made on a total of more than 600 job elements. Also, ratings for 102 human attributes were made ("attribute ratings"). Factor analyses were conducted separately for the element ratings and the attribute ratings. The resultant factors were intuitively meaningful. Furthermore, the OAI dimensions were significantly related to tested abilities of relevant job holders. Most striking is that the OAI seems to offer a certain generalizability in spite of the fact that it is "job oriented."

The fourth approach to the problem is, in fact, the classical way of working (Guion, 1965). The exploration aims at a tryout of criteria and predictors and the generation of hypotheses about possible relationships. Unfortunately, because of the rise of validity generalization, exploratory methods have lost much of their significance. But explanatory validation remains the only viable alternative when a relevant database is lacking or when jobs and situational contexts are rapidly changing.

In the preceding discussion, methods for choosing predictors that are based on expert judgments were excluded. Experienced researchers in the field of personnel selection seem to be able to make accurate estimates of the validity of predictors. According to Schmidt and associates (1983), the combined judgment of 20 experts yielded a validity estimate as accurate as a "criterion-related" validity study with N = 981. Experts with less experience, although "trained professionals in the field of personnel selection," do perform considerably less well, but still the combined judgment of 20 persons is as accurate as a local validity study with a sample size of 217 (Hirsch et al., 1985). From a logical point of view, it does not make sense to make use of expert judgment if sufficient data about comparable jobs are available. When such data are lacking, the use of expert judgments could be considered, but, since their value depends on several factors (Hamilton & Dickinson, 1987), the results would need to be interpreted carefully. Exploratory v beyond this case.

With regard to the progress has been made by 1986; and Zedeck & Cas described. First, one can lead to a certain reappraisal of the sign approach, i.e., a matter of growing cor the development of the developm validity generalization s same time, it is noted the lower than those of bia.

Secondly, there is a devices," the instrument tion. Probably, this is n but also of the good rep the United States in leg method are the elaborat the empirical content of method as such is a critical incidents meth (Smith & Kendall, 1963 tion. Nevertheless, seve there exists a well-develpment for a specific situ principle. Schmitt and steps: job analysis, task edge, skills, aptitudes, an KSAOs, selection tests. Sackett (1987) ha of instruments goes be presentation of stimu important aspects as we formulated more preci the development of pr Lawshe (1975) develop (CVR). Other indices, be found in Jones and :

Although the behavi to job samples, several
carefully. Exploratory validation seems to be a more dependable strategy in this case.

With regard to the development of operational predictors, little progress has been made during recent years (Guion & Gibson, 1988; Hakel, 1986; and Zedeck & Cascio, 1984). The main trends can be summarized as follows. First, one can observe that validity generalization research has led to a certain reappraisal of the predictors that are associated with the sign approach, i.e., ability and aptitude tests. However, this is more a matter of growing confidence in existing instruments than an increased interest in the development of new instruments. In fact, publications on validity generalization seem to have hampered test development. At the same time, it is noted that predictive validities of tests tend to be somewhat lower than those of biodata and job samples (Schmitt et al., 1984).

Secondly, there is a still-growing interest in “content-based selection devices,” the instruments that are based on the sample approach to prediction. Probably, this is not only a consequence of higher validity estimates but also of the good reputation that this kind of method has gained within the United States in legal cases on discrimination. Characteristics of this method are the elaborate job analysis as the first step of construction and the empirical content of the consequent construction steps. The construction method as such is not very new, as the roots of this approach lie in the critical incidents method (Flanagan, 1954), the retranslation technique (Smith & Kendall, 1963), and the sampling theory of statistical generalization. Nevertheless, several refinements have been proposed, and at present there exists a well-developed method for constructing a predictor instrument for a specific situation on the basis of the “behavioral consistency” principle. Schmitt and Ostroff (1986) distinguish between the following steps: job analysis, task generation, KSAOs generation (KSAO = knowledge, skills, aptitudes, other characteristics), importance ratings of tasks and KSAOs, selection of most important KSAOs, and construction of tests. Sackett (1987) has drawn attention to the fact that the development of instruments goes beyond the construction of stimulus materials. The presentation of stimulus materials and the scoring of responses are important aspects as well. Not only have the construction methods been formulated more precisely, but progress has also been made with regard to the development of procedures for the assessment of content validity. Lawshe (1975) developed a quantitative index, the Content Validity Ratio (CVR). Other indices, measuring some form of interrater agreement, can be found in Jones and associates (1983).

Although the behavioral consistency approach has initially been limited to job samples, several other selection methods have been developed on
this basis: biographical inventories and biodata (Owens, 1976; Pannone, 1984), application forms for training and experience evaluation (Ash & Levine, 1985), accomplishment records (Hough, 1984), and interviews (Latham et al., 1980).

Third, assessment centers (see chapter 8) which combine the sign and sample approaches are still of considerable interest. But, according to the review by Schmitt and associates (1984), the validity has not been so high as initially expected. The highest validity is found against rating criteria \( r = .43 \) which is not impressive because of common method variance and possible contamination of the criterion (Klimoski & Brickner, 1987). Objective criteria such as training results and salary level are predicted less well by assessment centers. In a recent meta-analysis by Gaugler and associates (1987) of 50 assessment center studies containing 107 validity coefficients, a mean validity coefficient (corrected for several statistical artifacts) of .37 was reported. Higher validities were found in studies in which potential ratings were the criterion (corrected \( r = .53 \)); promotional criteria are less well predicted (corrected \( r = .30 \)). The validities for other types of criteria (overall performance rating, dimensional ratings, success in training, career advancement criteria) and other purposes (early potential identification, selection, research) lie somewhere in between. All in all, these findings illustrate that assessment centers are rather successful in making valid predictions.

**Further Model Specification**

**Structure and Format.** Once the criteria and predictor variables have been identified, the structure of the model has to be established and parameter values have to be specified. From the various possibilities that exist here (e.g., conjunctive and disjunctive models), the linear compensatory model is the most well known and seems to be most frequently applied for both sign and sample-based prediction models. Discrete models can be mentioned as well, but their applicability seems limited: critical requirements that can be considered as “conditions sine qua non” do not occur frequently and are, at best, exemplary exceptions. The format of prediction models is usually algorithmic, which makes them suitable for computer processing. The tabular and graphical formats that were favored in classical textbooks seem to have lost much of their popularity.

A class of models that has been mentioned quite often during the last two decades is the moderator model. Initially, this type of model was proposed as a means for a more fair selection. The moderator variable, then,
is a racial or other external characteristic. More recently, moderator models have been mentioned in the context of validity generalization. In these cases, the moderator variable is a situational or job characteristic that could serve to increase the homogeneity of a set of predictor variables.

Regarding the first application, it can be concluded that efforts to demonstrate moderator-effects for racial and group characteristics have generally failed (Schmitt & Noe, 1986). Whenever moderator-effects are observed, they are caused by differences in regression constants rather than regression weights of separate predictors. In the rare instances that differences are found, criterion scores of minority group members are, surprisingly, not underestimated, but rather overestimated. Although this has been observed several times, a satisfactory explanation is still lacking. For one part, one could think of statistical explanations: small sample sizes and little power to detect moderators. This explanation is not completely satisfactory, however, since no evidence for systematic underestimation has been found in larger samples.

Regarding the second point (situation and/or job characteristics as moderators for predictor criterion relationships within occupational classes as used for validity generalization purposes), the discussion is still going on. A study by Gutenberg and associates (1983) showed that the “information processing/decision making” job dimensions of the PAQ moderate the validity coefficients of several GATB dimensions (general ability, verbal and numerical ability). In addition to job characteristics of this type, the situational arrangement of a job, i.e., the work setting, is also a potential source of moderators (Greuter, 1988).

In applying the moderator approach, several serious statistical problems arise. A search for moderators is usually carried through by introducing a product or interaction term (predictor × moderator) in ordinary least-squares multiple regression analysis, so-called moderated regression analysis. Morris and associates (1986) pointed to the high correlation between this product-term and its constituent predictors, introducing linear dependencies in the set of regressor variables (see also Sockloff, 1976). The power of the (traditional) F-test for interaction can be quite low because of the relation among the regressor variables. As a consequence, moderator effects may have a diminished opportunity for detection, even when sample size is adequate. Morris and associates recommended principal regression analysis on the principal components of the predictor set in which the smallest principal component was deleted. However, their remedy may lead to an overestimation of the interaction effect as has been put forward by Cronbach (1988). Dunlap and Kemery (1988) demonstrated in a series of Monte Carlo simulations that detection of
moderator effects in regression analysis is hampered by unreliability in either the predictor or the moderator variable. Although (un)reliability can always be acknowledged as a problem in prediction, it has a greater impact on the more complex moderator model: the reliability of the product-term is partly determined by the product of the reliabilities of its constituents. More precise and accurate measurement of regressor variables is called for in order to have a reasonable chance of detecting moderating effects.

**Parameter Estimation.** Parameters can be estimated with the help of either empirical or rational methods. The empirical methods can only be applied when a complete set of data is available. This poses a problem when validity generalization (or another type of meta-analysis) has been used as an alternative to an empirical validity study. Although the method could, in principle, be used for estimating intercorrelations as well, the required data are usually lacking, thereby forcing the designer to perform an empirical study after all, or to resort to rational methods.

The empirical methods for parameter estimation include the classical (unbiased) multiple regression analysis, and (biased) Stein and ridge regression methods. Multiple regression, according to Stein (1960), differs from classical multiple regression, because the regression weights are corrected with a factor that is similar to the shrinkage correction for multiple correlations. Since this correction factor is identical for all regression variables, the multiple correlation remains the same. The corrected weights are more reliable estimators of their population counterparts. Ridge regression is based on the same principle. Again, a correction of regression weights is carried through, but here the correction applies to the principle components, and its magnitude is variable depending on the eigenvalues. According to Darlington (1978), ridge regression leads to more reliable estimates of parameters than traditional methods, especially when there is a high degree of validity concentration, i.e., large differences in validities between principle components. Rational methods assign weights to predictors on the basis of expert judgments. The simplest method is unit weighing, but one can also ask judges to assign differential weights.

The Bayesian approach can be seen as a combination of empirical and rational approaches to parameter estimation. It offers a general method to specify a priori hypotheses on regression weights and revise these on the basis of empirical data. By applying this method iteratively, the estimates converge to their final values when more data are added. General Bayesian methods have been described by Lindley and Smith (1972) and Laughlin (1979). The method by Laughlin starts from a prior distribution of equal regression weights. These weights can be adjusted on the basis of empirical evidence. The method weight to the important of the Bayesian method comparable groups (n) have improved the group's associates (1971). It be used to establish w prediction of grades 1985 showed, with t generalizable within se

**Evaluation of Models**

Evaluation should be specified in advance. The case of empirical para model is usually obtain additional analysis. The biased regression met

Recently there has been Groenendijk, 1985; into components for e
dity of the overall pr methods such as the J Dickinson, 1987). Acc components may be sp or skills, provided that of behavior. The defi measured with criteria estimated rationally wi and supervisors estima job elements and total relati parameters l

The rationale for the validities when an en when a relevant datab
The concept can be a
sizes, inadequate perfo
evidence. The method contains a factor by which the designer can assign a weight to the importance of the empirical data. A well-known application of the Bayesian method concerns the estimation of regression weights in comparable groups (m group regression). Molenaar and Lewis (1979) have improved the method that was originally proposed by Jackson and associates (1971). It should be noted that this m group method can be used to establish whether validities are generalizable. In a study on the prediction of grades from different curricula, Dunbar and associates (1985) showed, with the help of this technique, that validities are only generalizable within selected groups of criteria.

**Evaluation of Models**

Evaluation should be done against the requirements and constraints specified in advance. The main requirement is predictive validity. In the case of empirical parameter estimation, a validity estimate for the total model is usually obtained simultaneously. Rational models may require an additional analysis. There are few new developments here, apart from the biased regression methods mentioned above.

Recently there has been renewed interest in *synthetic validation* (Algera & Groenendijk, 1985; Mossholder & Arvey, 1984). A job is broken down into components for each of which a validity estimate is made. The validity of the overall procedure can be synthesized by rational/empirical methods such as the J-coefficient of Primoff (1955; see also Hamilton & Dickinson, 1987). According to Hamilton and Dickinson (1987), the job components may be specified in a job analysis as behaviors, traits, abilities, or skills, provided that the elements used can be conceived as determinants of behavior. The defining component relations of the J-coefficient can be measured with criterion-related validation designs but they can also be estimated rationally with the help of expert judgment, i.e., job incumbents and supervisors estimate indicators for job performance (relations between job elements and total job performance) and test experts give estimates of relational parameters based on test performance (validities).

The rationale for the J-coefficient offers great potential for estimating validities when an empirically based validation study is not feasible, or when a relevant database for conducting a form of meta-analysis is lacking. The concept can be applied most conveniently with inadequate sample sizes, inadequate performance ratings (or no ratings at all), new jobs without incumbents, rapidly changing jobs, or (unknown) situational variables that modify job performance or validities, and so on.
Decision Making

Major Design Considerations

While prediction models serve to yield information on future performance of candidates, the function of decision models is to transform this information into plans for action. In order to fulfill this function, some analysis has to be made, taking into account such factors as expected criterion performance, utility of these performance outcomes, and possibilities for planned action. All this can be done with the help of special analytical schemes, several of which are available in the literature. In addition, a decision strategy is needed that embodies a certain principle for dealing with utilities and uncertainty. Such strategies can be borrowed from (normative) decision theory. Some examples are:

1. Maximization of utility (in case of certain outcomes);
2. Maximization of expected utility (in case of decision making under risk);
3. Maximin, Maximax, Minimax regret (in case of decision making under full uncertainty).

Below we will discuss how features of decision models depend on the analytical scheme and the decision strategy that has been adopted. To be sure, decisions can be made, and actually quite often are made, on an informal judgmental basis, without the help of a decision model. Studies on the judgmental processes involved and on efforts to model them (so called paramorphic representation) have been discussed by Wiggins (1980) and Roe (1983).

For describing the design of decision models we can once again use the design methodology that was presented at the beginning of the chapter. By doing this, we obtain the following general description of the design of decision models:

1. Defining the problem and specifying requirements: how is the predictive information on candidates transposed in decisions such as accepting or rejecting an applicant?
2. Defining the model’s content: what criterion and utility variables must be defined, and what are the possible actions to decide upon?
3. Choosing the model’s structure: how are utilities related to criterion performance?
4. Choosing the model's format: again, usually an algebraic function is chosen.

5. Estimating model parameters: specifying the precise form of the relationship between each utility variable and the set of criterion variables under consideration by estimating weights of criterion variables (relational parameters) and calculating cutoff scores (instrumental parameters).

6. Evaluation of the model: analyzing the differences between expected utilities and actual utilities accomplished by applying the decision model.

**Defining Decision Models' Content**

Decision models can take many different forms. Some commonly used elements are:

1. Two or more possibilities for action, i.e., accepting, rejecting, assigning one or more candidates;
2. One or more criterion variables referring to job performance, on which expected scores or probabilities of scores can be expressed;
3. One or more utility dimensions, expressing the value of job performance for the organization, either subjectively or objectively.

Sometimes the models are extended by incorporating:

4. One or more predictor variables that can be associated directly with either utility dimensions or actions;
5. Costs associated with collecting and processing information on candidates.

Possible actions follow from the analysis of the organization's personnel plan and the current workforce situation. Utilities and criteria can be found in a utility analysis. This covers: (1) establishing the number and nature of utility dimensions; and (2) assessing the relationship of utility dimensions with specific criteria (e.g., by making use of published criterion studies; see, for example, Brogden & Taylor, 1950; van Naerssen, 1962). Criteria and predictors can be chosen on the basis of previously established prediction models.

It should be noted that decision modeling for personnel selection purposes is far less developed than prediction modeling. Some general schemes for analyzing decision situations that can be applied to personnel selection
have been published in the context of decision theory. In addition, there are a number of methods and models that have been developed for personnel selection, but the older models tend to be poorly conceptualized, and little developmental conceptual work has been performed during the last decade.

Further Model Specification

Structure and Format. The structure of the decision model can only be established after the elements have been defined. As has been shown, the elements that take a central place are utility dimensions and criteria. Hence, a crucial question is how to relate them to each other and how to express the relationship. When the usual algebraic format is adopted, this comes down to the specification of a utility function. The general form of the utility function is \( U = f(Y) \). By making certain assumptions regarding its shape or properties, one can specify this function. Well-known types of utility functions are the threshold function (e.g., Gross & Su, 1975), the linear function (Mellenbergh & van der Linden, 1977), and the normal ogive function (van Naerssen et al., 1986; Novick & Lindley, 1978).

The relationship between criteria and utility can also be established by the direct assignment of numerical values. This implies that only minimal assumptions have to be made (Hull et al., 1973).

Within this latter category, one finds objective methods based on human resources accounting (see, for example, cost-accounting approach as described by Cascio, 1982, pp. 154–156), as well as subjective methods based on ratings. A usable rating method has been proposed by Mellenbergh (see Vrijhoff et al., 1983; Gaag et al., 1986). This method has been subjected to a tryout in the context of pass-fail decisions in education. In the 1986 study, 10 teachers and 10 students were asked to give their judgment about the utility of passing and failing on examinations in the English language, French language, and biology. The utilities were scaled according to the constant sum method. It was concluded by the researchers that the method was quite reliable, that there were only limited differences between the utility functions of teachers and students, and between the utility functions for the different examinations, and, furthermore, that two of each three utility functions could be approximated by a linear function.

In order to arrive at decisions, a relationship should be laid between the utility functions and possible actions. This may be done with the help of criterion cutoff scores. The procedures for finding such cutoff scores will be discussed later.
Decision models may include predictor variables as well. Such extended models offer the possibility to define cutoff scores on the predictor variables, which simplifies the total selection process. In fact, it means that the predictor model is incorporated, either partially or completely, within the decision model. For the relevant choices regarding the structure of the model, we can therefore refer to an earlier section of this chapter.

Incorporating costs in the decision model means splitting up fixed and variable cost components of selection, and relating variable costs to the parts of the procedure. Usually a linear cost function is employed, which relates the number of candidates to be examined to the overall expenditure. To be sure, costs have also been related to the length and composition of the examination procedure, but this is of no relevance in the framework of decision making on candidates unless an adaptive procedure is followed.

Parameter Estimation. Because the utility of selection outcomes is essentially a matter of subjective judgment, procedures are called for by which such judgments can be made explicit. The rating method, as well as accounting methods, have already been referred to. When the decision has been made to use utility functions, the parameters of these functions have to be established separately. This can be done by regression methods, taking the performance criteria as "predictors" and utility ratings as "criteria." The regression function can be of any linear or nonlinear type.

The same approach can be used for finding the parameters of cost functions. When predictor variables are included, the earlier material applies.

The determination of cutoff scores on either criteria or predictors requires the application of some decision strategy. As a rule, this is "maximization of expected utility" (MEU). The usual procedure to establish criterion cutoff scores for individual decisions is to determine the criterion score at which the two utility functions intersect, i.e., the utility of rejection equals that of acceptance. A related procedure exists for collective decisions. Here, one looks for the criterion score at which the accumulated utility for the group to be accepted equals that of the group to be rejected.

De Grujter and Hambleton (1984) and van der Linden (1984) have studied various utility functions with regard to the determination of cutoff scores. A general problem seems to be that the accuracy of utility functions for passing and failing (on examinations) is not sufficient to assess reliably the point of intersection, and hence the cutoff score. According to De Grujter and Hambleton (1984), a solution can be found in "robustness studies" which specify the range of population cutoff values within which one single cutoff score can be used. These authors also discuss problems that can arise in relation to utility functions and cutoffs for subpopulations.
Splitting up the total population into subpopulations can lead to small sample sizes which preclude the establishment of stable cutoff scores. The procedures for establishing predictor cutoff scores resemble those for criterion cutoff scores. In fact, the latter can be seen as a special case of the former, when predictors are assumed to have a linear relationship to criteria. A general assumption-free method for determining predictor cutoffs has been described by Roe (1983) both for the case of individual decisions and for collective decisions. The methods described by Stone and Kendall (1956) and Mellenbergh and van der Linden (1977) can be considered as special cases of the method for individual decisions. The methods described by Guttman and Raju (1965), Guilford (1965), Rorer and associates (1966), Darlington and Stauffer (1966), Alf and Dorman (1967), Cronbach and Gleser (1965), Naylor and Shine (1965), and Taylor and Russell (1939) can be seen as specific instances of the method for collective decisions.

The aforementioned methods do not apply to fixed quota situations. In such situations, an important consideration is whether the number of applicants should be considered as a given premise.

If the number of applicants cannot be extended (for example, by intensifying recruitment procedures), the only possibility is to manipulate the selection ratio, i.e., the corresponding cutoff score. If predictor, criterion, and utility variables are linearly related to each other, this comes down to selecting those applicants with the highest predictor-scores until the quota has been reached (top-down strategy). A serious drawback of this method is that cutoff scores can fluctuate considerably in time, due to quantitative and/or qualitative changes in the supply of applicants.

If the number of applicants is not to be considered as fixed, one can act as is indicated above for the quota-free situation, i.e., placing the cutoff scores at the intersection of the utility functions for rejection and acceptance. In this situation, it is necessary to define an optimal selection ratio that can be found, for instance, following the methods of Cronbach and Gleser (1965) or van Naerssen (1962, p. 363). A recent review on cutoff scores, including legal issues, has been given by Cascio and colleagues (1988).

**Evaluation of Decision Models**

A basis for the evaluation of decision models can be found in the deviation of estimated utilities from actually obtained utilities. In an analogy of the concept of predictive validity, which expresses the deviation between predicted and obtained values of decision validity as the degree to which the decision model works (Roe, 1983, p. 25).

Various specific approaches have been proposed (Hambleton 1974, 1975; Huynt, 1973, 1976) and external elaborated by Mels (1974, 1975). Another approach is the utility of decision: methods have been proposed by Naylor and Shine (1949; Cronbach & Meehl 1955).

The approach of by far the most practical psychology is the following function

\[
\Delta U = \text{dollar value of random selection} \times \text{number of utilities} \times \text{standard deviation of predictor scores} \times \text{validity of selection} \times \text{ordinate of point on cost function}
\]

Confidence intervals are typically computed by Alexa (1983) and other assumptions have (as are predictor scores) been considered to be contrasted with a priori.

The Brogden/C cost except for some
predicted and obtained criterion performance, one can define a concept of decision validity. In the same way, one could define decision reliability as the degree to which equivalent sets of estimated utilities correspond (Roe, 1983, p. 250).

Various specific coefficients for evaluating decisions have been proposed (Hambleton & Novick, 1973; Swaminathan, Hambleton, & Algina, 1974, 1975; Huynh, 1976). A general coefficient for evaluating the "internal and external optimality" (reliability) of a decision procedure has been elaborated by Mellenbergh and van der Linden (1977).

Another approach would be to compare the overall, i.e., cumulated, utility of decisions taken with a given model. For this purpose, several methods have been developed, for example, by Taylor and Russell (1939), Naylor and Shine (1965), and Brogden, Cronbach, and Gleser (Brogden, 1949; Cronbach & Gleser, 1965).

The approach of Brogden and Cronbach and Gleser is, at this moment, by far the most popular method in the field of industrial and organizational psychology. In this model, the total utility is expressed with the following functions:

\[ \Delta U = N_s \cdot SD_y \cdot r_{xy} \cdot \frac{\gamma}{\phi} - C \cdot N_s \]

where:
\[ \Delta U = \text{dollar value payoff from the selection program (as contrasted with random selection);} \]
\[ N_s = \text{number of selectees;} \]
\[ SD_y = \text{standard deviation of performance in dollars (calculated in the group of present employees);} \]
\[ r_{xy} = \text{validity of the predictor;} \]
\[ \phi = \text{selection ratio;} \]
\[ \gamma = \text{ordinate of the normal curve associated with } \phi; \]
\[ C = \text{per applicant costs of the selection program (variable and fixed costs).} \]

Confidence intervals for \( \Delta U \) can be calculated following a procedure developed by Alexander and Barrick (1987). In deriving this model, several assumptions have been made: utility functions are assumed to be linear (as are predictor-criterion relations), utilities for rejecting employees are considered to be zero, and \( \Delta U \) expresses the net gain in utility as contrasted with an a priori procedure of random selection.

The Brogden/Cronbach/Gleser model has remained almost unchanged except for some minor embellishments. Schmidt and associates (1979)
multiplied the benefit component of the model by the expected tenure of the hired cohort (i.e., $T$), pointing out that revenues are accumulating for each year of tenure. Further improvements have been suggested by Boudreau (1983a; see also Boudreau, 1989) and Cronshaw and Alexander (1985), such as including several financial/economic concepts such as costs of capital (discount rate), taxes, service value, and so on. Also, the effects of recruitment have been made explicit (Boudreau & Rynes, 1985). The accumulation of effects due to the repeated application of selection programs can be calculated with the so-called Employee Flows Utility Model by Boudreau (1983b).

Most of the research efforts are now directed at developing accurate procedures for estimating $SD_y$ in dollars. For a review of such methods as the 40 percent to 70 percent rule, the CREPID-procedure, and the percentile method of Schmidt and associates (1979), the reader is referred to Cascio (1982; see also Cascio, 1987). As far as the empirical evidence demonstrates (e.g., Burke & Frederick, 1986; DeSimone et al., 1986; Reilly & Smither, 1985; Weekly et al., 1985), no method seems clearly superior to others. A cautious conclusion may be that the more global estimation procedures (for example, the percentile method of Schmidt et al., 1979) give higher estimates. However, even these higher estimates can sometimes be lower than an empirically counted value (DeSimone et al., 1986). All together, it is not all clear where the research on $SD_y$ is leading (Guion & Gibson, 1988). If the goal is to get realistic and accurate estimates, there is a long way to go considering the very large standard errors of utility estimates as reported by Alexander and Barrick (1987). If the main concern is calculating relative utilities of alternative selection procedures, utility research has already accomplished too much; these latter inferences can probably also be made (fairly accurately!) on the basis of comparing validities alone.

**Designing Selection Procedures**

In the previous paragraphs, we have focused on two functions of selection procedures, i.e., prediction and decision making, discussing some technical aspects of modeling. It should be recalled that there are other functions as well, and also that every function can be implemented in a variety of ways when designing selection procedures. The problem faced by the designer is to make a choice from the numerous technical options in such a way that the singularities of the particular selection problem are optimally met. Given the reductive nature of the design task, it is impossible to give specific recommendations with regard to the composition of selec-

...
tion procedures. Much depends on the "program of requirements" and the ideas and experience of the designer. For an illustration, we would like to refer to Roe (1989), who gives a short account of a project that aimed at the redesign of a set of selection procedures for a psychological selection consultancy.

Generally speaking, there are at least four types of considerations that can play a role in the design of selection procedures (Roe, 1989):

1. Effectiveness considerations, which have to do with the appropriateness of the predictive information, and the correctness of decisions yielded by the procedure. These considerations may, for example, lead to the inclusion of certain types of questions on the application form, the use of certain specific aptitude tests, or a combined use of tests, biodata, and work samples.

2. Efficiency considerations, relating to the overall costs and benefits of the use of the selection procedure, and the contribution to both by its distinctive components. Efficiency may be increased by adopting cheaper information (e.g., school grades), reducing test length, or introducing multiple selection stages (Cronbach & Gleser, 1965).

3. Ethical considerations, relating to such aspects as intrusion of privacy, right to appeal, and nondiscrimination. Leaving questions on sensitive information for the end of the procedure, including a procedure for filing and processing complaints, or setting different selection ratios for applicants of different cultural backgrounds, may be the result of these considerations.

4. Managerial considerations, concerning the commercial side of personnel selection, as well as the organization of psychologists' and staff activities. These may lead to a standardization of test duration, to a test administration program that optimizes the use of manpower, or, alternatively, to automation of particular parts of the procedure for reasons of cutting labor costs.

Many of these aspects surpass the psychometric and decision-theoretic views that have dominated personnel selection literature during the last decades. Taking into account the results from the study referred to above, it seems to us that the technological approach to selection offers an adequate framework for dealing with all of these aspects simultaneously.

References


---


---


DEVELOPMENTS IN PERSONNEL SELECTION METHODOLOGY


Interest in the new a for special reasons to models of General M ogy is no exception t centers, the situation tings long before psy The breakthrough an the adoption of work well-known intelli ourselves to assessm goes back some 45 ye Thornton and Byh center methods, descen tly in military setti Office Selection Boar (OSS) in the early fort OSS to adopt a progr that of the English V assessment center pro