DOCUMENT RESUME

ED 039 576 CG 005 367

AUTHOR	Novick, Melvin R.; Jackson, Paul H.
TITLE	Research Bulletin: Bayesian Guidance Technology.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-69-83
PUB DATE	Oct 69
NOTE	56p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$2.90
DESCRIPTORS	Academic Achievement, *Computer Science, Educational Environment, Educational Research, Educational Testing, *Guidance Services, *Predictive Measurement, *Statistical Analysis, *Testing Programs

ABSTRACT

This paper presents a discussion of the use of educational tests in guidance services as seen in the light of modern developments in statistical theory and computer technology, and of the increasing demands for such services. A focus and vocabulary for this discussion is found in Turnbull's recent article on "Relevance in Testing." Following an introductory discussion of the need for guidance services some very recent work in Bayesian inference is reviewed and the implications of this work for educational research methodology are noted. Special attention is given to the Lindley equations which provide solutions for a number of problems in the comparative prediction of academic achievement. The suggestion here is that in a changing educational environment the Bayesian methodology can provide an increase in the effectiveness and applicability of such programs as Horst's monumental Washington Pre-College Testing Program. Comparative prediction is seen as an idea whose time has come. (Author)

RESEARCH BULLETIN

RB-69-83

BAYESIAN GUIDANCE TECHNOLOGY

Melvin R. Novick
and
Paul H. Jackson

Educational Testing Service
Princeton, New Jersey
October 1969

BAYESIAN GUIDANCE TECHNOLOGY

Melvin R. Novick and Paul H. Jackson

Educational Testing Service

This paper presents a discussion of the use of educational tests in guidance services as seen in the light of modern developments in statistical theory and computer technology, and of the increasing demands for such services. A focus and vocabulary for this discussion is found in Turnbull's recent article on "Relevance in Testing". Following an introductory discussion of the need for guidance services some very recent work in Bayesian inference is reviewed and the implications of this work for educational research methodology are noted. Special attention is given to the Lindley equations which provide solutions for a number of problems in the comparative prediction of academic achievement. The suggestion here is that in a changing educational environment the Bayesian methodology can provide an increase in the effectiveness and applicability of such programs as Horst's monumental Washington Pre-College Testing Program. Comparative prediction is seen as an idea whose time has come.

Some Vocabulary of Educational Technology

Personnel problems may be grouped together in many ways and the typical problem can usually be considered from more than one point of view. We shall be restricting ourselves here to those problems that can conveniently be viewed as problems in guidance and/or selection. Another very useful and contrasting grouping is that involving problems of classification. The classification problem has recently been studied in depth by Rulon, Tiedeman, Tatsuoka and Langmuir (1967). Further references to this area may be found in that book. Wolfe's (1969) review of that book gives some glimpse of the mathematically more sophisticated allocation methods now being used by the military services. An up-to-date review of Guidance and Counseling appears in the April 1969 issue of the Review of Educational Research.

The standard classification problem involves a closed system of assignment of each member of a group to one of several subgroups or classifications. These classifications are often defined by a subsequent training program, job assignment or, more generally, by subsequent treatments. The military services are typically concerned with a classification problem whenever a group of recruits completes basic training. Each recruit must then be given an assignment which involves sending him either to one of a number of service schools or to one of a number of on-the-job training programs. If each of these schools and each of these programs is viewed as a classification, then this personnel problem may be viewed as one of classifying each of the recruits. There is usually present in this situation a quota for each of the classifications which must be filled and a maximum number that may be assigned. Often these requirements are not totally demanding so that some latitude of choice may be permitted to the individual. Whenever substantial choice is present the problem may be viewed in the context of the guidance-selection paradigm.

The kinds of selection and guidance problems with which we shall be concerned occur each year at the point of transition for students from secondary school to university. The typical student will wish to consider entering one of a number of colleges, universities or other institutions of further education. Many factors will affect his final choice and one of these will surely be his expectation of his potential success within each of the various programs. The *guidance* problem with which we shall be concerned is that of developing statistical methods which will enable the student's high school, the *sending system*, to make accurate predictions for each student with respect to each college, university or other training program that he may be considering. Indeed an important task of any guidance service will be to suggest to students that they may be qualified to enroll in programs that they had not previously considered. Such services can encourage potentially qualified students whose background has not given them an expectation of college attendance, to consider this alternative. Predictions of performance will provide each student with one useful piece of information that will help him, with the assistance of his guidance counsellor, to make an informed and rational decision. In the pure guidance problem, as described here, the student is free to enroll in any of the programs he may be considering. In practice this will not be true for most students. However, the statistical methods developed for the pure guidance problem are equally valid when there are restrictions, provided only that a large measure of choice is left to most students. For example a peacetime volunteer army might find the guidance paradigm to be very useful while a mobilization army would surely find the classification paradigm more appropriate.

At the same time that the students and their counsellors are concerned with guidance, the university admissions officers are concerned with the *selection* of the "best" possible entering freshman class. In a pure selection problem it is assumed that there are more applicants than vacancies and that each *accepting system* is free to take just those students that it believes to be best qualified. The pure selection model (which in the educational context might well be called an acceptance model) is often approximated rather well and the statistical methods developed for this paradigm will find equally wide application even in those situations when it is not, provided only that some positions exist for which multiple applications have been received. Actually in most instances the better students receive acceptances from many colleges so that no college can be sure of getting every student it selects.

We shall use the term comparative guidance to describe any system of information transmittal designed to provide a student with information about two or more possible career opportunities. Horst's techniques of multiple absolute prediction (1955) and multiple differential prediction (1954) are two important techniques useful in comparative prediction.

Scientific Method and Humanistic Goals

The important distinction between the guidance-selection and classification paradigms is the degree of compulsion characterizing each system. The classification paradigm adopts a purely actuarial outline which, in the extreme, delegates to the computer irrevocably the task of assigning each person to an "optimal" treatment. The guidance-selection paradigm, however, leaves the choice of college by the student and the choice of students by the college to a relatively unstructured but informed interactive process. In the extreme the classification paradigm is completely mechanistic. The guidance-selection paradigm, however, is fundamentally humanistic. Yet it adopts a quantitative scientific approach to the greatest possible extent consistent with the realization of the aspirations of the largest possible number of individuals and a degree of overall efficiency of selection from society's point of view.

Formal classification models reflect the point of view that if assignments are good, on the average, then a satisfactory state of affairs has been attained. The student, however, is unconcerned with such average good, but is concerned with whether or not his particular assignment is good. If he perceives that he belongs to some subgroup for which, on the average, poor assignment decisions are made, it will not comfort him to know that the system works well for almost everybody else.

It is thus essential, from this point of view, that the overall personnel decision procedure take into account not only the needs of society, which must reflect the needs of the individuals, but also specifically the individual needs of its people. Thus personnel decisions must be both efficient and fair. Cronbach and Gleser (1965) have pointed out,

> ...that an abstract conception of "justice" lies behind much of the concern [in testing] about error of measurement. An ability test is expected to rank persons from best to poorest, and error distorts the ranking. Since such distortion is "unfair" to the individuals who are ranked lower than they deserve, testers want to reduce errors of measurement.

Reflecting the accepted point of view at the time of the first edition of their book (1955) they argued further that

> ...from a utilitarian point of view, these errors can be ignored unless they alter the goodness of whatever decisions are to be made.

The first part of Cronbach and Gleser's statement is clear and must be accepted as an important contribution to our thinking about what constitutes both good tests and good testing procedures. The second part, however, must be interpreted in the light of the social and political developments of the last decade and as a result the last phrase of this remark must bear heavy emphasis. Recent developments have resulted in Cronbach's more recent writing (Cronbach & Snow, 1969) indicating more specifically that in present day American society a more elaborate utility structure must be considered than has been in the past. The utilitarian point of view that Cronbach and Gleser spoke of was that of the testing organization and those selecting students. It is not necessarily that of most examinees. It is now recognized that the student's point of view must be considered more carefully than it was a decade ago.

We recognize that the utility of a procedure will be an increasing function of its overall mean effectiveness. We may also feel, however, that its utility will be lessened if its effectiveness is very low for certain recognizable subgroups. If so, then the concept of fairness becomes a component of utility and cannot be ignored. A procedure that is manifestly and grossly unfair to any subgroup of people will not be a satisfactory procedure even if "on the average" it is very good simply because it is very good for most people. By directly quantifying questions such as these in formal decision theoretic terms it would be possible to handle them within the classification-decision theoretic paradigm. It will be more natural, however, and more useful to treat these problems carefully but in a less structured way within the dual guidance-selection paradigm. This can be done by examining regressions within relevant subpopulations.

It is now generally recognized that the maximization of performance on any one particular criterion is seldom the only consideration relevant to a guidance or selection decision. It may well be that a particular student would have a higher grade-point-average in business school than in law school (or vice versa), but if he strongly prefers law to business (or vice versa) and if he can be assured that he can "make it" through law (or business) school, this may well be the best choice for him and for society. That is to say that his degree of career satisfaction and his overall long-term contribution to society may be greater if he attends law (business) school. In cases such as these the important contribution of a prediction technology will be to assure that his initial choice is a reasonable one.

Similarly, the university is seldom concerned only with getting the very brightest students. Most undergraduate colleges seek some kind of regional and sometimes ethnic and social balance and some diversity of goals in their student body, understanding that such diversity and balance creates a richer university experience for all of their students (e.g., see Whitta, 1968). Often American colleges accept students from underdeveloped areas, both domestic and foreign, not because they necessarily believe that these students will be "better" than others that are turned down but simply because society, at large, has a greater need to train these people. The pertinent question in relation to these people is not how well they will do in any absolute or even relative sense but whether they will profit sufficiently from the program. Operationally this often reduces to the simple question of trying to predict whether or not these students will be able to complete satisfactorily the training program, even at the most minimal level.

This humanistic tradition (Katz, 1966) also takes as a basic precept the notion that an individual will not necessarily be most happy doing the kind of work for which his aptitudes best qualify him. The fact that a high school senior is the best typist in his class and only the second best mathematician should not automatically suggest advanced training at secretarial school. Most people will be well qualified to enter and pursue successfully more than one vocation. The scientific approach to personnel guidance views the task of prediction as one of informing the individual as to the extent of his probable "success" in those training courses and vocations that interest him. The humanistic tradition allows that the choice, whether it be of college or vocation, is left to the individual, to the extent that that choice does not make unacceptable demands on society.

This humanistic tradition also takes as a precept the belief that neither the mechanical efficiency of society nor its gross material output are the sole or even the primary goals that personnel technology should serve. If it is agreed that the function of society is to serve all of its people, then, rightfully, any maximization must be of the benefits to these people rather than to the <u>structure</u> of society. Very rigid manpower policies can guarantee having neatly ordered tables of organization but orderly structure does not guarantee either work efficiency or career satisfaction. Democratic societies are always less orderly than totalitarian ones but somehow this lack of structure has proven both productive and satisfying.

One feature of the problem that becomes apparent immediately is that short term optimization is often at the expense of long term good. A mature college graduate will not necessarily select that graduate program in which he feels he can do best. Rather he will ask himself what kind of training he needs to get, to do the kind of work he wants to do. He may find that this training can be obtained only at a university at which, and in a course in which, it is predicted that he will do at best moderately well. But if he can satisfy himself that there is a sufficiently high probability of successfully completing that program then he may well find it to his advantage to forego the attainment of immediate honors from an easy program in favor of the long-term benefits from the program that is more difficult for him.

More broadly it is now recognized that decisions must always reflect the desires and aspirations of the individual, and the needs of society as they represent the combined aspirations of its people. Much work has been done to develop formal mathematical systems that incorporate both probabilities and utilities. A knowledge of these methods is very useful and no person's education in personnel technology can be complete until he has familiarized himself with a thorough treatment of the application of decision theoretic models such as is contained in Cronbach and Gleser's Psychological Tests and Personnel Decisions (1965).

At present it does not seem to be feasible to handle the quantification of utilities as part of a centralized comparative prediction service. When meaningful and accurate, formal quantification can be very useful. But a strained, inaccurate quantification will be mechanistic and stifling. The treatment of utilities should at present be left to the student and his guidance counselor. What can now be done is to provide a well explicated probabilistic system which the student, guidance counsellor, and admissions officer can use as one concrete basis for their own relatively informal utility analysis. No doubt an expansion in guidance services should be accompanied by increased training in utility analysis for guidance counselors. This position is entirely consistent with that of Cronbach and Gleser who, in the second edition of their book, remark that:

> Work since 1955 has reinforced our judgment that decision theory is more important as a point of view than as a source of formal mathematical techniques for developing and applying tests.

On the other hand, the development of simple guides to precise and meaningful utility analysis for career guidance is certainly a research area that should now be receiving high priority.

The possibility of turning important personnel decisions over to a computer has been a tempting one. For a time this approach exuded an aura of relevance, objectivity and precision, the very qualities that justify educational testing. But contemporary youth demand a greater personal participation in the determination of their future. They now rebel against any vestige of authoritarianism in education, even when it is one more of form than of substance. In this atmosphere the dehumanizing effect of unmoderated computer-made classifications will be enormously costly. There is a limit to the benefit that can be obtained from investment in computer hardware. More efficient computers may, in fact, be needed for work in personnel guidance but the need for more and more thoroughly trained and equipped guidance personnel and for more relevant and acceptable quantitative tools is far greater. Improved computer facilities which students can manage directly in an interactive mode however may prove useful in relieving the counselor of some of the burdens of information storage and retrieval. But above all else the goal must be to maximize the informed participation of the student in the determination of his future.

Stages in the Development of Educational Testing Technology

The earliest successful work in educational testing was of a manifestly *empirical* character. By designing tests having *direct relevance* to the operational task set to him, Binet (see Chauncey & Dobbin, 1963) was successful in discriminating between those French schoolchildren who were and were not able to benefit from the particular school program available to them.

With the resources, technology and personnel available during the latter part of the nineteenth century it was not possible to develop a multiplex of testing procedures each tailor-made for a particular action decision. Partly for this reason Binet's methods did not enjoy wide application in Europe, but with little delay these methods crossed the Atlantic and, particularly with the beginning of World War I, found rich soil in which to grow. The names of Termin, Otis, and Yerkes stand out in this period.

An important theoretical step was taken when Spearman, in England, proposed a single ability factor theory to account for the relationships among test scores, and between them and academic success. According to Spearman each student could be thought of as having a unidimensional ability which accounted, in large measure, for his performance on various tests and on various academic tasks. According to Spearman each of these tests and tasks had its own specific component but the various tests and tasks shared only a single general and dominantly important factor. According to such a theory the purpose of testing was to measure this general factor, called intelligence, and to rank individuals so that the more able students could be identified. While it was acknowledged that specific tests and specific tasks might have specific features these were considered to be relatively unimportant.

Spearman's theory supported the development of the IQ test which, for a time, was the major component of educational testing. Undoubtedly this single factor or single trait approach to testing enjoyed popularity in part because it justified a basically unidimensional approach to educational testing, and such an approach was perhaps all that the resources, technology and personnel of its day could support at the operational level. By lifting testing from the specific task oriented prediction paradigm which available resources and technology could not support to the universal ability oriented concept of measurement, which resources and technology could support, psychologists made possible a dynamic and immensely useful growth in testing.

The unifactor theory did not long hold preeminence. The theoretical simplicity, and practical utility which buttressed it, in time gave way before the onslaught of Thurstone's succession of studies showing that human ability is not unidimensional and hence simple but multidimensional and hence complex. Thurstone demonstrated conclusively that it was useful to isolate many human ability factors and that persons' rankings on these factors could vary substantially. Technologically this meant that psychologists should construct multiscale tests and that in specific applications weighted composite scores should be used based on just those scales relevant to the short and long run implications of the intended decision.

Thurstone's theoretical position triumphed not only because it was theoretically superior to Spearman's, but also because some increase in available resources, and the consequent technological breakthroughs, made it possible for testing practice to partially reflect his ideas. However, in part because of questions of cost, many major educational testing programs as contrasted to industrial testing programs have adopted a compromise between the Spearman and the Thurstone positions. It has been found that a very workable procedure is to measure and report scores on two omnibus dimensions of human ability labelled verbal ability and quantitative ability. These measures have the crucial advantages of simplicity and understandability, the absence of which had limited the use of more complex methods in the field. Until recently these measures were the predominantly important part of the major educational testing programs. In addition to their reporting simplicity and demonstrated relevance to immediate academic decisions these scales have proved popular because they are believed to measure a broad spectrum of abilities relevant not only to immediate prediction in the academic situation but also to the more long term and important questions of future job success. In Cronbach and Gleser's (1965) terminology these tests have a wide bandwidth. These tests will not and should not be abandoned in the very near future, but they must be modified, extended and supported as they now are beginning to be in some testing programs by newer and more immediately relevant tools.

In recent years the technology of testing has been developing rapidly and the sophistication of persons in the field has also been rising, more slowly at first, but now with greater acceleration. Coupled with this has been a dynamic development in the computer systems available both to test publishers and test consumers. As a result of these advances educational testing now stands poised for major developments in programs and related services which may well have great significance for American education.

In discussing the College Board Program, Turnbull (1968) has identified possible future stages in the development of testing programs. The first of these will be of primary concern to us here because the methods surveyed in this paper are directly relevant to it. This next stage, as

Turnbull sees it, is the stage of multiplex external programs, which involves "an extension of the recent trend toward the diversity of testing programs and of tests within programs."

This stage involves a giant step in the "tailoring" of testing to meet the demands for decision making relevant to specific examinees and specific choices. The trend here is away from, or towards a supplementation of, the omnibus testing which has served as a workable compromise between the Spearman and Thurstone approaches, and directly toward a Thurstonian recognition of the multidimensional complexity of human abilities and the multidimensional requirements for effective personnel decision making. By producing some tests of narrower bandwidth it is proposed that their fidelity, i.e., predictive power for specific decisions, may be increased. Turnbull, however, suggested that there is a "missing element," "a way to express the results of both standardized tests and school performance in terms meaningful to post-secondary education, in a language at least as well understandable, . . . as the College Board scale".

The methods described here adopt a long available reporting language that is much more understandable than the College Board scale. These methods provide, for each student and for each college or program in which he is interested, understandable, meaningful and maximally accurate predictions of his potential performance in educational opportunities that are relevant to his goals. A reporting system with this objective has been in operation in the state of Washington since 1960 as part of Horst's Washington Pre-College Testing Program. Several testing organizations have also recently taken important steps in this direction. Some giant evolutionary steps must now be taken in the further development of such programs. These steps should lead to meaningful improvements on any existing system.

For one thing, predictions for college applicants should often be made in two forms. The student should be given a point and/or interval estimate of his future grade point average both for his first year in college and (at that point or later) for the entire four year program. He should also be given both point and interval estimates of the probability of his completing both the first year and (at that point or later) the entire four year program. These certainly are understandable quantities, and of immense immediate relevance to his problem of selecting a college or other further education program. Emphasis on the second kind of prediction is not found in current practice.

Students need much more information about college curricula than they are now getting and some training and guidance in decision making would be useful. A thorough discussion of these latter problems is given by Katz (1963, 1969a, 1969b). It would also be useful if students were informed of the probability that they would be accepted by each of the colleges to which they might wish to apply. All of this requires immense computer storage and computation speed but the needs are not beyond present day capabilities.

Thus, after more than 80 years, and only after major breakthroughs in statistical theory, testing technology and computer resources is it now possible to use on a broad scale the multiplex, direct task oriented system, validated empirically by Binet, rationalized theoretically by Thurstone and advocated for so many years by Horst.

At this point and despite our previous discussion readers may still feel that this approach sacrifices educational meaningfulness to attain statistical efficiency by focusing attention primarily on narrow criteria such as grade point average. It is important that an answer be given immediately to that thoughtful query. That answer is based on understanding the nature of the decision problem for which educational tests and the statistical prediction methods accompanying them are used.

Historically, selection methods at the university level have focused on the prediction of first year grade point average. This particular criterion has availability as a major virtue, though there have been studies showing some relationship between first year performance and subsequent academic performance and indeed later career performance. These latter correlations, however, are by no means impressive. For a brief summary of these results see Holland and Richards (1965).

The prediction of grade point average is undoubtedly tied up with a proper emphasis in universities on academic excellence, the desire of individual teachers to instruct good students and the desire of institutions to produce scholars. In part this is a carry-over from an earlier age when the pursuit of learning was considered to be its own reward. Such an attitude remains reflected in university policy because to discard it completely would destroy our universities as we now know them and particularly their essential roles in basic research and human enlightenment. However, our leading universities have shown that it is possible to maintain academic excellence and extensive programs of basic research and yet at the same time serve the larger needs of society. Thus, for example, not every graduate student doing work in mathematics is now pointed toward a career in basic research and teaching in mathematics. Rather, a large percentage of students taking such courses are doing so only to pick up needed skills for technological application. This has always been true, but only in recent years have educators been willing to speak directly in these terms.

It would be an unwarranted digression to explore here all of the complicated ramifications of that development. But the demands that this thinking places on score reporting-prediction procedures are relevant. One clear requirement is the reporting of an estimate of the probability of success in the particular university or particular course of study for each student. A student may well not wish to attend the university at which he would do best, and a university may not necessarily wish to take only those students who will do best. Rather each may seek a matching of student to program so as to offer the prospect of the student making a significant contribution to society, provided the student has a reasonable probability of completing the course. Thus the guidance and acceptance of a student will depend upon the relevance of a particular program to the abilities and goals of the individual student and those of the individual institution. The Bayesian methods discussed in this paper are oriented towards this approach to score reporting and prediction. The guidance and selection models currently available must be extended to provide interval estimates both of grade point average and of successful completion of the program of study.

There are, of course, more standard methods of prediction and the reader undoubtedly now wonders why it is necessary to have a new statistical methodology. The problem arises, in part, because of the present dynamic nature of American education. Previously curricula within colleges remained relatively unchanged for many years and colleges themselves changed their natures even more slowly. Therefore data could be collected over a period of several years with the assurance that regression equations determined from the data would be applicable and useful for another several years. Thus the size of the sample available for any study was limited mainly by administrative difficulties in gathering data.

American colleges no longer always exhibit such stability. Programs within colleges can change dramatically in just a year or two and thus historical data may now have only descriptive value. If, for example, a graduate psychology department were to substantially increase the mathematical content of its curriculum the use of regression equations from previous years would be very unsatisfactory. Since criterion data are available only after several years acceptances have already been made there is a clear demand for some way of writing prediction equations that takes account of whatever small amount of relevant data there is with respect to that college and also the experience that other similar colleges have had in such a situation. If a particular college can identify its new program as being similar to that of certain other colleges it would undoubtedly want to draw on the experience of these colleges. This would also be true of a college undertaking prediction problems for the first time.

The Bayesian methods discussed here have a virtue peculiar to them. These methods make it possible to increase the accuracy of predictions for the individual not only by gathering additional data about him and the college to which he is applying, but also by gathering additional data about the group of which he is a member and about colleges similar to the ones to which he is applying. It is a significant virtue of the Bayesian method that our knowledge concerning groups of students and of colleges gives us, probabilistically, information that can be translated into more accurate predictions for each individual. The statistical basis for this will be discussed in the next section.

Thus the Bayesian regression approach to be presented here, with its increased sensitivity and potentially with dual predictive modes, seems to be the natural approach to the guidance-selection problem and, as we have seen, the mode of score reporting is both easily understood and maximally informative. Some details of this proposed reporting system are given in later sections.

The step now being taken in the evolution of educational testing technology is one that is firmly grounded in a succession of historical developments. This step is in no sense revolutionary and while there have been many important contributions it is not the child of any single person. Nevertheless if this next step is taken specifically in the direction suggested in this paper, it will be a giant step. It will involve the embracing of statistical methodology that is only slowly losing its controversial status. It will also mean that though some of its techniques will remain important and useful the entire measurement tradition will lose its primacy as a basis for developing operational testing procedures. But here again we do no more than echo the prescription contained in Cronbach and Gleser (1965) to abandon the view expressed by Hull (1928) that "the ultimate purpose of using aptitude tests is to estimate or forecast aptitudes from test scores." Surely it must be recognized that relevance in testing cannot be inferred from the estimation of true score.

Bayesian Methods in Educational Testing

A review of Bayesian methods has recently been given by Meyer (1966). We shall now describe Bayesian analyses for two important new models. This presentation is meant to provide a technical basis for an improved guidance-selection system. A parallel verbal presentation of this material is given at the beginning of the section on An overview of new developments in testing services. Many readers may prefer to see this verbal summary before examining the explicit quantitative statement of the models.

The first Bayesian analysis we shall describe is that of the classical test theory model and the second is a regression model for several subgroups. The first of these will be familiar to measurement specialists who should see from this discussion the intimate relationship between test theory and Bayesian inference. The second model provides results which are similar to those of the first model and which are directly relevant to comparative guidance services. Finally we survey Bayesian methods in the analysis of variance components.

Within the classical test theory model each person's observed score x on a test may be used as an estimate of his true score τ. If this is done, the standard deviation of the errors over persons for such a procedure (the standard error of measurement) will be $\sigma_X(1 - \rho_{XX'})^{\frac{1}{2}}$, where σ_X is the observed score standard deviation and $\rho_{XX'}$ is the reliability of the test. This provides a measure of the inaccuracy, on the average, of this particular method of estimating true score. An alternative method of estimation is to use the weighted average regression estimate $x\,\rho_{XX'} + \mu_X(1 - \rho_{XX'})$ where μ_X is the mean of the observed scores in the population of persons. If this is done, the population standard deviation, over persons, of the resulting errors (the standard error of estimation) is $\sigma_X\rho_{XX'}^{\frac{1}{2}}(1 - \rho_{XX'})^{\frac{1}{2}}$.

As can be seen on comparing formulas, the standard error of estimation is always less than the standard error of measurement, and substantially so when the reliability of the test is not large. Thus by incorporating supposedly known values of the reliability and the mean observed score into the estimate of true score by means of the weighted average regression estimate a better estimate is obtained than that based solely on the observed score.

Over and above its mathematical derivation the regression formula (Kelley, 1927) makes intuitive sense. If we have little or no information about a person, and we can assume that he has in effect been randomly selected from the population at hand, it seems reasonable to use the mean ability level in the population as our estimate for that person. When the length of a test is very short the reliability of the test will be near zero and the regression estimate will be very nearly equal to that population mean score. When the test is very long the reliability of the test will be near unity and the regression estimate for each person will be very nearly equal to that person's mean observed score. Thus, as we would expect, when very reliable information is available about a person's true score we would need to put little weight on the mean population value.

Unfortunately, there is a difficulty in attempting to apply the Kelley formulation in most practical applications. The problem is that the population mean is typically not known before measurements are taken and hence the regression formula cannot be used in its given form. In effect what is needed is a regression estimate based not on the person's observed score and a <u>known</u> mean observed score, but rather one based on the person's observed score and the average observed score of a random sample of people from the population. Results of this type are available in the framework of Bayesian methods with normality assumptions. The first of these was given by Box and Tiao (1968) and a later one was given by Lindley (see Novick, 1969a). These estimates are of the form, $w_{XX'}x + (1 - w_{XX'})\bar{x}$, i.e., a weighted average depending on weights $w_{XX'}$, the person's observed score x and the mean observed score $\bar{x}$ in the sample. In this formulation the quantity $w_{XX'}$ is an estimate of the reliability of the observed score. It tends to unity as the number of observations on the person increases without limit and to zero as the number of observations on the person tends to zero. For intermediate cases its value depends on the relative number of observations on the particular person, the number of observations on all persons and also on the number of persons on whom observations are available.

For our purposes it will be useful to consider the Bayesian estimates obtained by Lindley since this method easily generalizes to the case of unequal replications. Under moderate conditions and using the specific prior distribution suggested by Novick (1969a) to characterize a situation in which we have no prior information, Lindley shows that the mode (or most probable value) of the conditional distribution (the posterior Bayes distribution) of the true scores τ_s after obtaining all observed scores can be calculated as the solution of the m equations

$$mn \frac{\tau_s - x_{s.}}{\Sigma s_i^2 + \Sigma(x_{i.} - \tau_i)^2} + \frac{(m - 1)(\tau_s - \tau.)}{\Sigma(\tau_i - \tau.)^2} = 0 \qquad [1]$$

where x_{ij} is the j -th observation on the i -th person, m is the number of persons, n is the number of replications on each person, $s_i^2 = \sum_j (x_{ij} - x_{i.})^2/n$, $x_{i.} = \sum_j x_{ij}/n$ and $\tau. = \sum_i \tau_i/m$ and where it is assumed that τ_i are not all equal. These are the simplest of the <u>Lindley equations</u>. Because the quantity τ_s is a part of the mean value $\tau.$, these equations cannot be solved directly. An approximate solution to these equations for large m and n having the general form described above is

$$\hat{\tau}_s = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \frac{1}{n}\hat{\sigma}_E^2} x_{s\cdot} + \frac{\frac{1}{n}\hat{\sigma}_E^2}{\hat{\sigma}_T^2 + \frac{1}{n}\hat{\sigma}_E^2} x_{\cdot\cdot} \ , \quad s = 1,2,\ldots,m \ , \qquad [2]$$

where $\hat{\sigma}$'s are the usual ANOVA estimates and $x_{\cdot\cdot} = m^{-1} \sum_i x_{i\cdot}$. For small samples equations [1] and [2] do not give the same results. Further details of a method for obtaining the exact solutions to the Lindley equations by iteration, the modest conditions under which they are valid and reasons for preferring the Lindley method are given by Novick and Thayer (1969b). A generalization of these equations to include the case of unequal replication numbers and including a technical improvement to guarantee convergence has recently been provided by Lindley (1969b).

The true score estimates given in [1] were obtained from a Bayesian structural model which assumes that the observed scores for each individual are normally distributed with mean equal to that person's true score and with homogeneous error variance σ_E^2 , and that the true scores are normally distributed with mean μ_T and variance σ_T^2 . It was further assumed that there was no information available about the true or error score variances or the mean true score. Formally this was accomplished by using the indifference prior distributions for μ_T , σ_T^2 and σ_E^2 suggested by Novick (1969a) as developed from the work of Novick and Hall (1965). These indifference priors consist of independent uniform distributions on μ_τ , log σ_E , and log σ_τ . However, if some prior information is available either about the distribution of true or error scores, this information can be incorporated into the prior distribution using the procedure suggested by Novick (1969a) as developed from the work of Novick and Grizzle (1965). Often it is useful and sometimes it may be essential to do this. However it seems to be true that when the number of persons being tested is large, prior information can be largely disregarded (Novick & Thayer, 1969a).

The choice of the prior distribution for this analysis reflects prior information and beliefs (or lack of the same) concerning the mean true score in the population, the spread of true score values and the average variability, within person. These, in total, imply a prior distribution on the individual true scores τ_i . After obtaining observations on persons we have a new Bayes distribution for the τ_i and we also have a new Bayes distribution for the mean true score, the variance of the true scores, and the variance of the error scores and all of this information is available to guide any decision that must be made at any stage of testing. Lindley's methods and the very similar ones of Box and Tiao provide improved techniques for estimating true and error score variances and reliability. The details are given in a paper by Novick, Jackson and Thayer (1969).

The point to be emphasized here is that at any point in the data gathering the Bayes distribution for any particular τ_i reflects more than just the observations on person i. Rather it reflects the combined information relevant to all of the τ_i . Thus after we have information on some τ_i we are no longer completely uninformed about a new τ_s , rather our prior distribution for this new τ_s would effectively be our estimated distribution of τ values in the population of people. As has been seen the effect of this is to regress estimates of true score towards a common mean. This regression provides the Bayesian solution to a number of statistical problems. Thus for this rather complex

Bayesian structural model the actual use of a vague prior for data analysis seems appropriate when the number of persons is large, while for less complex models objections can be raised (e.g., Novick & Grizzle, 1965). This is so because the buildup of information is much more rapid with the structural model than with simpler models.

This same kind of argument has been applied by Lindley (1969a) to the estimation of regression coefficients. Suppose that a number of similar graduate departments of psychology wish to use the GRE advanced psychology examination to predict a student's performance on a final written examination and hence to supply one useful piece of information for their selection process. Then a student with a score of x has an expected score y in the i -th department given by the linear model

$$\mathcal{E}(y_i|x) = \alpha_i + \beta_i x$$

where the parameters α_i and β_i depend on the particular psychology department and typically vary among departments. Thus we have possibly different linear regressions in each department. We assume the distribution of y_i given x_i to be normal with mean as given above and, for present expository purposes, with known residual variance σ_i^2.

Whatever experimental data are available and deemed relevant can be expressed in the form

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots n_i$.

Here we have supposed that there are m departments, that data are available for n_i previous students in each of these departments, that

x_{ij} is the test score of the j-th student in the i--th department, that y_{ij} is similarly his final written examination score and that the residuals ϵ_{ij} are independently normally distributed with mean zero and known variance σ_i^2 .

The usual statistical analysis of the data would proceed in the following way. Each department would be considered separately and the regression line for that department estimated by the usual least squares procedure giving estimates

$$\hat{\beta}_i = S(x_i,y_i)/S^2(x_i) \qquad [3]$$

and

$$\hat{\alpha}_i = y_{i.} - \hat{\beta}_i x_{i.} \quad , \qquad [4]$$

where

$$S(x_i,y_i) = \sum_{j=1}^{n_i} (y_{ij} - y_{i.})(x_{ij} - x_{i.}) \quad ,$$

$$S^2(x_i) = \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2 \quad ,$$

$$x_{i.} = \sum_{j=1}^{n_i} x_{ij}/n_i$$

and

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i \quad ,$$

these being the usual sum of products, sum of squares and means for the i -th department.

Lindley points out that this standard procedure is open to the following criticism. "In estimating the regression for any [department] it fails to take into account experiences gained with similar [departments]. For example, suppose one found that the regression slopes, β_i , were typically around one, then one would expect the slope for [another department] to have about the same value and would be astonished if it differed sharply from it. This is perhaps most clearly seen by considering what one might intuitively do if no data were available for one [department] beyond an x -score for a single student; one would reasonably estimate his y -score as $\alpha + \beta x$ where α and β were some sorts of means of the values obtained for similar schools for which there were data. This criticism (of the classical method) can be overcome, and the suggested procedure just described made precise, by using a Bayesian argument in place of the standard one."

The dissatisfaction with the orthodox approach springs from the fact that one knows a priori that the slopes β_i have similar values and that (departments) are not terribly heterogeneous; similar remarks apply to the ordinates α_i . We therefore suppose that this prior knowledge is made precise by assuming that the individual β_i are independent and identically distributed with a normal distribution of unknown mean β and known variance; that the α_i come from a similar distribution with known variance but unknown mean α, and that these are independent. Furthermore the knowledge of α and β is supposed vague.

With these additional assumptions the full model is that

$$E(y_{ij}) = \alpha_i + \beta_i x_{ij} \qquad [5]$$

with variances σ_i^2, all the y 's being independent normal; that

$$E(\alpha_i) = \alpha \ , \quad E(\beta_i) = \beta \qquad [6]$$

with known variances, all α_i and β_i being independent normal; and that the prior knowledge of α and β is vague. Lindley then shows how for this model the posterior distributions of α_i 's and β_i 's may be found. In particular he shows how to obtain the modes of the posterior distributions given the data, these modes providing modified estimates in equations [3] and [4]. The expression for the Bayesian estimate of β_i has the form

$$\hat{\hat{\beta}}_i = \frac{S(x_i,y_i) + c_i}{S^2(x_i) + d_i} \qquad [7]$$

where the precise nature of c_i and d_i need not concern us here. "The terms c_i and d_i represent corrections to be applied to the sums of products and squares respectively in the light of the "prior" information we have about the parameters. Without c_i and d_i, the right hand side of [7] is the usual estimate $\hat{\beta}_i$, equation [3]. Furthermore c_i and d_i depend not just on the pooled data for the i -th department but also on pooled data from the other departments. In particular they depend on all the other estimates $\hat{\hat{\beta}}_j$, $j \neq i$, and have the effect of regressing the estimates from $\hat{\beta}_i$ to $\hat{\hat{\beta}}_i$, where $\hat{\hat{\beta}}_i$ is nearer to the average slope than is $\hat{\beta}_i$. Hence the extreme slopes are modified by being moved towards the central values. The formula is even valid for a department about which there are no data; the estimated slope takes the form c_i/d_i, though, in this case the ratio c_i/d_i does not depend on i. This form essentially says that we can estimate the slope for this department by regarding it as typical of the other departments for which data are available. Similar methods and results apply to the estimation of the intercepts α_i ." Further work needs to be done to obtain a suitable measure of dispersion so that interval estimates can be made. When the variances σ_i^2 are unknown the exact Bayesian structural estimates of the α_i and β_i must be obtained as a solution to a set of Lindley equations (Lindley, 1969c), similar to, but more complex than, those given in [1]. Further work must be done to extend these techniques for multiple regression.

The analysis of variance components is a frequently occurring statistical problem whose solution leads to unexpected complications. A familiar example in educational testing is the estimation of the true score variance σ_T^2 and error score variance σ_E^2 in the classical test theory model, the data being n parallel measurements on a sample of m persons.

Posterior distributions of σ_T^2 , σ_E^2 and the reliability coefficient could, of course, be obtained as a byproduct of the analyses described at the beginning of the section, and an advantage of the Bayesian methodology is that there can be no inconsistency between the conclusions reached about the various parameters of interest, as there might be if each were estimated separately by some classical method which appeared "good" for it alone. However, considerable light has been shed on the variance components problem by a number of Bayesian investigations in recent years.

The information about σ_T^2 and σ_E^2 provided by the data is summarized by the pair of sufficient statistics

$$S_W = \Sigma\Sigma(x_{ij} - x_{i.})^2 , \qquad S_B = \Sigma n(x_{i.} - x_{..})^2 \quad ,$$

where x_{ij} is the j -th parallel measurement on the i -th person,

$x_{i.} = \sum_j x_{ij}/n$ is the i -th person's average score,

$x_{..} = \Sigma x_{i.}/m$ is the overall average score.

S_W and S_B are commonly referred to as the within-persons and between-persons sums of squares, and have associated with them $m(n - 1)$ and $(m - 1)$ degrees of freedom respectively. Dividing the sums of squares by their degrees of freedom we obtain the mean squares MS_W and MS_B with expectations

$$EMS_W = \sigma_E^2$$

$$EMS_B = \sigma_E^2 + n\sigma_T^2 \quad .$$

Usual classical practice is to take MS_W as an unbiased estimate $\hat{\sigma}_E^2$ of σ_E^2 and $n^{-1}(MS_B - MS_W)$ as an unbiased estimate $\hat{\sigma}_\tau^2$ of σ_τ^2 . Clearly $\hat{\sigma}_\tau^2$ can be negative, which is felt to be somewhat absurd, and many modifications of classical methods have been proposed to deal with this situation, none of them entirely satisfactory.

The Bayesian method always leads to a nonnegative estimate of σ_T^2 . Also a number of writers, using Bayesian methods, have brought into clearer focus the implications of a <u>classical</u> estimate $\hat{\sigma}_T^2$ substantially less than zero. Their work graphically illustrates the fact that such a result casts grave doubts on the assumptions of the model, particularly the assumptions of parallelism and experimental independence of the replicate measurements. This work also highlights the weakness of the classical estimate of error variance in that it fails to use the information in the between sum of squares. A survey of the technical details of this work has been given by Novick, Jackson and Thayer (1969).

Multiple Comparisons and the Choice of Predictor Variables

Two problems that have been of intense and continuing interest to data analysts in education, psychology and other behavioral sciences and which are important in the development of a Bayesian guidance technology are those of multiple comparisons and the choice of predictor variables. For each of these the Bayesian position seems so sound, even compelling, when viewed in the context of the models discussed earlier that some comments on these topics seem appropriate here. A more detailed treatment of these problems is given by Novick (1969b) on the basis of work by Lindley (1965).

The multiple comparisons problem arises when we attempt, say, to simultaneously make individual comparisons of the differences among a large number of means. If classical statistical methods based on a sequence of n , α level "t" tests are used, one must "expect" approximately $n\alpha$ erroneous decisions of difference when no differences exist. Similar objections arise when less trivial formal procedures are used. Of course, there are "experiment-wise", as opposed to "pair-wise", methods of treating this problem classically, and these are preferred by most statisticians, but this approach often appears to be unduly conservative. Advocates of such approaches, on the other hand, argue that Bayesian methods are insufficiently conservative.

It is a fact that Bayesian methods can b insufficiently conservative, but they need not be. As indicated by Novick and Grizzle (1965) when uniform prior distributions are used a priori on mean parameters then the Bayesian method yields results very similar to the classical pair-wise approach, and this is unacceptable. Novick and Grizzle exhibit a sound and effective but somewhat crude method of constructing more acceptable prior distributions and demonstrate that when this is done more satisfactory results follow.

If we view each "treatment" (i.e., mean) effect as an effect randomly selected from a population of such effects, then any decision we wish to make concerning any treatment pairs, based on observed between-treatment differences, should be tempered by our a priori knowledge of, or information or beliefs concerning, the mean value and spread of treatment effects in that hypothetical population. Bayesian methods that use independent uniform prior distributions on the individual treatment means imply that the spread of treatment effects in the population is arbitrarily large. This can never be deemed reasonable, and can be deemed acceptable only when very large fixed sample sizes are used. Such priors prejudice the posteriors toward large differences unless the sample is sufficiently large so that the likelihood swamps the prior Bayes density.

However, when the structural model is used the Bayesian method will regress observed differences to the overall mean difference among treatments and this regression will diminish with increasing sample size. If we feel, in any application, that we should adopt very conservative procedures, it is probably in part because we do not "expect" to find many differences among means except by chance, i.e., we believe the mean difference among treatments to be zero and the variance among treatment differences to be small.

To give our procedure the desirable amount of conservatism we need only quantify those prior beliefs that justify this desire. Should we be inaccurate in our assessment of the parameters of the distribution of treatment means we will at least know that on the average our posterior beliefs concerning these parameters will be more accurate than our prior assessments, as will our decisions based on them. Actually for reasons already given, the Bayesian structural model even without presumed prior information avoids the problems encountered with uniform prior distributions on individual parameters.

The problem of the choice of predictor variables is a variant on the theme of multiple comparisons, attention being shifted from a consideration of treatments having nonzero difference from the mean of the treatment effects, to those variables having nonzero partial correlation with some criterion. It is well known that in a classical approach, when sample sizes are small relative to the number of predictors, it is often best in a predictive efficiency sense not to use all variables in the multiple regression, but rather to use some lesser number. In the Bayesian approach, unless one or more of the partial correlations is zero, it is always better to use all variables, when the evaluation of the efficiency of the procedure is made with reference to the Bayes criterion and with the assumed prior distribution. These apparently contradictory conclusions can be reconciled. Since a Bayesian analysis with uniform prior distributions used for all regression parameters is essentially equivalent to a classical analysis, it should not surprise us that the use of such unreasonable prior Bayes distributions lead to unreasonable frequency results. However, suppose that we assume that the regression coefficients have been sampled from a population of regression coefficients, and suppose that our prior distribution on the mean of these coefficients is centered at zero and the prior distribution on the variance does not place undue weight on infinite values. What we will then find is that our posterior estimates of the individual regression coefficients will themselves be regressed to the mean of the regression coefficients so that for small samples the Bayesian will have quite different estimates than will the classicist. Again, use of the Bayesian structural model will accomplish this end. For small sample sizes some of these regression coefficient estimates may in fact regress very nearly to zero. Again, as in the previous estimation problem, both Bayesians and non-Bayesians use prior information, but only the Bayesian explicitly quantifies this aspect of his work and only the Bayesian method permits the data to modify these prior beliefs. Perhaps this "explains" why the Bayesian uses all variables when a prior distribution is available.

An Overview of New Developments in Testing Services

The Bayesian regression model developed by Lindley is based on the simple notion that the ability of persons and the grading and performance standards of educational institutions can be more efficiently estimated by taking into account our knowledge that particular values of these parameters will be highly related within homogeneous groups. The first application was to the estimation of true scores. Here a Bayesian justification was found for Kelley's classical weighted average regression estimate of true score based on observed score and mean observed score in the population. The Kelley estimate in its Bayesian extension was found to be a weighted average of the observed score for the person and the mean observed score in the sample of persons tested with the weights being, respectively, the sample reliability of the test and one minus that reliability. In effect this estimation procedure is based on the notion that when the observed score is relatively unreliable an improvement in estimation can be obtained by using information about the mean value in the sample of examinees. The Bayesian estimate regresses the individual observed scores back toward the mean of all of the observed scores and this regression is large when the unreliability of the test is large.

This same idea was then used to provide a model for new guidance and selection methods for situations in which an external criterion is available. The standard situation is that we have information, in the form of test scores and/or high school grades, about students who come from different schools and who express an interest in different colleges. Just as students exhibit different true ability levels on a particular test or school performance records, so schools and colleges have different grading standards and hence differing difficulty levels which must be taken into account if accurate and unbiased prediction is to be accomplished. When we wish to estimate the parameter values relevant to schools or colleges it will be useful to use expert judgment to group schools and colleges homogeneously and use a Bayesian regression type estimate for each school and each college parameter with each parameter value being regressed back towards the average value c' such parameters for all schools or colleges in that group. (Such groupings can, of course, be modified on the basis of subsequent information.) One application was discussed briefly to illustrate the use of the formal model. We now describe other applications of this same model. Our purpose is to indicate how the use of this model can and should (and should not) substantially affect testing practice in the very near future. In effect we shall be providing a brief prospectus of academic prediction systems.

The first modification of test score reporting that might come to mind would involve the reporting of Bayesian regression estimates of true score instead of the actual observed scores. We might be tempted, for example, to identify each examinee as a student in a particular high school and then regress his observed test score toward the mean of the observed scores obtained by persons from that school. This procedure, however, can be faulted on several counts.

In the first place, the reliabilities of most academic aptitude tests are very high in populations in which there is a broad range of ability levels and even in restricted subpopulations they do not typically differ substantially from one group to another. When reliabilities are high the regression effect is so small that there is little point to regressing the estimates. Moreover, even if the reliabilities were not large the regression estimates would not substantially change the ordering of the examinees unless there is a substantial difference in the test reliabilities in the different groups or the number of replications across persons varied. Since, again, these differences tend in practice to be small there is little point in making these corrections.

Furthermore if the reliabilities were not large and if the mean value in the groups were more than trivially different, serious objections concerning the fairness of this procedure would need to be raised. As Robbins (1960), in effect, points out, it is unfair to penalize a student by lowering our estimate of his ability because we know that he can be identified with a low ability group. Thus there is good reason to question the fairness of rejecting this student and accepting another student who did less well on the test simply because we identify the second student as coming from a high ability group. There is both common sense and theoretical justification for thinking that a student with a SAT Verbal score of 600 coming from a very poor school is, in fact, a better choice for many colleges, than a second student with a score of 610 coming from a very good school. Thus the reporting of Bayesian regression estimates of true score both lacks significant virtue and has possibly serious defects. A resolution of the fairness question can be obtained only when the relevance of the test score to the pending action decision is taken into account. (See Cleary, 1968, for an intelligent discussion of the problem of bias in testing.)

The application of most immediate interest to testing organizations will undoubtedly be that of using test scores to provide comparative prediction of success at various colleges and within various curricula within a college. In this application the regression of grade point average on test score for each college and each college program is estimated and this is done without taking into account the high school affiliation of each student.

The purpose of such an exercise is rationally to arrive at the kind of judgments now generally being made on the basis of rather poorly gathered, poorly organized and poorly transmitted information as to the overall difficulty level of various colleges or college programs and the particular traits necessary for success at these colleges and in these programs. The exercise, however, is not a simple one and great care must be exercised to define the problem precisely. Successful work along these lines has been done by Horst and his associates at the University of Washington for many years. The new Bayesian methodology, however, promises a substantial increase in the accuracy with which such predictions can be made.

One precisely stated problem would be the prediction of first year grade point averages at various colleges on the basis of academic aptitude test scores. By limiting the problem to the prediction of first year grades, difficulties arising because of differences in standards among departments within a college are minimized. At many colleges all freshmen take very much the same program so that there will be no interdepartmental differences. However, if an attempt is made to predict four year grade point average, this may generally need to be done within departments. At the graduate school level it would undoubtedly be necessary to work field by field rather than across fields within school. The Bayesian method would be especially useful in this application because of the relative smallness of individual programs.

Guidance and selection problems for professional schools are particularly suited for treatment by centralized prediction methods. The relative smallness of the programs and the greater community of interest among the participants would make these programs ideal field laboratories during the developmental stage of a Bayesian guidance-selection project. Ultimately comparative prediction should be a continuing process beginning in the earliest years of education in the assessment of reading readiness and continuing throughout a person's active work years.

Consider another rather simple guidance-selection problem. We have a single college selecting students from a fixed group of high schools. Suppose that studies have not been done relating high school performance to college performance but scores on a battery of tests are available on all students. A standard approach would be to relate such test scores to college performance and to make tentative selections on the basis of the particular combination of test scores which appear to best predict college performance. This classic criterion-oriented approach to combining a number of test scores will generally prove superior to any single scale unidimensional latent trait approach which attempts simply to order students on the basis of their "intelligence", presumably so that the more "intelligent" can be selected without regard to the peculiar character of the individual college. The Bayesian-regression approach holds out the promise of even further significant improvement.

In such situations it will often be the case that the schools differ substantially in the mean test scores of their students and typically this will be concomitant with the general level of instruction and grading within the schools. Schools that get good students can teach more than schools that do not and in turn can put out better trained students who will score more highly on tests than students from other schools. However, present level of training, in itself, is not necessarily an adequate predictor either of performance at the next level of training or in career potential.

Again if we believe the possibility that a 600 student from a poor school may be a better potential selection than a 610 student from a good school we need some formal mechanism for evaluating this hypothesis. A Bayesian differential predictability regression model provides the needed tool. Using the model described in the previous section the slopes and intercept of the regression of college performance on test scores are computed for each school as an estimate that is regressed from the usual least squares value back toward the average value among all schools. This accomplishes two things. First it improves the overall accuracy of the estimation procedure, often even providing reasonably good estimates when only small amounts of data are available on some schools. Second, it permits differences in regression slopes and intercepts to emerge in a continuous fashion from the data as the amount of data increases. In this application such differences will not often exist, but when they do they will be very important. Recall that when no information is available about a particular school the Bayesian regression estimate is the average value among schools and when "infinite" information is available the Bayesian estimate is the usual least squares estimate.

Now it may happen that clear differences among schools in slopes and intercepts emerge from a particular data set. For example we might well find that the slope of the regression line for one "good" school is less than the slope for a "poor" school. If this were the case, and if the crossing of these regression lines occurred in the region of obtainable scores then for high scores the predicted value for the student from the poorer school would be higher than that for the student from the better school. The opposite could, of course, be true. In any event this is something about which accurate information rather than speculation must be made available. If it were found that prediction was substantially poorer in some groups than in the overall population, there would be good reason both to examine the aptness of the training program and the criterion for this subgroup and to seek more effective predictors for this subgroup. More generally, differential predictability methods can be used whenever students can be grouped in a meaningful way so that different regression lines are relevant in different groupings.

Thus we have another simple situation in which the Bayesian regression model might make a substantial contribution to predictive efficiency and in this application it is difficult to anticipate any charge of unfairness against the method. The method is fair because it accurately performs its function of predicting a person's ability to succeed in a proposed program of study. It would be unfair to mislead a person by knowingly furnishing him with over-predictions of his probability of success in a given program. Under certain circumstances, however, it might be argued that the accepting systems should only be furnished with predictions based on data from the entire population. Unfairness in its most objectionable form arises when the range of available programs is restricted so as to exclude from training some who can profit from further formal education and when selection for available programs is based on measurements that are not relevant to the prediction of success in that program or in the career opportunities to which this program leads. Just as tests must be tailored for different kinds of decisions so educational opportunities must be tailored to different kinas of people (Cronbach & Snow, 1969). One important outcome of differential predictability studies should be an increase in the variety of programs available to students.

If the Bayesian comparative prediction and differential predictability models both prove useful with test scores as predictors there is the possibility of combining the two systems. This involves the use of a more complicated mathematical model than the one described above for use in comparative prediction or differential predictability alone.

Another potential application of the Bayesian structural model is to the adjustment of high school grades to provide optimum prediction of college grades for various curricula and the adjustment of college grades to reflect intercollege differences in grading standards. The first major study involving such central prediction methods was that of Bloom and Peters (1961). This work has led to the belief that substantial increases in correlation can be effected by adjusting high school and college grades. Some later studies conducted by testing organizations (e.g., Lindquist, 1963; Watkins & Levine, 1969) have failed to support the early promise of such methods. A review by Linn (1966) does not provide a favorable appraisal of such methods. However a very recent study by Cory (1968) does support one non-Bayesian method.

Clearly a necessary condition for any adjustment technique to be of value is that there exist substantial school and/or college effects and possibly a large interaction effect (i.e., differing slopes within school college pairing). If both schools and colleges exhibit negligible between unit variation it can hardly be surprising when no benefit is obtained from adjustment methods. In such situations research suggests that the unadjusted high school grade average is the best single predictor of college grades, even better typically than academic aptitude test scores. When differences do exist among high schools the use of test scores rather than grades may largely do away with the need for student source adjustments. Indeed this simplification has been a major reason for the existence of a testing industry.

Moreover, even when substantial differences in high school and college grading standards and differences in within pair regression slopes do exist, classical methods can prove to be less than useful when many parameters must be estimated with little data. In the case of individual regression slopes such paucity of data will typically exist. It may be possible to gather much data on P. S. #1 in New York and on the Black Hills Teachers College in Spearfish, South Dakota, but we are unlikely to have much data on students from that specific high school attending that specific college.

This causes no problem when the Bayesian Structural model is used because information on similar school-college combinations can be used to provide statistically optimum prediction weights even when no information is available on a particular school-college combination. Until such time as the Bayesian structural model has been used in situations in which school and college differences do exist, it will be premature to presume final judgment on such methods, particularly since they may be specifically the tools needed for what Turnbull (1968) calls the <u>school-based system</u>. In some applications it may prove useful to use both test scores and high school grades with differentiation being made both as to high school source and to prospective college choice. Work of French (1963) and of Lunneborg and Lunneborg (1966, 1967) suggests that the use of course grades and academic aptitude tests as predictors of college performance can with profit be supplemented by the use of other cognitive and noncognitive measures.

In summary, then, the Bayesian methodology can be used to provide direct predictions of success using test scores and/or previous grades for one or more possible training programs and treating the applicant group as a whole or dividing it into relevant subgroups when appropriate.

A Score Reporting and Guidance Service

There are a number of features of the proposed score reporting and guidance service that should make it attractive to both students and school administrators. To pinpoint these features let us consider the report being furnished a student who took the Academic Aptitude Test on December 15, 1974.

Score Report

Academic Aptitude Test

AAT Verbal Scaled Score	500	AAT Quantitative Scaled Score	600
Percentile Rank	50%	Percentile Rank	89%

Guidance Information

		If Accepted			
University	Probability of Acceptance	Probability of Completing First Year	First Year Predicted Grade Point Average	Probability of Attaining Degree in Normal Period	Overall Predicted Grade Point Average
Ivy League University	.10-.20	.85-.95	2.4-2.7	.60-.80	2.0-2.8
Underdeveloped Area Technical College	.95-1.00	.90-.95	3.4-3.8	.75-.95	3.2-3.8
North Atlantic State University	.00-.05	.95-1.00	3.5-3.9	.80-.95	3.3-3.9
Rocky Mountain State University	.85-.95	.80-.90	3.0-3.5	.65-.85	2.8-3.5
Community Junior College	1.00-1.00	.85-.90	2.9-3.3	.65-.75	2.6-3.3

The low probability of acceptance for this student at Ivy League University reflects primarily the low selection ratio at that university. Note, however, that if accepted, his probability of successfully completing the first year is higher than at Rocky Mountain State University or at Community Junior College. This is not an unusual finding. Many highly selective universities are very protective of those that are accepted while many universities with open door policies leave entering students to fend for themselves with the result that many fail for academic or other reasons.

The negligible probability of the student's acceptance at North Atlantic State University and his certain acceptance at Community Junior College reflect only the fact that North Atlantic State University accepts almost no out-of-state students and that Community Junior College is required to accept all residents with high school diplomas. Differences in predicted grade point averages reflect largely differences in curricula at the various universities and varying degrees of emphasis on verbal and quantitative skills together with the more obvious differences in the overall difficulty levels at the various universities. Differences in lengths of prediction intervals reflect largely the amount of prior information on each of these universities. Note that the prediction intervals for the four year GPA are longer than for the first year GPA. In order to obtain more accurate four year predictions it may be necessary to work within departments or major departmental groupings (arts, sciences, business administration, etc.) since the requirements may differ among such groupings more than they do among universities. This could certainly be done more accurately after the student has completed his first year of college--the point at which he is beginning to make his decision as to major subject.

These remarks make it clear that the accurate interpretation and successful use of the information contained in the guidance report will require precision and care. A heavy responsibility will fall both on those who prepare the explanatory materials that accompany the report and on the individual guidance counselor. It must be made clear to the student that the predictions made available and resulting from any particular pattern of test scores apply to the typical examinee receiving these scores. Another way of putting this is to say that the stated predictions will be good predictions for randomly sampled students from those attaining the particular pattern of scores. These predictions, however, should be used only as a base line. It must be emphasized that all predictions are based on data from groups in which there has been much self-selection on variables not measured in a validity study. Therefore, regression coefficients determined by any method, Bayesian or other, must be treated cautiously with reference to any particular applicant. The guidance counselor must bear the final responsibility for combining the information contained in this report with all other information on the student taking into account any special knowledge that may be available on the student. He must be sure that these predictions do not push a student into a program that his own self-understanding would indicate to be a poor choice. He must also look carefully for special qualities that a student may have that would make him particularly attractive to a college. In addition to this the guidance counselor must be able to help the student understand his preferences and utilities and to combine these with the predictions to arrive at a rational decision.

The score reporting form described above is unlikely to be appropriate in that precise form for any specific program. Rather it has been designed to exhibit and emphasize the contributions of the Bayesian methodology. The precise composition of any reporting form will depend on the specific requirements of each individual program. It is the thesis of this paper that the Bayesian prediction methodology will make it possible for such programs to place heavier emphasis on predictions methods. The major advantages of the Bayesian prediction methodology described here are an increased sensitivity when data are scarce and a resulting ability to discard obsolete data and thus keep up with current trends. The fact that this essentially clerical function is done centrally frees guidance personnel to examine the individuality of their own problems more carefully and to devote an increased effort to the task of helping each student to formulate and understand his own goals and to find possible means of attaining these goals.

Implications for Test Construction Methodology

As we have indicated earlier, it is common practice in some testing programs to report only omnibus verbal and quantitative aptitude scores despite the fact that it has long been recognized that human ability is multidimensional rather than two dimensional. In most programs the verbal and quantitative scores are composites based on several different components of human ability. Parenthetically we might say that if either of these scales were unifactorial, serious questions would need to be raised as to their appropriateness. If we accept the fact that ability is multidimensional it would seem strange to use a two dimensional academic aptitude test.

Over the years the suggestion has been repeated many times that multiple scale scores be reported on all academic aptitude tests. This would mean that those charged with guidance and selection responsibilities could combine these scores using the particular regression weights appropriate to the prediction problem that concerns them. There has been a great reluctance on the part of the largest testing organization to do this. The objection has been that because of the relative shortness of the subscales, they would individually be relatively unreliable. Certainly they would not attain the degree of reliability of the usual composite scores. The fear has been that inexperienced test score interpreters would overinterpretate and overemphasize any peculiarities of any of the individual subtest scores. This worry is a legitimate one.

It has not been possible to break this impasse, despite the fact that, in theory, the individual reliabilities of the subscales are unimportant provided that the composite used for prediction (whether that be the usual unit-weight composite or a multiple regression composite) is reliable. The proposed reporting and guidance service resolves this problem by doing the multiple correlation work based on subtests centrally so that only predicted grade point averages and scaled total test scores need to be reported to the students. Each of these quantities will typically have very high reliability.

When prediction work is centralized testing organizations will be encouraged to use a variety of item types in every verbal and quantitative scale. The particular choice of item types and the resulting composite can be made relevant to the particular uses to which the scale is going to be put. Thus most tests could still ostensibly consist of omnibus verbal and quantitative scales, but behind these would stand a varying multiplex of subscales selected and used through regression methods for the particular problems for which the test is being used in a particular application.

This does not mean that the idea of multiple score reporting need be abandoned. Indeed when it can be presumed that students will have adequate professional guidance in interpreting these scores, previous objections may be overcome. An important contribution of centralized prediction methodology is that it lessens the necessity though not the desirability of multiple scale reporting.

With the shift of emphasis from the estimation of ability to the prediction of performance there should be a simultaneous shift in test construction techniques. While such psychometric properties of tests as item difficulty and biserial correlation with total test soore will remain important they will need to be supplemented by questions of item-criterion and subscale-criterion correlation. The empirical validities of individual scales will become at least as important as their reliability. Subscale length, for example, would be manipulated by methods derived from Horst's and Calvin Taylor's work to maximize composite score validity (e.g., see Woodbury & Novick, 1968; Jackson & Novick, 1967; Novick & Thayer, 1969a; Thayer & Novick, 1969) rather than reliability. The inescapable fact is that as essential as are considerations of the psychometric properties of tests, they are not sufficient for the assessment of "relevance in testing".

Acknowledgments

The authors are research statisticians who have worked on Bayesian methods for a number of years. During a year-long stay in Britain and as a result of conversations with Professor D. V. Lindley on certain topics on the foundations of Bayesian inference the senior author became convinced that certain specialized Bayesian techniques would be relevant to the recurring problem of central prediction. Since Professor Lindley had done much work on these techniques an arrangement was made for him to continue and accelerate his development of the specific tools required for Bayesian central prediction.

As the present authors jointly began to consider specific applications it became clear to them that the classical central prediction paradigm was just one aspect of a more encompassing and more relevant guidance-selection paradigm. This reconceptualization was facilitated by discussions over a period of more than a year with Dr. Leonard Cahen. The authors wish to express their appreciation to Dr. Cahen for his effectiveness as a critical discussant and for his perceptive reading of the first draft of this paper.

Since such competence as the authors may have is in statistics rather than guidance they felt it imperative to seek the assistance of persons concerned with guidance, specifically on the problem of communicating the virtues of the Bayesian method to the guidance profession. The response to this appeal was very gratifying and without in any way implying their endorsement of the ideas contained in this paper we should like to acknowledge our debt for important contributions to the following persons: Robert Boldt, Irving Broudy, Richard Burns, Charles Consalus, Harry Harman, Alice Irby, Lyle Jones, Martin Katz, William Laidlaw, Richard Levine, Winton Manning, Robert Solomon, Elizabeth Stewart, and John Winterbottom.

References

Bloom, B. S., & Peters, F. R. Academic Prediction Scales. New York: The Free Press of Glencoe, 1961.

Box, G. E. P., & Tiao, G. C. Bayesian Estimation of Means for the Random Effects Model. Journal of the Americal Statistical Association, 1968, 63, 174-181.

Chauncey, H., & Dobbin, J. Testing--Its Place in Education Today. New York: Harper & Row, 1963.

Cleary, T. A. Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cory, C. H. A Comparison of Four Models for Making Predictions Across Institutions. ONR Technical Report. Seattle: University of Washington, 1968.

Cronbach, L. J. Assessment of Individual Differences. In P. Farnsworth and Q. McNemar (eds.), Annual Review of Psychology. Stanford, Calif.: Stanford University Press, 1956.

Cronbach, L. J., & Gleser, G. C. Psychological Tests and Personnel Decisions. (2nd ed.) Urbana: University of Illinois Press, 1965. [First edition 1955]

Cronbach, L. J., & Snow, R. E. Individual Differences in Learning Ability as a Function of Instructional Variables. Stanford, Calif.: Stanford School of Education, 1969. (ERIC No. ED-029-001.)

French, J. W. Comparative Prediction of College Major-Field Grades by Pure Factor Aptitude, Interest, and Personality Measures. Educational and Psychological Measurement, 1963, 23, 767-774.

Holland, J. L., & Richards, J. M. Academic and Nonacademic Accomplishment. Journal of Educational Psychology, 1965, 56, 165-174.

Horst, P. A Technique for the Development of a Differential Prediction Battery. Psychological Monographs, 1954, No. 380.

Horst, P. A Technique for the Development of a Multiple Absolute Prediction Battery. Psychological Monographs, 1955, No. 390.

Horst, P. Differential Prediction in College Admissions. College Board Review, Fall 1957, No. 33.

Horst, P. The Statewide Testing Program. Seattle: University of Washington, 1961.

Hull, C. L. Aptitude Testing. Yonkers: World Book, 1928.

Jackson, P. H., & Novick, M. R. Maximizing the Validity of a Unit-Weight Composite as a Function of Relative Component Lengths with a Free Total Testing Line. Psychometrika, to appear.

Katz, M. Decisions and Values. New York: The College Entrance Examination Board, 1963.

Katz, M. A Model of Guidance for Career Decision-Making. Vocational Guidance Quarterly, September 1966.

Katz, M. Can Computers Make Guidance Decisions for Students. College Board Review, Summer 1969, No. 72. (a)

Katz, M. Theoretical Foundations of Guidance. Review of Educational Research, 1969, 30, 127-140. (b)

Kelley, T. L. Interpretation of Educational Measurements. Yonkers-on-Hudson, N. Y.: World Book, 1927.

Lindley, D. V. Introduction to Probability and Statistics. Part 2: Inference. Cambridge: University Press, 1965.

Lindley, D. V. A Bayesian Solution for Some Educational Prediction Problems--I. Research Bulletin, 69-57. Princeton, N. J.: Educational Testing Service, 1969. (a)

Lindley, D. V. A Bayesian Estimate of True Scores that Incorporates Prior Information. Research Bulletin 69-75. Princeton, N. J.: Educational Testing Service, 1969. (b)

Lindley, D. V. A Bayesian Solution for Some Educational Prediction Problems--II. Research Bulletin 69-00. Princeton, N. J.: Educational Testing Service, 1969. (c)

Lindquist, E. F. An Evaluation of a Technique for Scaling High School Grades to Improve Prediction of College Success. Educational and Psychological Measurement, 1963, 24, 623-646.

Linn, R. L. Grade Adjustments for Prediction of Academic Performance: A Review. Journal of Educational Measurement, 1966, 3, 313-329.

Lunneborg, C. E. A Research Review of the Washington Pre-college Testing Program. Journal of Educational Measurement, 1966, 3, 157-166.

Lunneborg, C. E., & Lunneborg, P. W. Uniqueness of Selected Employment Aptitude Tests to a General Academic Guidance Battery. Educational and Psychological Measurement, 1967, 27, 953-960.

Lunneborg, P. W., & Lunneborg, C. E. The Differential Prediction of College Grades from Biographic Information. Educational and Psychological Measurement, 1966, 26, 917-925.

Meyer, D. L. Bayesian Statistics. Review of Educational Research, 1966, 36, 503-516.

Novick, M. R. Multiparameter Bayesian Indifference Procedures (with discussion) Journal of the Royal Statistical Society: Series B, 1969, 31, 29-64. (a)

Novick, M. R. A Bayesian Approach to the Selection of Predictor Variables. Paper presented at a conference in honor of Professor Paul Horst at Seattle, Washington, June 22, 1969. Research Bulletin 69-58. Princeton, N. J.: Educational Testing Service, 1969. To appear in the Horst Commemorative Volume. (b)

Novick, M. R. Bayesian methods in psychological testing. Research Bulletin, 69-31. Princeton, New Jersey: Educational Testing Service, 1969. (c)

Novick, M. R., & Grizzle, J. E. A Bayesian Approach to the Analysis of Data from Clinical Trials. *Journal of the American Statistical Association*, 1965, *60*, 81-96. (a)

Novick, M. R., & Hall, W. J. A Bayesian Indifference Procedure. *Journal of the American Statistical Association*, 1965, *60*, 1104-1117. (b)

Novick, M. R., Jackson, P. H., & Thayer, D. T. Bayesian Estimation and the Classical Test Theory Model. (In preparation)

Novick, M. R., & Thayer, D. T. Some Applications of Procedures for Allocating Testing Time. Research Bulletin 69-1. Princeton, N. J.: Educational Testing Service, 1969. (a)

Novick, M. R., & Thayer, D. T. A Comparison of Bayesian Estimates of True Score. Research Bulletin 69-74. Princeton, N. J.: Educational Testing Service. (b)

Robbins, H. A statistical screening problem. In Olkin, et al. (Ed.), *Contributions to probability and statistics--Essays in honor of Harold Hotelling*. Stanford: Stanford University Press, 1960.

Rulon, P. J., Tiedeman, D. V., Tatsuoka, M. M., & Langmuir, C. R. *Multivariate Statistics for Personnel Classification*. New York: Wiley, 1967.

Thayer, D. T., & Novick, M. R. A General Computer Program for Optimally Allocating Testing Time. Research Bulletin 69-50. Princeton, N. J.: Educational Testing Service, 1969.

Turnbull, W. W. Relevance in Testing. Science, 1968, 160, 1424-1429.

Whitla, D. K. Evaluation of Decision Making: A Study of College Admissions. In D. K. Whitla (Ed.), Handbook of Measurement and Assessment in Behavioral Sciences. Reading, Mass.: Addison-Wesley, 1968.

Wolfe, J. H. A Review of Rulon, P. J. et al., "Multivariate Statistics for Personnel Classification". Educational and Psychological Measurement, 1969, 29, 541-544.

Woodbury, M. A., & Novick, M. R. Maximizing the Validity of a Test Battery as a Function of Test Length for a Fixed Total Testing Time. Journal of Mathematical Psychology, 1968, 5, 242-259.