ACT RESEARCH REPORT

No. 38

38

November, 1970

BAYESIAN CONSIDERATIONS IN EDUCATIONAL INFORMATION SYSTEMS

Melvin R. Novick

PUBLISHED BY THE RESEARCH AND DEVELOPMENT DIVISION

THE AMERICAN COLLEGE TESTING PROGRAM

P. O. BOX 168, IOWA CITY, IOWA 52240

ABSTRACT

The development during the past decade of guidance-oriented educational testing programs is surveyed together with the resulting need for statistical methods for educational information systems. Bayesian methods are described as being uniquely capable of combining prior, collateral, and direct experimental information to provide probabilistic statements about parameters descriptive of students, educational programs and their relationships. The Bayesian statistical methods needed for these applications are described.

BAYESIAN CONSIDERATIONS IN EDUCATIONAL INFORMATION SYSTEMS[1]

Melvin R. Novick

For many years students' scores on academic aptitude tests have provided selective colleges and universities with one important piece of information relevant to their decision of whether or not to select a particular applicant. Such tests have had the desirable effect of making admissions decisions for these institutions more dependent on academic promise and less dependent on status and influence. The result has been a broadening of the base of educational opportunity in this country. I am confident that these tests will continue, for some time, to serve this function.

Our educational system now, however, is in the process of redefining its constituency at the postsecondary level (Munday & Rever, in press) to include essentially all students who can effectively benefit from any additional education (Harcleroad, 1970; Novick & Jackson, 1970). This trend is best seen in the recent and projected growth in the number of students attending community colleges. One result of this trend is the growing number of students in nonselective colleges. Decisions of consequence for such students center largely around the choice of program of study.

Concomitant with this growth there has been a broadening of the range of available educational opportunities. If this broadening continues, and if there is an increase in the diversity of training

[1]Paper read at the 1970 Invitational Conference on Testing Problems to be published by Educational Testing Service in the *Proceedings* of the conference. Reproduced by permission of Educational Testing Service, Princeton, New Jersey.

methods to accommodate students with different ability profiles, we shall approach a *meaningful* national policy of open admissions. This does not suggest that any one institution will need to encompass any greater range of programs or any greater number of students than it can effectively handle. It means only that the educational system as a whole will serve a much wider constituency.

In this situation it will be both possible and desirable to maximize the informed participation of each student in the decisions that affect his educational career (Novick & Jackson, 1970). Indeed, to a very great extent it will be the student, not the college, who is the primary decision maker. It will be the student who requires information about himself, the colleges, and the particular programs that may be relevant to his goals. In this context, educational testing becomes just one component of a decision-oriented information transmittal system having a guidance, rather than a selection, orientation.

The American College Testing Program (ACT), since 1964, has provided a guidance-oriented information system, which is now used annually by approximately one million college applicants to both 2- and 4-year colleges and universities. This program provides the student with test scores and a variety of other information about himself. It also provides him with predictions of his potential performance at colleges in which he is interested.

The College Entrance Examination Board has recently begun offering an information system, The Comparative Guidance and Placement Program (CGP), specifically for use in the community colleges. The ACT and CGP programs are alternatives appropriate for students in academic curricula in the community colleges. A new guidance-oriented information system, the Career Planning Program (CPP), is currently under development by ACT for use by students in vocational-technical curricula. The CPP and CGP programs are alternatives for students in these curricula.

Thus, for the past decade, we have been witnessing a continuing reorientation of services offered at the postsecondary level by the major testing organizations (Turnbull, 1968). The present trend will undoubtedly continue, and Bayesian statistics can, I think, make an important contribution in this new setting (Novick & Jackson, 1970).

The Bayesian method is unique in providing a formal mechanism for combining observational information with *prior* information or beliefs to provide *posterior,* or after the sample, probability distributions for parameters of interest such as student abilities, institutional mean values or regression coefficients relating performance criteria to test scores. A typical Bayesian statement made after observing a small random sample of persons would be of the following form: the probability is .95 that the mean ACT English score of examinees from the State of Iowa in the year 1969 lies between 20.4 and 23.2. The length of such a *credibility interval* would depend largely on the number of observations in the sample.

The posterior probability distribution is interpreted by Bayesians as a formal numerical representation of the state of knowledge about the parameter of interest. (It literally carries all of the available information about the parameter.) Certain characteristics of this posterior Bayes distribution are of particular interest. For example, such measures of central tendency as the mean, the median, and the mode are useful as general descriptors, the mode being the most probable value of the parameter. The reciprocal of the variance of the posterior distribution is a measure of the *precision* of available information.

The heart of the Bayesian method is Bayes theorem which says that, given the data, the posterior distribution of the parameter is proportional to the product of (a) the distribution of the data, given the parameter and (b) the prior (or before the sample) distribution of the parameter. The first of these distributions is what is often called the model distribution and is just that used in classical forms of parametric inference. Bayes theorem itself is a straightforward application of the basic theorem of conditional probability and hence enjoys general acceptance. In effect, Bayes theorem adds sample information to prior information to provide a formal representation of posterior information. The Bayesian method may thus justifiably be thought of as a formal system of

information accumulation.

In many simple applications Bayesian credibility interval statements either coincide numerically with classical confidence interval statements or differ only by trivial amounts. The two kinds of interval statements, however, have quite different meanings. The classical statement is "the probability is .95 that the obtained confidence interval will cover the true mean." This is a statement about the interval not the mean. The Bayesian statement is "the probability is .95 that the true mean lies in the specified credibility interval." The Bayesian statement is a direct statement about the mean; many people find it preferable.

The price one pays for the elegance of the Bayesian analysis is the need for specifying a prior Bayes distribution summarizing prior information or beliefs. There is controversy on this point because (a) some people do not wish to interpret probabilities as degrees of belief, but only as relative frequencies as in classical theory and, (b) even accepting a belief interpretation for probabilities there is still a very real problem of just how to quantify these beliefs. The latter problem is particularly acute because in any important study experts will disagree on the evaluation of prior information. Indeed the purpose of the study is typically to resolve such disagreements.

In 1963 a major paper by Edwards, Lindman, and Savage describing Bayesian methods appeared in the *Psychological Review.* This paper described the Bayesian method as an explication of a theory of personal probabilities with which the names of Ramsey (1963), de Finetti (1964), and Savage (1954) are most prominently associated. The impact of this paper was enhanced by the enormous popularity that Bayesian methods were enjoying in business applications, primarily as a result of the efforts of Schlaifer (1959).

The Bayesian personal probability method is described as resting on two foundational supports. The first of these, developed in the *Review* paper, is a theorem showing that if each investigator uses a reasonable prior distribution, all posterior distributions will eventually converge and we will thus have *stable* estimation. Thus, the Bayesian method is shown to have the requisite property of eventually resolving prior differences of opinion.

The second support for the theory is based on an argument due to de Finetti and formalized in a theorem by Savage (1954). In essence the theorem says that if you wish to be sure of behaving in a logically consistent or *coherent* manner in any decision situation, then you must effectively behave as if you have a prior distribution and you must effectively use Bayes theorem. An implication of Savage's theorem is that if you behave in a non-Bayesian way in a betting situation your opponent can specify a sequence of bets that would appear favorable to you and that would, in the long run, almost certainly lead to a loss by you. One might expect these arguments to be compelling, for who would choose to bear both the professional scorn and the economic ruin that logical inconsistency promises to bring.

Many papers have also appeared showing that well accepted principles of classical inference can lead to very unsatisfactory results (Bock & Wood, in press; Cornfield, 1970). For example the usual classical unbiased estimate of a between-group variance component can be negative even though a variance component must, by definition, be non-negative (Novick, Jackson, & Thayer, in press). In contrast, the Bayesian estimate is always non-negative. Despite this, the Bayesian method did not receive on the spot acceptance because of a perceived weakness involving the selection of the prior distribution. According to the personal probability theory each investigator constructs his own prior distribution by means of a self-interrogation or introspection of how he would bet on various possible values of the parameter. No attempt is made to attain any sort of pre-experiment consensus among investigators; rather, great reliance is placed on the principle of stable estimation.

The usual objection raised to personal probabilities is that it is the antithesis of science to let each experimenter select his own prior distribution. Somehow, it is thought, the prior information must depend on prior data. This is very difficult, however, because prior information is typically fragmented and the evaluation of it is

subject to individual interpretation and bias.

It also seems evident that while the business entrepreneur need convince only himself of the reasonableness of his action, the scientist is typically trying to convince someone else—a journal editor, a research grant committee, or an audience such as gathered here today. It seems to me, that this necessitates, that in scientific publication one of two things must be done. Either the prior distribution must be as well justified as anything else in the study, or, for argumentative purposes, the scientist must present a parallel analysis showing that even with a prior distribution that others might specify, the results of the present experiment support his contentions.

The technique I now wish to discuss makes it possible to construct a prior distribution from the data at hand and thus to largely depersonalize personal probabilities. This technique can be used whenever inferences are made simultaneously about a large number of persons, schools, or other experimental units, for example, in estimating the true scores (i.e., expected scores) of members of a well-defined group of examinees. We know that the observed score for a person has an error distribution centered at his true score. But since we treat our examinees as having come from a population of potential examinees, we also have a distribution of (unobservable) true scores. Thus we have the well-known model II, the variance components or random effects model, which has been well-studied along classical lines by many statisticians including Cornfield and Tukey (1956). The model has been used in a semi-Bayesian way to estimate means by Robbins (1954/55) and by Stein (1962). Earlier still, this model was used to estimate means in educational work by Kelley (1927). Recently Bayesian analyses for the estimation of means with this model have been provided by Box and Tiao (1968) and by Lindley (in press) and applied in the field of public health by Cornfield (1969). A comparison of some Bayesian and classical methods has been done by Novick, Jackson and Thayer (in press).

The Kelley (1927) regression estimate of true score given observed score has a form that closely approximates other model II solutions. That estimate is just a weighted average of the person's observed score and the mean observed score in the population, the weights being, respectively, the reliability of the test and one minus the reliability. Thus the regression estimate of true score depends not only on the direct observations on the particular person but also on the indirect or *collateral* information gained from all other observations in the specified group.

This regression estimate makes sense. If we have an unreliable measurement on any person, a heavy weight is given to the mean value of the population of which he is a member and the estimate is regressed back nearly to that value. If our measurement is very reliable it gives little weight to this population value and there is very little regression. In intermediate cases there is only partial regression to the overall mean. Kelley (1927) showed that the overall mean squared error is substantially reduced by using this procedure when the reliability itself is low or moderate.

The various Bayesian and semi-Bayesian approaches to this problem yield results that are very similar to those obtained by Kelley. Robbins (1954/55) captured the spirit of what was being done when he pre-empted the name *empirical Bayes* for his procedure. In effect what is being done here is to use the collateral observations to estimate the parameters of the prior distribution for each person and then to use the direct observations to get the posterior distribution. Robbins' procedure differs from the full Bayesian model II analysis in that he uses a classical method to estimate the parameters of the prior distribution for the Bayesian analysis, while the full Bayesian analysis also does this in a Bayesian way. My own feeling is that the new Bayesian procedures are as empirical as is Robbins' procedure, possibly more so. They are certainly more illuminating theoretically, and only they provide a formal method for combining both prior and collateral information.

A third foundational support for Bayesian work, and particularly for Bayesian model II analysis, is contained in a theorem, due to de Finetti (1964) and generalized by Hewitt and Savage (1955). If our prior information about the

various persons is identical, then we have what de Finetti calls a symmetric or *exchangeable* prior distribution for the person parameters. The de Finetti-Hewitt-Savage theorem states that any exchangeable prior distribution is equivalent to a prior distribution obtained under the assumption that the persons were randomly sampled from some population, and hence that model II is applicable. The strength of this theorem now seems very great. It means that a model II analysis will *typically* be preferable to a model I, i.e., fixed effects analysis (Lindley, in press).

Despite our well-displayed fondness for the Bayesian model II estimation of means, we must acknowledge there can be a problem. It may add to overall efficiency to reduce our estimate of a person's true score because we identify him with some population that has a lower mean true score, but it may not appear fair. Suppose, in a selection situation, one person has his score lowered by this regression to the population mean and a second person from a population with a higher mean true score has his score raised. Suppose further that this results in an inversion in the ordering of the reported scores and that, as a result, the second person is selected for college admissions and the first is not. We would certainly be hard put to convince the first examinee, his parents *and his lawyer* that he had been treated fairly.

We do not mean to suggest that model II cannot be used in a selection situation, only that to do so fairly may require a much more careful selection procedure; one, for example, that considers in a full decision theoretic analysis the differential utility of accepting persons from the different groups. The important point though is that the whole situation changes when the student becomes the decision-maker, i.e., when we are considering a guidance rather than a selection situation. The decision of what to do with this information then falls to the student. He may, for example, want to modify our estimate using information available to him but not to us.

Actually, the above discussion is largely academic with a test like the SAT, which is very long and reports only two scales, and therefore has high subtest reliability. The regression estimates of true scores will then differ little from the observed score. In multi-scale batteries of short subtests the effect on subtest scores will be more pronounced. In such situations one might find merit in reporting the Bayesian multiple regression estimate of each true score given all of the observed scores. This approach has been suggested by Cronbach and Furby (1970) for the estimation of change scores. Since only a single overall population is identified there will be no unfairness to any individual. When the intercorrelations of the subtest scores are more than trivial this can result in a substantial increase in the reliability of each subtest.

When used to estimate institutional parameters or regression coefficients, in either a guidance or a selection context, the model II estimates are also not subject to any unfairness criticism. This application is important because by using prior and collateral information in a Bayesian analysis we can typically obtain any specified degree of precision with a smaller sample size than a model I analysis would require. It really makes no sense to estimate each institutional parameter, or for that matter to do every validity study, as if we were starting from a state of ignorance.

The immediately most important application of the Bayesian model II analysis, in my judgment, is to the estimation of regression parameters. Each of the guidance-oriented testing programs mentioned earlier incorporates predictions of academic performance as an important piece of information to be supplied to the student. The growth in the number and diversity of programs at the community college level and the relative smallness of individual programs suggest that we shall often not have enough data on a particular curriculum within a particular college to estimate the partial regression weights with satisfactory accuracy. Analyses that we have done on data from each of the three guidance testing programs confirm this expectation. The problem will become even more acute as we sharpen our focus on post-training criteria and are then inevitably faced with drastically reduced sample sizes.

What we will need to do is recognize that in carefully specified groupings of community colleges, for example, regression coefficients for a

particular curriculum do not differ too greatly across colleges. We can expect some differences in the regression weights because of minor differences in curriculum content and grading standards, but a great deal of similarity can be expected.

Recently Professor D. V. Lindley of University College London has supplied us with a full Bayesian model II analysis for regression in m colleges. The result of this analysis in the single predictor case is to regress the regression weight for each college towards the average of the regression weights across colleges. Here the amount of regression depends largely on the true variance of the regression weights across colleges and on the sample size within the particular college. According to statistical theory, the Bayesian estimates of the regression weights should, on the average, be more accurate than the usual model I estimates. We have now completed the programming of Lindley's very complex solution to this problem and have applied the technique extensively to the estimation of regression parameters obtained from one testing program. We have done this for both simple linear regression and for multiple regression.

Table 1 gives the results of one such analysis. The usual least squares estimates of model I are given in the first column. Notice that two of these estimates are negative. Neither I nor any person I have consulted really believe that the true values are negative. In the second column the estimates obtained from Lindley's model II Bayesian analysis are given. These values certainly more nearly correspond with what we think the true state of affairs to be.

In order to check the reasonableness of our Bayesian solution, we have also developed a classical model II analysis (Jackson, Novick & Thayer, 1970; Jackson, in press). The third column of Table 1 gives the values obtained from this analysis. The relative closeness of the solutions in columns 2 and 3, and their substantial difference from the solution in the first column, suggest to us that the Bayesian solution is both accurate and useful. Recent data analyses that we have done suggest that predictions based on the ACT Test will similarly benefit from a Bayesian treatment. I should also mention that an empirical Bayes procedure for this problem (Martz & Krutchkoff, 1969) has also recently been published but we have not yet completed our study of this work.

The assumptions upon which the Lindley derivation is based require that this kind of analysis be done by a Bayesian statistician only in close collaboration with an educational specialist. The grouping of colleges into homogeneous groups in order to satisfy the exchangeability assumption may be very important. We have high expectation that empirical work will show that when the Bayesian method is carefully applied it will yield very meaningful improvements in prediction over the classical model I analysis. If this is true, Professor Lindley's work will prove to be a major contribution to guidance technology, and more generally, to the development and use of educational information systems.

Table 1

Comparison of Three Estimates of Regression Coefficients
Comparative Guidance Program—Education Curriculum
Regression of GPA on Vocabulary Score[a]

College No.	Least Squares Estimates	Bayesian	Classical Model II	College No.	Least Squares Estimates	Bayesian	Classical Model II
1	2.2	2.9	2.7	11	1.5	2.7	2.2
2	−1.6	2.0	0.4	12	3.1	3.1	3.1
3	5.1	3.6	4.0	13	2.6	3.0	2.7
4	4.9	3.9	4.4	14	3.4	3.1	3.4
5	2.6	3.0	2.8	15	3.8	3.4	3.5
6	−0.1	2.2	1.7	16	2.2	2.8	2.6
7	9.3	4.4	6.3	17	1.1	2.4	1.7
8	3.4	3.2	3.3	18	3.9	3.6	3.7
9	3.7	3.4	3.5	19	4.0	3.5	3.8
10	0.1	1.9	1.1	20	4.7	3.9	4.3
				21	5.9	4.0	5.0

[a]Acknowledgment is made to Educational Testing Service for making data available for this analysis.

References

Bock, R. D., & Wood, R. Test theory. *Annual Review of Psychology,* **22**, in press.

Box, G. E. P., & Tiao, G. C. Bayesian estimation of means for the random effect model. *Journal of the American Statistical Association,* 1968, **63**, 174-181.

Cornfield, J. The Bayesian outlook and its application (with discussion). *Biometrics,* 1969, **25**, 617-658.

Cornfield, J. The frequency theory of probability, Bayes theorem and sequential clinical trials. In D. L. Meyer & R. O. Collier, Jr. (Eds.), *Bayesian Statistics.* Ninth Annual Phi Delta Kappa Symposium on Educational Research. Peacock Publishers, Inc., 1970.

Cornfield, J., & Tukey, J. W. Average value of mean squares in factorials. *Annals of Mathematical Statistics,* 1956, **27**, 907-949.

Cronbach, L. J., & Furby, L. How should we measure "change"—or should we? *Psychological Bulletin,* 1970, **74**, 68-80.

de Finetti, B. Forsight: Its logical laws, its subjective sources. *Annales de l' Institut Henri Poincare,* 1937, **7**. Reprinted in H. E. Kyburg, Jr. & H. E. Smokler. *Studies in subjective probability.* New York: Wiley, 1964.

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review,* 1963, **70**, 193-242.

Harcleroad, F. F. (Ed.) *Issues of the seventies.* San Francisco: Jossey-Bass, 1970.

Hewitt, E., & Savage, L. J. Symmetric measures on Cartesian products. *Trans. American Mathematical Society,* 1955, **80**, 470-501.

Jackson, P. H. The estimation of many parameters—some simple approximations. *ACT Technical Bulletin.* Iowa City, Iowa: The American College Testing Program, in press.

Jackson, P. H., Novick, M. R., & Thayer, D. T. *Bayesian inference and the classical test theory model II. Validity and prediction.* ETS Research Bulletin 70-32. Princeton, N. J.: Educational Testing Service, 1970.

Kelley, T. L. *Interpretation of educational measurements.* Yonkers on Hudson, New York: World Book, 1927.

Lindley, D. V. The estimation of many parameters. *Proceedings of the Waterloo Conference on the Foundations of Statistics,* in press.

Martz, H. F., Jr., & Krutchkoff, R. G. Empirical Bayes estimators in a multiple regression model. *Biometrika,* 1969, **56**, 367-374.

Munday, L. A., & Rever, P. R. Perspectives on open admissions. In P. R. Rever (Ed.), *Monograph Four: Open admissions and equal access.* Iowa City, Iowa: The American College Testing Program, in press.

Novick, M. R., & Jackson, P. H. Bayesian guidance technology. *Review of Educational Research,* 1970, **40**, (No. 4), 459-494.

Novick, M. R., Jackson, P. H., & Thayer, D. T. Bayesian estimation and the classical test theory model: Reliability and true scores. *Psychometrika,* in press.

Ramsey, F. D. *The foundations of mathematics and other logical essays.* London: Kegan, 1963.

Robbins, H. An empirical Bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* Vol. I. *Theory and statistics,* 1954/55, 157-163.

Savage, L. J. *The foundations of statistics.* New York: Wiley, 1954.

Schlaifer, R. *Probability and statistics for business decisions.* New York: McGraw-Hill, 1959.

Stein, C. M. Confidence sets for the mean of a multivariate normal distribution (with discussion). *Journal of the Royal Statistical Society* B, 1962, **24,** 265-296.

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

ACT Research Reports

This report is the thirty-eighth in a series published by the Research and Development Division of The American College Testing Program. The first 26 research reports have been deposited with the American Documentation Institute, ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Photocopies and 35 mm. microfilms are available at cost from ADI; order by ADI Document number. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress. Beginning with Research Report No. 27, the reports have been deposited with the National Auxiliary Publications Service of the American Society for Information Science (NAPS), c/o CCM Information Sciences, Inc., 22 West 34th Street, New York, New York 10001. Photocopies and 35 mm. microfilms are available at cost from NAPS. Order by NAPS Document number. Advance payment is required. Printed copies may be obtained, if available, from the Research and Development Division, The American College Testing Program. The reports are indexed by the *Current Contents, Education* Institute for Scientific Information, 325 Chestnut Street, Philadelphia, Pennsylvania 19106.

The reports since January 1969 in this series are listed below. A listing of previous reports is included in each of several items published by The American College Testing Program: *Your College Freshmen* (pp. 158-160), *Your College-Bound Students* (pp. 107-109). A complete list of the reports can be obtained by writing to the Research and Development Division, The American College Testing Program, P. O. Box 168, Iowa City, Iowa 52240.

No. 28 *A Description of Graduates of Two-Year Colleges,* by L. L. Baird, J. M. Richards, Jr., & L. R. Shevel (NAPS No. 11306; photo, $3.00; microfilm, $1.00)

No. 29 *An Empirical Occupational Classification Derived from a Theory of Personality and Intended for Practice and Research,* by J. L. Holland, D. R. Whitney, N. S. Cole, & J. M. Richards, Jr. (NAPS No. 00505; photo, $3.00; microfilm, $1.00)

No. 30 *Differential Validity in the ACT Tests,* by N. S. Cole (NAPS No. 00722; photo, $3.00; microfilm, $1.00)

No. 31 *Who Is Talented? An Analysis of Achievement,* by C. F. Elton, & L. R. Shevel (NAPS No. 00723; photo, $3.00; microfilm, $1.00)

No. 32 *Patterns of Educational Aspiration,* by L. L. Baird (NAPS No. 00920; photo, $3.00; microfilm, $1.00)

No. 33 *Can Financial Need Analysis Be Simplified?* by M. D. Orwig, & P. K. Jones (NAPS No. 01210; photo, $5.00; microfilm, $3.00)

No. 34 *Research Strategies in Studying College Impact,* by K. A. Feldman (NAPS No. 01211; photo, $5.00; microfilm, $2.00)

No. 35 *An Analysis of Spatial Configuration and Its Application to Research in Higher Education,* by N. S. Cole, & J. W. L. Cole (NAPS No. 01212; photo, $5.00; microfilm, $2.00)

No. 36 *Influence of Financial Need on the Vocational Development of College Students,* by A. R. Vander Well (NAPS No. not available at this time.)

No. 37 *Practices and Outcomes of Vocational-Technical Education in Technical and Community Colleges,* by T. G. Gartland, & J. F. Carmody (NAPS No. not available at this time.)