DOCUMENT RESUME

ED 457 427 CG 031 163

AUTHOR	Rudner, Lawrence M.
TITLE	Responding to Testing Needs in the Twenty-First Century with an Old Tool.
PUB DATE	2001-00-00
NOTE	13p.; In its: Assessment: Issues and Challenges for the Millennium; see CG 031 161.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Adaptive Testing; Computer Assisted Testing; *Criterion Referenced Tests; *Educational Assessment; Educational Theories; Elementary Secondary Education; *Evaluation; *Measures (Individuals); Testing; Theory Practice Relationship
IDENTIFIERS	Bayes Theorem

ABSTRACT

This articles discusses ways to respond to the need for criterion referenced information using Bayes' theorem, a method coupled with criterion referenced testing in the early 1970s. The theorem determines the most likely classification for an examinee from a dichotomous choice or through placement on a categorical or interval scale. A discussion is included on the application of the theorem to computer adaptive tests, in which an examinee's ability level is estimated during the testing process. Bayesain adaptive testing requires less pretesting; needs smaller item pool; can be applied to criterion referenced and diagnostic testing; can generate classifications based on multiple skills; and requires relatively little statistical knowledge relative to item response testing. The basic framework of this testing can be applied to a range of settings, and several classroom applications are presented. The framework can also be embedded in an intelligence tutoring system to determine mastery after each instructional unit. (Contains 16 references.) (JDM)

Responding to Testing Needs in the Twenty-First Century With an Old Tool

By
Lawrence M. Rudner

Chapter Three

Responding to Testing Needs in the Twenty-First Century With an Old Tool

Lawrence M. Rudner

Abstract

Bayes' theorem is introduced as a method for criterion-referenced testing. This theorem determines the most likely classification for an examinee from a dichotomous choice or through placement on a categorical or interval scale. The application of Bayes' theorem to computer adaptive tests, in which an examinee's ability level is estimated during the testing process and items selected accordingly, is discussed. Relative to item response testing, Bayesian adaptive testing requires less pretesting, needs smaller item pool, can be applied to criterion-referenced and diagnostic testing, can generate classifications based on multiple skills, and requires relatively little statistical knowledge.

Much of modern assessment research and development concentrates on norm-referenced tests, which by definition are designed to rank-order students by placing them on broad continua representing unidimensional traits. The summative information from norm-referenced assessments serves many purposes, but as we enter the twenty-first century, there is a rising call for criterion-referenced information concerning what students know and can do relative to clearly defined desired outcomes of instruction. Although criterion-referenced interpretations of norm-referenced tests are commonplace, the literature from the 1970s and 1980s on criterion-referenced tests can provide some insights to guide current research and practice. As Hambleton and Sireci (1997) point out, the differences between the performance tests of today and the criterion-referenced tests of the 1970s are not fundamental. Both are focused on assessment of what students know and can do.

This article introduces ways of responding to the current clamor for criterion-referenced information using Bayes' theorem—a method that was coupled with criterion-referenced testing in the early 1970s (see Hambleton and Novick, 1973). After introducing Bayes' theorem, I provide some detail demonstrating how it can provide the basis for computer adaptive criterion-referenced tests. I then briefly discuss other potential classroom applications of Bayes' theorem. Specific advantages of using this model are that relatively small data sets are required and that the necessary computations are surprisingly simple.

Bayes' Theorem: A Brief Overview

Rather than placing a student on an ability scale, the goal of a Bayesian approach is to identify the most likely classification for the examinee. This classification may be dichotomous (e.g., master/non-master), polychotomous (e.g., master/at-risk/non-master) or a placement on a categorical or interval scale. A simple example in which the goal is to classify an examinee as being either a master or a non-master is used to illustrate Bayes' theorem. Responses to previously piloted items are used to determine the probabilities of mastery P(M) and non-mastery P(N) and then to classify the examinee based on those probabilities. Lacking any other information about the examinee, let us assume equal prior probabilities, i.e., P(M) = .50 and P(N) = .50. After each item is scored, we will update P(M) and P(N) based on the response to the item.

As givens, we will start with a collection of items for which we have determined the following four probabilities:

1. Probability of a correct response given that the examinee has mastered the material
2. Probability of an incorrect response given that the examinee has mastered the material
3. Probability of a correct response given that the examinee has not mastered the material
4. Probability of an incorrect response given that the examinee has not mastered the material

I will denote these as P(C|M), P(I|M), P(C|N), and P(I|N), respectively. Note that there are different conditional probabilities for each item. These conditional probabilities can be determined from very small-scale, low-cost pilot testing; for example, one approach is to use the percentages of examinees in each group responding correctly or incorrectly. Suppose that on item 1 of the pilot test, 90% of the masters and 40% of the non-masters responded correctly. Because a person responds either correctly or incorrectly, P(C|M) = .90, P(I|M) = .10, P(C|N) = .40, and P(I|N) = .60.

The task then is to update P(M) and P(N) based on the item responses. The process for computing these updated probabilities is referred to as *Bayesian updating, belief updating* (probabilities being a statement of belief), or *evaluating the Bayesian network.* The updated values for P(M) and P(N) are referred to as the *posterior probabilities.* The algorithm for updating comes directly from a theorem published posthumously by Rev. Thomas Bayes in 1763:

$$P(M|C) \times P(C) = P(C|M) \times P(M)$$

Let us suppose our examinee responds correctly to item 1. The probability of a correct response, P(C), is thus 1.0 and by Bayes' theorem, the new probability that the examinee is a master given a correct response is

$$P(M|C) = (.90 \times .5) / 1.0 = .45$$

Similarly, $P(N|C) = P(C|N) \times P(N) = .40 \times .5 = .20$. We can then divide by the sum of these joint probabilities to obtain posterior probabilities, as follows:

$$P'(M) = .45 / (.45 + .20) = .692$$

and

$$P'(N) = .20 / (.45 + .20) = .308.$$

We use these posterior probabilities as the new prior probabilities, score the next item, and again update our estimates for P(M) and P(N) by computing new posterior probabilities. This process continues until all the items have been scored. Equivalently, we could have computed the product of the relevant probabilities (correct or incorrect) for masters and non-masters, then divided by the sum to obtain the last posterior probability.

The Bayesian network defined here is a simple diverging graph. The master/non-master state is causally connected to the set of item responses. When applied to decision-support systems and other expert systems, Bayesian networks are typically much more complex, involving hundreds of interconnected and cross-connected variables (see Lauritzen & Spiegelhalter, 1988; Pearl, 1986). Evaluating such networks is computationally complex. As I have shown here, however, the computations for basic applications are quite simple.

Bayesian Computer Adaptive Testing

Paper-and-pencil tests are typically fixed-item tests in which all examinees answer the same questions within a given test booklet. This is terribly inefficient. Bright individuals have to endure items that cover skills and knowledge they clearly possess. Less able individuals have to suffer through material that is above their ability. These "too easy" and "too difficult" items function like adding constants to an individual's score, providing relatively little if any information about the examinee's true ability. Consequently, large numbers of items and examinees are needed in order to obtain a modest degree of precision, reliability, and validity.

With a computer adaptive test, the examinee's ability level can be iteratively estimated during the testing process, and items can be selected based on a precision-based real-time estimate of the individual's ability. From the pool of items, examinees can be presented with those items that maximize the information about their ability levels. Thus, examinees will receive few items that either are very easy or very hard for them. This tailored item selection results in reduced standard errors and greater precision with only a handful of properly selected items. The time required for testing is greatly reduced, and examinees receive valid, reliable, and legally defensible estimates of their ability. In addition, retesting can occur more frequently without requiring that massive, entirely new item pools be developed and validated.

With the growth of expert systems and the use of artificial intelligence, there has been increasing interest in the use of probability theory and Bayesian networks as a tool to help synthesize observations and generate probabilistic assumptions about current student ability. This information, in turn, may be used to guide the presentation, sequencing, and pacing of instruction. The same mathematical principles have also been proposed as the basis for an attractive form of adaptive testing applicable to a wide range of situations. Relative to item response theory computer adaptive testing (IRT CAT), Bayesian adaptive testing (B-CAT), requires little pretesting and a small item pool. B-CAT can be used with criterion-referenced tests, used to make mastery–non-mastery classifications, incorporated into diagnostic testing, and easily applied to multidimensional assessments. Further, the mathematics of B-CAT are much simpler than those of IRT CAT.

The traditional paradigm for computer adaptive testing is an iterative process with the following steps:

1. A tentative ability estimate is made.
2. All the items that have not yet been administered are evaluated to determine which will be the best one to administer given the current estimate of ability.
3. The best item is administered and the examinee responds.
4. A new ability estimate is computed based on the responses to all of the administered items.
5. Steps 2 through 4 are repeated until a stopping criterion is met.

Bayesian computer adaptive testing follows the same five steps. Instead of estimating ability, however, B-CAT estimates classification probabilities. Frick (1992), and Madigan, Hunt, Levidow, and Donnell (1995) explain how Bayesian networks can be used as the CAT framework. Welch and Frick (1993) provide a excellent and readable overview of the topic. With B-CAT, the goal is to determine the most likely classification for the examinee. This classification may be dichotomous (e.g., master/non-master) or may involve placement on a categorical or interval scale. With B-CAT, conditional probabilities are the givens and posterior probabilities are iteratively estimated. Possible stopping criteria include time, number of items administered, or change in ability estimate. With Bayesian adaptive testing, a desired alpha and beta level can be employed.

To explain B-CAT, I provide an example where the goal is to classify an examinee as being either a master or a non-master. Basically, the new posterior probabilities are computed after each item is administered. One stops administering items when the probability of mastery is sufficiently high or low. Items are selected from the pool of remaining items to maximize information or minimize a loss function.

As givens, let us assume a collection of items for which the four probabilities outlined previously have been determined. We will use the database of four items shown in Table 3.1 (the data for this example come from Welch and Frick, 1993). For the example, we will assume these items are administered sequentially. Ideally, the next item to be administered would be the item that minimizes $P(C_i|M) - P(C_i|N)$; that is, the item most likely to yield the largest change in the posterior probabilities.

Table 3.1. Sample Probabilities of Correct and Incorrect Responses by Masters and Non-masters

	Masters (M)		Non-masters (N)	
Item (i)	$P(C_i|M)$	$P(I_i|M)$	$P(C_i|N)$	$P(I_i|N)$
1	.89	.11	.65	.35
2	.81	.19	.24	.76
3	.92	.08	.47	.53
4	.98	.02	.86	.14

Note that for each *i*, $P(C_i|M) + P(I_i|M) = 1.00$ and $P(C_i|N) + P(I_i|N) = 1.00$. Responses are dichotomous states—an examinee responds either correctly or incorrectly. The goal is to classify the examinee as most likely being a master or a non-master based on his or her responses to selected items. Again, lacking any other information about the examinee, we will assume equal prior probabilities of being a master or non-master (i.e., P(M) = .50 and P(N) = .50). After each item is given, we will update P(M) and P(N) based on the response to the item.

Let us suppose our examinee responds incorrectly to item 1. By Bayes' theorem, the new probability that the examinee is a master given an incorrect response is

$$P(M|I_i) = P(I_i|M) \cdot P(M) / P(I_i)$$

We know that the examinee has responded incorrectly, so $P(I_i) = 1.00$ and $P(M|I_i) = .11 \text{ x } .5 = .055$. Similarly, $P(N|I_i) = P(I_i|N) \text{ x } P(N) = .35 \text{ x } .5 = .175$. We can then divide by the sum of these joint probabilities to obtain posterior probabilities, as follows:

$$P'(M) = .055 / (.055 + .175) = .239$$

and

$$P'(N) = .175 / (.055 + .175) = .761$$

We next use these posterior probabilities as the new prior probabilities, select a new item, and again update our estimates for P(M) and P(N) by computing new posterior probabilities. We iterate the process until some specified stopping criterion is reached. Wald's (1947) Sequential Probability Ratio Test appears to be favored in the literature.

To continue the example, let us assume that the examinee responds correctly to item 2, incorrectly to item 3, and incorrectly to item 4. Table 3.2 shows the resultant probabilities for all four items.

Table 3.2. Calculations for Probability of Mastery Based on Four Sample Responses

Item (i)	Response (R_i)	State(S)	Prior Probability	$P(S|R_i)$	Joint Probability	Posterior Probability
I	I	Master	.500	.11	.055	.239
		Non-master	.500	.35	.175	.761
2	C	Master	.239	.81	.194	.515
		Non-master	.761	.24	.183	.485
3	I	Master	.515	.08	.041	.138
		Non-master	.485	.53	.257	.862
4	I	Master	.138	.02	.003	.024
		Non-master	.862	.14	.121	.976

At each iteration, the subsequent item can be selected to maximize the expected change in the posterior probability. After administering these four items, the probability that our examinee is a non-master given this response pattern is .976. Had we set a minimum posterior probability of .975 (1 - $\alpha/2$) as the stopping rule, we could then terminate item administration.

In theory, this approach to CAT has the advantages of IRT CAT plus several crucial advantages of its own:

- It can incorporate a small item pool.
- It is simple to implement.
- It requires little pretesting.
- It can be applied to criterion-referenced tests.
- It can be used in diagnostic testing.
- It can be adapted to yield classifications on multiple skills.
- It is easy to explain to non-statisticians.

In recent years there has been growing theoretical interest in B-CAT among the educational testing community (De Ayala, 1990; Frick, 1992; Lewis & Sheehan, 1990; Segall, 1996; Spray & Reckase, 1996; van der Linden & Hambleton, 1997). There have also been a handful of small studies evaluating B-CAT. De Ayala (1990), Jones (1993), Spray and Reckase (1996), and Welch and Frick (1993) all found advantages to B-CAT relative to other forms of adaptive testing. These studies, however, were typically limited to one examination and to relatively small samples. B-CAT is also featured as the engine behind at least one large company offering intelligent tutoring system development services (Gemini Learning Systems Inc.: http://www.gemini.com).

Classroom Applications

The basic framework described in this article is applicable to a wide range of settings. For example, the framework can be used to score a diagnostic pretest. Here the pretest would cover a variety of skills. A pilot test would determine the probabilities of responding correctly for people who have mastered each skill and the probabilities for those that have not done so. After the test is given to an individual, the probabilities of mastery for each skill could be computed. The resultant list would identify which skills have been mastered and which are likely in need of attention. One could go further and model specific misconceptions (e.g., the examinee sums denominators when adding fractions). Here the relevant probability would be likelihood of selecting a particular incorrect option (or generating a particular type of wrong answer) given that an examinee has a specific misconception. Such a test would not only provide mastery information but identify specific areas to correct.

The framework is also applicable to multidimensional items and tests. One could write items, for example, that require the application of mathematical skills to solve a science problem. A pilot test would need to be administered to compute the probability of responding correctly to each item given mastery of the mathematics skills and the probability of responding correctly given mastery of the science skills. The single test with complex items could then be scored, using the Bayes' theorem and information about each skill area.

Finally, the framework can be embedded in an intelligent tutoring system to determine mastery after each instructional unit, tailor individualized instruction to characteristics of the student, and adapt that instruction as the student learns material. This would again require a collection of pretested items that assess the concepts covered by each instructional unit.

Research Questions

Some concern has been raised concerning the sensitivity of Bayesian networks to misspecified prior probabilities. This is not really a concern with B-CAT, as the system will converge after the administration of only a few items, as it does with IRT CAT. Our concerns are (a) whether B-CAT truly leads to efficient and accurate state classifications, and (b) the sensitivity of Bayesian networks to misspecifications of the conditional probabilities. Probability theory defines expectations over large data sets and large samples. Yet, with B-CAT, we are interested in making inferences about individuals based on small data sets. Thus, B-CAT is a theory that has yet to be demonstrated to work in realistic situations. Bayesian conditional probabilities are based on either qualitative judgments or sampled empirical data. In either case, the specified conditional probabilities are not the same as true conditional probabilities. One is working with estimates, not true values, and the resulting inherent error can seriously bias the results. Shrinkage could be an issue, and the effect that error in the conditional probabilities may have on the posterior probabilities and the number of items needed is not clear.

References and Resources

One can easily experiment with simple Bayesian networks using any of a large variety of readily available, free software packages. A search on the Internet in November 2000 for "Bayesian Network

Software Packages" yielded more than 20 free packages that could potentially be applied. Two that I have tried are Hugin Lite and Genie.

Bayes, T. (1763). Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* 53, 370–418.

Chamiak, E. (1991). Bayesian networks without tears. *AI Magazine,* Winter.

De Ayala, R. J. (1990). A simulation and comparison of flexilevel and Bayesian computerized adaptive testing. *Journal of Educational Measurement,* 27(3), 227–239.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research 8*(2), 187–213.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*(3), 159–170.

Hambleton, R. K., & Sireci, S. G. (1997). Future directions for norm-referenced and criterion-referenced achievement testing. *International Journal of Educational Research, 27*(5), 379–393.

Jones, W. P. (1993). Real-data simulation of computerized adaptive Bayesian scaling. *Measurement and Evaluation in Counseling and Development, 26*(2), 143–151.

Lauritzen, S. L., & D. J. Spiegelhalter. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society,* Series B, *50*(2), 157–224.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*(4), 367–386.

Madigan, D., Hunt, E., Levidow, B., & Donnell, D. (1995). *Bayesian graphical modeling for intelligent tutoring systems.* (Technical Report). Seattle: University of Washington.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks *Artificial Intelligence, 29,* 241–288.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 6*(2) 331–54.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405–414.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* London: Springer Verlag.

Wald, A. (1947). *Sequential analysis*. John Wiley and Associates.

Welch, R. E., & Frick, T. (1993). Computerized adaptive testing in instructional settings. *Educational Training Research and Development, 41*(3), 47–62.

About the Author

Lawrence M. Rudner is director of the ERIC Clearinghouse on Assessment and Evaluation and assistant director of the Maryland Assessment Research Center for Education Success at the University of Maryland, College Park. He earned his Ph.D. in educational psychology and evaluation from Catholic University of America and also holds an MBA in finance from the University of Maryland, College Park. Among his areas of interest are educational measurement and information science. Co-author of *What Teachers Need to Know about Assessment* (Washington, DC: National Education Association), Rudner also serves as editor of *Practical Assessment, Research and Evaluation* and associate editor of *Applied Measurement in Education.* In addition, he chairs the AERA Telecommunication Committee.

Counselor Education

Assessment in Counselor Education: Admissions, Retention, and Capstone Experiences
Irene Mass Ametrano and Sue A. Stickel

Revitalizing the Assessment Course in the Counseling Curriculum
Albert B. Hood

The Pedagogical Basis for Multifaceted Assessment in Counselor Education
Barbara D. Yunker and Mary E. Stinson