DOCUMENT RESUME

ED 401 302 TM 025 851

AUTHOR Bejar, Isaac I.
TITLE Generative Response Modeling: Leveraging the Computer as a Test Delivery Medium.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-96-13
PUB DATE Apr 96
NOTE 44p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Computer Assisted Testing; *Cost Effectiveness; Estimation (Mathematics); Monte Carlo Methods; *Responses; *Test Construction; Test Items; Writing Evaluation; Writing Skills
IDENTIFIERS Response Model

ABSTRACT

Generative response modeling is an approach to test development and response modeling that calls for the creation of items in such a way that the parameters of the items on some response model can be anticipated through knowledge of the psychological processes and knowledge required to respond to the item. That is, the computer would not merely retrieve an item from a database, as is the case in adaptive testing, but would compose it, or assist in doing so, according to the desired specifications. This approach to assessment has implications for both the economics and validity of computer administered tests. To illustrate the concept, a system for measuring writing skills is outlined in which the examinee is expected to rewrite sentences, rather than just recognize errors in a sentence, using a multiple-choice format. The possibility of estimating the psychometric parameters of items based on a psychological analysis of the response process is examined, and shown to be feasible. A Monte Carlo study with 100 simulated examinees at each of 6 ability levels is presented, which investigated the possibility of compensating for that imprecision when estimating ability or proficiency. It is concluded that a generative approach is feasible, and can be a mechanism for taking advantage of the considerable investment required for computer-based testing. (Contains 5 figures, 7 tables, and 59 references.) (Author/SLD)

RESEARCH REPORT



# GENERATIVE RESPONSE MODELING: LEVERAGING THE COMPUTER AS A TEST DELIVERY MEDIUM

**Isaac I. Bejar**



**Educational Testing Service**
**Princeton, New Jersey**
**April 1996**

Running Head: GENERATIVE RESPONSE MODELING

# Generative response modeling: Leveraging the computer as a test delivery medium

Isaac I. Bejar

Educational Testing Service, Princeton, New Jersey

Abstract

Generative response modeling is an approach to test development and response modeling that calls for the creation of items in such a way that the parameters of the items on some response model can be anticipated through knowledge of the psychological processes and knowledge required to respond to the item. That is, the computer would not merely retrieve an item from a database, as is the case in adaptive testing, but would *compose it,* or assist in doing so, according to desired specifications. This approach to assessment has implications for both the economics and validity of computer administered tests. To illustrate the concept, a system for measuring writing skills will be outlined where the examinee is expected to rewrite sentences, rather than just recognize errors in a sentence, using a multiple choice format. The possibility of estimating the psychometric parameters of items based on a psychological analysis of the response process will then be examined and shown to be feasible. Such estimates are less precise than estimates based on large samples of test takers. A Monte Carlo study is presented to investigate the possibility of compensating for that imprecision when estimating ability or proficiency. The paper concludes that a generative approach is feasible, and can be a mechanism for taking advantage of the considerable investment required for computer-based testing.

# Generative response modeling: Leveraging the computer as a test delivery medium

Although the use of computers in testing was envisioned several decades ago (e.g., Green, 1964; Weiss, 1978), it is only recently that nationally administered tests have become operationally feasible for licensing exams (e.g., Lewis and Sheehan, 1990), placement (e.g., Ward, 1984), and admission tests (e.g., Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., Thissen, D., 1990). Interestingly, the major hurdles along the way to implementation were not psychometric per se, but rather organizational, economical, and technological. Indeed, from the point of view of a testing organization, the implementation of computer-based testing requires a substantial "reengineering" of major aspects of the test production process. However, had a research foundation for creative use of the computer not been laid out well before the wide availability of computing power (e.g., Lord, 1980), it is likely that today's computer-based tests would not have taken full advantage of the computer as a medium of delivery. Research must continue to insure that this will be the case in the future and that better tests, not just more economical or convenient ones, will be developed. One possibility is to take advantage of the computer to administer performance-based or open-ended exams that may not have been feasible in a paper and pencil format. A major hurdle in the implementation of performance-based exams is the cost of scoring. Thus, the use of computers to score open-ended responses (e.g., Bejar, 1991; Bennett, 1993; Braun, Bennett, Frye & Soloway, 1990; Breland & Lytle, 1990; Freedle, 1988; Kaplan, 1992; Oltman, Bejar & Kim, 1993) should be encouraged. In this paper, I discuss the possibility of a generative approach to assessment, whereby the computer is used to assist in the creation and prompt evaluation of items.

Generative response modeling (Bejar, 1993) is an approach to test development and response modeling that calls for the creation of items in such a way that the parameters of the items on some response model can be anticipated through knowledge of the psychological processes and knowledge required to respond to a test. That is, the computer would not merely retrieve an item from a database, as is the case in adaptive testing, but would *compose it,* or assist in doing so, according to desired specifications. In the next section, I will expand the concept of generative response modeling, which, as we will see, has implications for both the economics and validity of computer administered tests. Next, I will outline a system for measuring writing skills, where the examinee is expected to rewrite sentences rather than just recognize errors in a sentence, using a multiple choice format. The possibility of estimating the psychometric attributes of items based on a psychological analysis of the response process will then be examined and shown to be feasible, but also less precise than estimates based on large samples of test takers. The inherent imprecision of item parameter estimates based on an a priori procedure could be compensated when estimating person or ability parameters through an adaptive testing procedure. That possibility is investigated through a Monte Carlo experiment, in order to assess the psychometric feasibility of such an approach to test design. Finally, some conclusions regarding the feasibility of generative modeling are offered.

## Historical and current context

Generative response modeling has its roots in computer-assisted instruction (e.g., Hively, 1974; Uttal, Rogers, Hieronymous & Pasich, 1970). An early example can be found in Fremer and Anastasio (1969), an extensive summary of those early efforts can be found in Roid and Haladyna (1982), and a shorter one in Bejar (1983). Uttal, et al. (1970) used the term "generative instruction" to describe an alternative to the machine learning efforts of the 1960s, which were

based on Skinnerian principles, and viewed learning as a matter of reinforcing the bond between stimulus and response. By contrast, generative instruction aimed to diagnose the source of difficulties in learning. This more cognitive objective has prevailed and has been subsequently elaborated in the context of arithmetic instruction by Brown and Burton (1978), among others, into a branch of cognitive science known as student modeling (e.g., Clancey, 1986; van Lehn, 1988; Martin & van Lehn, 1995; Mislevy, 1995). In short, a generative approach takes advantage of psychometric and psychological science for both content and response modeling and can be implemented in a variety of domains. These domains include ability and achievement testing, as well as the measurement of complex skills, such as troubleshooting, clinical diagnoses, and pedagogical skills (Bejar, 1993). There are a growing number of projects demonstrating the feasibility of the generative approach (e.g., Bejar 1990; Bejar & Yocom, 1991; Irvine, Dunn & Anderson, 1990; Meisner, Luecht & Reckase, 1993; Wolfe & Larson, 1990; Hornke, 1986). An area that seems ready for generative modeling is statistical problem solving, where Nitko and Lane (1990) have laid a foundation for item generation.

Because of the current views on validity (Messick, 1989), generative modeling has obvious validity and efficiency implications. The validity evidence for a test developed under a generative framework is improved because a detailed model of performance underlies the construction of items. The model of performance is often recast as a set of attributes of features, and as the relation of those features to item statistics. The efficiency implications follow because it is possible to by-pass the pretesting process entirely, or in part, once the system has been sufficiently validated.

The delivery of tests by computer can be leveraged--indeed, can be essential for--the development of such systems. For example, consider the case of a computer-assisted item writing system being developed in support of an existing item type. An item designer, in principle--although I do not underestimate the complexity of the logistics--can submit a set of new items for "live pretesting" by taking advantage of the network of computers in place to administer the tests as well as a suitable item design system. The data would then be returned to the item designer within a few days, rather than months, and analyzed in such a way as to judge the accuracy of item statistics. At that point, the item designer could revise the items, or the model relating item statistics to item attributes, or the theory or framework that generates the attributes, and the process repeated as necessary. (Although in principle, data can be returned in days, sampling considerations may suggest longer periods of data collection to insure that the data are from a representative sample from the entire test taking population (C. Tucker, personal communication, April 6, 1995). Eventually, when the model of performance has reached a sufficiently high level of refinement, it may not be necessary to pretest items at all. Two major practical advantages of such a system are: 1) Learning on the part of the item designer takes place, so that in time it is possible to generate items with well-estimated item statistics; 2) The process takes advantage of the existing infrastructure for test delivery and serves to make its development more affordable. Making maximum use of the existing computing resources may be essential to the long-term economic success of computer-based testing. Computer-based testing is more vulnerable to maintaining item security, because there is less control over the rate at which a given item is "exposed", and therefore items need to be produced at a much higher pace.

In short, generative modeling can be seen as an approach to educational and psychological measurement that integrates content and response modeling into a unified framework. Item creation is guided by knowledge of the psychology of the domain, and concomitantly psychometric descriptions (e.g., parameters estimates on an appropriate response model) are attached to the generated items. Then, every time a test is administered, the model of performance is challenged by contrasting predicted and observed psychometric description. This approach has much in common with other efforts to develop and validate psychologically-inspired tests or batteries (e.g., Frederiksen, 1986; Guttman, 1969, 1980; Kyllonen, 1990; Nichols, 1994), and seems especially useful in an environment where computer-based test delivery is already in place.

## Generating and scoring authentic test material: writing assessment

Users of test information would like a single test to be as informative as possible. For example, in an educational context, a test is ideally a diagnostic instrument. Almost by definition, a diagnostic test must probe very detailed aspects of a student's knowledge (see e.g., Tatsuoka & Tatsuoka, 1992.). However, more recently, educators have become interested in the authenticity of tests (e.g., Baron, 1991) and the need to integrate assessment and instruction (Snow & Mandinach, 1991). According to Dwyer (1994), the concept of authentic assessment is by no means unitary. It refers not only to the content and format of the assessment, but also to the control of the assessment process. Concern with control of the assessment process is clearly "metapsychometric", but it appears that, in principle, there need not be disharmony between demands for authenticity and measurement theory. In fact, they converge in at least one important respect, namely, the need for a multiplicity of measures to fully capture students' current skills and achievement. The assessment of writing skills illustrates this point.

The value of the multiple measures in the assessment of writing was demonstrated by Ackerman and Smith (1988). They began with a model of the writing process postulated by Flower and Hayes (1980), and from an analysis of that model, constructed direct (i.e., essay) and indirect measures of writing, including open-ended discrete tasks, as well as multiple choice items. The authors concluded that both types of assessment were needed to fully assess writing: "practitioners interested in reliably measuring all aspects of the writing process characterized by this continuum may require the use of both indirect *and* direct methods of assessment." (p. 126-127). Ackerman and Smith's (1988) results suggest that to comprehensively assess individual differences in some content domain, it may be necessary to assess facets of that domain through modes appropriate to the different facets.

A useful, but less demanding first step toward that goal is to focus first on single sentences. Figure 1 shows the components of a generative model proposed by Bejar (1988) for the diagnostic assessment of writing, where examinees are presented sentences and are expected to detect those that exhibit a flaw. The test taker is expected to rewrite the sentence to remove the flaw. By not providing clues as to the nature of the flaw, such a system has the potential to provide more authentic assessment. (Ironically, as software products directed towards writers improve, the skills tapped by a multiple choice format may become more authentic! Specifically, grammar correction programs, which are now built into most word processors, offer suggestions as to what might be wrong with a sentence. As the manual accompanying such programs stress, it is the user's responsibility to *choose* from several possibilities how to correct the sentence).

---

Insert Figure 1

---

One component of the system is a database of sentences which have been chosen because of their suitability for assessing writing skills. In this hypothetical system, an item is created by retrieving a sentence from the database and transforming a grammatically acceptable sentence by introducing a specific error into it. The sentence is then presented to the examinee who must then decide whether it has an error or not. The four outcomes of the interaction of the examinee with the sentence are shown in Figure 2. If the sentence has no error and the examinee rewrites it, partial credit would be given, provided the rewrite did not alter the meaning of the sentence. Otherwise, no credit would be given. Full credit would be given if the examinee leaves the sentence unchanged. However, if the sentence has an error, full credit would be given for rewriting the sentence and removing the error in such a way as to preserve the meaning of the sentence.

---

Insert Figure 2

---

The scoring module needs to be able to detect a large number of writing errors and assess the equivalence in meaning of different sentences. Bejar (1988) studied the effectiveness of various programs for "grammar checking" by studying their ability to recognize the errors tested by the Test of Standard Written English (TSWE). Off-the-shelf consumer software developed at the time, was found to be inadequate for that purpose. However, one package known as WordMap (Lytle, 1986) was shown to be far more advanced and capable of expansion. The

results of the analysis can be found in Bejar (1988), and suggest that WordMap was sufficiently advanced to deal with many of the constructions used in the TSWE. Moreover, WordMap was easily expanded by the developer to deal with errors or constructions with which it was not already familiar.

The determination of equivalence of meaning is substantially more complex and would have to be solved before the system in Figure 1 could become operational. Although the problem of determining the meaning equivalence of two sentences is not a trivial one, the task may be simplified in this case, because a representation of the meaning of the sentence on which the item is based could be stored along with the sentence. The problem is then reduced to representing the meaning of the rewritten sentence and comparing the two representations. The problem could be facilitated further by imposing constraints on the rewrite, such as not allowing the introduction of new words. Nevertheless, mapping a sentence to a different representation is likely to be difficult, although progress is inevitable (see e.g., Kaplan & Burstein, 1994). Additionally, computational linguistics have taken a statistical turn (e.g., Charniak, 1993) that appears useful for measurement purposes, by facilitating the inferring of scoring keys from a corpus of answers to a question. Developments in statistical theory known as *statistical learning theory* (e.g., Vapnik, 1995), are also conducive to the solution of this problem.

## Modeling item difficulty

In the previous section, I discussed how items might be generated and scored. By itself, item generation does not constitute generative modeling, which also requires the assignment of item parameter estimates to generated items. In this section, I will discuss that assignment. This is where psychology and psychometrics have a chance to dovetail. Psychology and

psychometrics can be brought together through the parameters on some response model. For example, through an understanding of the psychology of sentence comprehension, it might be possible to model the recognition of error in sentences and their correction. The data presented below regarding this possibility are limited but promising.

In an earlier investigation, Bejar (1983) demonstrated that expert test developers encountered unexpected problems in estimating the difficulty of items from the Test of Standard Written English (TSWE). The TSWE is a test designed to assess writing skills through multiple-choice items, where the examinee is expected to choose the location of an error in a sentence such as, subject verb agreement, comma splice, etc., or to indicate a rewrite of a sentence that removes an error. In a subsequent investigation, Bejar, Stabler and Camp (1987) studied the possibility of estimating item difficulty of such items through a psycholinguistic analysis of the sentence. Sentences corresponding to 40 items were syntactically analyzed and several features of the sentences were computed, ranging from the number of words in the sentence to various characterizations of the parse tree of a given sentence. The analysis suggested that two features of the parse tree were a potent predictor of difficulty, the "depth" of the sentence and the "depth of the error."

These depth measures can be illustrated by means of two examples. Figure 3 and 4 show the two sentences and their parse tree. The depth of the sentence refers to the maximum number of segments from the top node, labeled "S", down to the lowest level of the tree. Figure 3 shows a very easy item, with difficulty of 6.6 on the delta metric. (The "delta metric" is a nonlinear transformation of proportion correct in use at Educational Testing Service. It is set to a mean of 13 and standard deviation of 4.) The depth is 6. Figure 4 shows an item of medium difficulty,

11.9 on the delta metric. The depth of the sentence in that case is 10. The depth at which the error is located can be similarly defined as the number of segments from the top node down to the error itself.

---

Insert Figure 3 and 4

---

Of course, in practice, the relationship between difficulty and depth measures is not a perfect one, as suggested by these examples. However, it was a more potent predictor of difficulty than the combined ratings of four test development experts. Bejar, et al. (1987) also found that the depth in the sentence at which the error was found was also an important determinant of difficulty. Table 1 shows the analysis for regressing delta on the combined rating of experts, the depth of the sentence, and the depth of the error. As can be seen, the combination of expert judgment, depth, and depth of error, leads to a fairly high level of prediction. In principle, any collateral information about a sentence could be stored in the system described in the previous section, and could be used to help predict the difficulty of items (e.g., Mislevy, Sheehan, & Wingersky, 1993). Although, ideally, the collateral information would be motivated by a theory or framework concerned with the constructs under measurement (e.g., Sheehan, 1995).

---

Insert Table 1

---

The foregoing illustrates an approach to predicting the difficulty of items generated by the system described in the previous section. In addition to syntactic measures such as depth measures, other features might be identified from research in sentence comprehension. For example, an obvious factor likely to affect difficulty is the type of error. At the moment, we lack a comprehensive model of performance on error detection in sentences. A source of information for identifying such a model is the collection of items that have previously appeared on tests like the TSWE. By examining the difficulty of previously administered TSWE items, it may be possible for a team of researchers including linguists, psycholinguists, writing teachers, and psychometricians to infer such a model, which may be of as much interest to both linguists and psycholinguists. The initial model need not be perfect, because it can be updated as items are generated and responses to those items analyzed, with an eye to fine tuning the system.

The scheme described above is a general one and not limited to TSWE-like items. Other item types can be handled in a similar manner, but with a different set of variables entering into the model of difficulty. In general, any item that can be represented as the embedding of a feature to be detected, such as an error embedded in a sentence, as in the case of TSWE; a missing word, as in the case of sentence completion items; a subfigure embedded in a larger figure, as in hidden-figure items; and a design flaw, as in "red lining" in architecture, etc., could be implemented by means of the system like the one shown in Figure 1. How precise the difficulty model needs to be is a matter for further research. The methodology of expected response functions (Mislevy, Wingersky, & Sheehan, 1994; Lewis, 1985), as well as Monte Carlo simulations, can be used to investigate the effect of uncertainty in item parameters and their impact on ability estimation. Results from a Monte Carlo study are presented below.

## A Monte Carlo Feasibility Study

In this section, I offer evidence to suggest that when the purpose of assessment is the estimation of a global ability or proficiency level, it may be possible through an adaptive testing procedure to compensate for the inherent inaccuracy of a priori item parameter "estimates". Thus, less precise estimates of item difficulty might be tolerable, if they could be compensated by a longer adaptive test. (Readers unfamiliar with adaptive testing methods are referred to Wainer, et al. (1990). In the following discussion, Item Response Theory (IRT) is assumed as the psychometric framework. In general, however, there is no necessary link between IRT and generative modeling. For example, a generalizability perspective could be adopted where the item pool consists of a fixed set of classes, each with a "master" item from which isomorphs are produced (e.g., Bejar & Yocom, 1991). A given examinee then receives an item, chosen randomly from each class. In this case, the interest would be on the generalizability of scores. In particular, the variance component due to the use of isomorphs would gage how well the "isomorphing" process was carried out. The generalizability or decision consistency to such data would assess the reliability of scores over all randomly parallel forms that can be constructed from the pool. Such a design is especially applicable in situations where the purpose of the exam is to determine whether the examinee is above some previously established cut score. Here we are concerned with a different test design, namely, an adaptive test design, which is often done from an IRT perspective, and often presupposes an item pool calibrated assuming one of several response models. The pool is usually calibrated, such that item parameters are estimated and put on a common metric through a pretesting or an on-line process (e.g., Jones & Zhiying, 1994). In a generative approach, instead of collecting item response data to estimate item statistics or parameters, items and parameter estimates are produced algorithmically, or at least according to a

set of principles. Parameter estimates produced in this fashion are inherently less precise that parameter estimates based on large samples. The major problem this presents is in the estimation of ability. In this section, I will present the results of a study to assess the effect of error in item parameter estimates on the estimation of ability and show that the effect of errors of estimation can be compensated, to some extent, through an adaptive testing procedure. In practice, a fully generative approach can be supplemented, such that newly created items would be pretested on a small sample. The results below are for a theoretical scenario where there is no pretesting data available to supplement the item parameter estimates obtained from the generative procedure.

Some researchers have addressed the effect of errors in item parameter estimates on ability estimation, usually to study misspecification error (e.g., Wainer & Thissen 1987; Vale 1985), whereby the data follow a model different from the assumed model. Here, however, we are interested in the effect of the imprecision of item parameter estimates on ability estimates when the underlying response model is correct but imprecisely estimated. In what follows, we assume a one-parameter logistic model for both the data and the estimation of ability, $\theta$. The absence of a guessing parameter is justifiable because, in practice, items from a generative procedure are likely to be open-ended. However, the assumption of equal discrimination required by a 1-parameter model may not be reasonable, in general, but will be assumed here for illustrative purposes.

The Monte Carlo experiment presented below poses the question of the effect of item parameter imprecision on ability estimation when items are administered adaptively. One hundred simulated examinees at each of six levels of $\theta$ (-3, -2, -1, 0, 1,2,3) were administered an adaptive test until a certain number of items were reached. The true parameters are used to

determine the probability of answering the item correctly, whereas the parameters with errors, or presumed item parameters, are used to estimate ability. The simulation assumes that there is an infinite item pool available. Although that assumption is unrealistic, the simulation is not meant to fully answer the question of the effect of item imprecision on ability estimates, but rather to assess the feasibility of a generative approach. The process used to generate the data is described in Figure 5:

---

Insert Figure 5

---

This process was repeated for each ability level and replicated 100 times. The outcome of the simulation consists of the mean and standard deviation of the ability estimates over 100 replicates for a given value of $\theta$. The program that implements the simulation allows the investigator to set several parameters of the simulation, including the level of error in the difficulty item parameter. The results presented below are the result of a 2x2 design. One factor is the number of items, which has two levels, 20 and 30 items. The other factor is the level of error in the difficulty parameter, characterized as the standard deviation of the error component added to the true value of the difficulty, with two levels, .50 and 1.0. The design appears in Table 2 and shows a description of each data set. For comparison, a simulation with 20 items and no error in the difficulty parameter was also run.

---

Insert Table 2

---

Table 3 summarizes the simulation study. The table consists of five panels corresponding to each of the five conditions. The first column is the ability level of the simulated cases or $\theta$. The second column, Mean($\theta$), reports the mean maximum likelihood estimate (MLE) of $\theta$ over 100 replicates. The column labeled SEM refers to the conditional standard error of measurement. It is computed as the mean of the reciprocal of the squared root of the information function estimated at the MLE of $\theta$ over 100 replicates (see Hambleton and Swaminathan, 1985). The column labeled Info refers to the mean of the information function evaluated at the estimated $\theta$ over 100 replicates. The column labeled OInfo is an empirical measure of information, namely the reciprocal of the variance of the MLE at a given value of $\theta$ over 100 replicates. The next column, Bias, reports the mean difference between the MLE and the true $\theta$. The next column reports the ratio of observed information to theoretical information. Values of less than 1.0 in this column suggest that the theoretical information, Info, is an optimistic measure of precision. The next to the last column reports the ratio of observed information in a given dataset to the observed information in the dataset under no error. In table 3e the ratio is 1.0 by definition.

---

Insert Tables 3a, b, c, d, e

---

First notice, in Table 3a, that the ratio OInfo/Info is close to 1.0, as it should be when there are no errors in the parameter estimates. That is, the theoretical and empirical measures of score precision converge. This is not the case in subsequent panels, suggesting that Info is not an accurate measure of precision when the item parameter estimates are fallible. For that reason, we

will focus on the last column as a means of judging the effect of fallible item parameter estimates. That is, we will judge the effect of error in item parameter estimates by comparing empirical rather than theoretical estimates of measurement precision, using the condition with 20 items and no error as a benchmark.

In Table 3b, the ratio ranges from 1.077 to 1.265. This means that an adaptive 30 item test with a standard error of .50 in the difficulty parameter yields more precise scores than a 20 item test with error-free item parameters estimates ranging from 7.7 to 26.5%. Thus, adding 50% more items to an adaptive test, more than compensates for errors in item parameters at that error level. The equivalent results that occur when the error in item parameter rises to 1.00 are shown under the same column in Table 3d. The ratio is now below 1.0 and ranges from .562 to .988. This suggests that when the imprecision of the difficulty parameter estimates is that high, adding 50% more items is not enough to compensate for the imprecision of the item parameter estimates.

Let us now look at the results for the conditions with 20 items, which tells us what losses in precision of ability estimation we can expect due to imprecision in difficulty estimates. The last column for tables 3c and 3e range from .667 to .991 and .373 to .860 respectively. Not surprisingly, the loss of precision is higher in table 3e, which simulates a higher level of error in the difficulty parameter.

On the surface, these results suggest that the imprecision introduced into ability estimates can be compensated, to some extent, by lengthening the adaptive test. However, before we reach that conclusion, we must examine the bias in the estimation of ability caused by the error in the item parameters. That information appears under the column Bias. Bias would manifest itself as differences in a consistent direction. Such a pattern is present in tables 3d and 3e, which

simulate a high level of imprecision, where the difference is negative for values of $\theta$ below 0.0. This means that beyond a certain level of error in item difficulty parameter estimates, a bias is introduced in the maximum likelihood estimates of ability. However, as seen at lower levels of error, it appears possible to compensate for the error by lengthening an adaptive test. These results are consistent with theoretical results by Mislevy (1992), who notes that the effect of error on item parameter estimates on the estimation of ability is less at values of ability where the probability of correct response is .50, as it would be in an adaptive test where the one-parameter response model is applicable.

## Summary and Discussion

A generative approach, such as the one presented in this paper, seems necessary to take maximum advantage of the investment needed for computer-based testing, and to be responsive to the increasing expectations concerning the quality and fairness of tests. Adaptive testing is a good idea (Reckase, 1989) that insures precise measurement with fewer items and, therefore, contributes to the goal of improving quality and fairness through greater efficiency. For example, the gains in testing time made possible through an adaptive testing approach make it possible to test others aspects of proficiency. However, adaptive testing does require a major investment because a network of computers is needed to implement it. I have suggested a generative perspective, whereby the investment necessary for computer delivery is used to produce items more efficiently, and--no less importantly--to act as a validity-enhancing mechanism. The application of generative modeling was motivated by a discussion of a system designed to assess a limited aspect of writing skills, and a discussion of a preliminary model of item difficulty. Because the system assumed an open-ended response format, it could be thought of as facilitating a more authentic approach to measuring "sentence mechanics", compared to the multiple choice

version of the same task. Evidence was presented suggesting that the less precise item statistics that might be expected from a generative approach can be compensated, up to a point, if tests are administered adaptively. Thus, a generative approach to test design seems a natural way to further leverage the investment in computers for test administration.

A generative approach presupposes a deeper understanding of the response process than "conventional" test design. Developing the knowledge to effectively model item performance is necessarily intense, but it is also salutary, especially in accounting for item difficulty. This approach can also be seen as a natural arena for relevant disciplines to interact, a process that can only improve a measurement instrument. The alternative is to rely solely on an approach unenlightened by a multidisciplinary perspective, which as Shepard (1991) has documented, is what often happens, and not necessarily with positive results. At a practical level, generative modeling provides a mechanism for taking advantage of the considerable investment needed to carry out computer-based testing. The approach becomes more natural as the availability of computing resources and the accessibility of relevant knowledge, such as lexical databases, increases. Thus, it becomes possible to envision a system where the item designer submits to the test delivery system newly created items to be administered to examinees representative of the test taking population. The system then swiftly returns the data, on those newly created items, to the item designer for analysis. If nothing else, this approach to test design contributes to quality and fairness by enhancing the efficiency of the process. When coupled with an attitude that test response can be used to refine models of the assumed response process, a generative approach should lead to a more fundamental improvement in test quality and fairness by continually improving the meaning and interpretation of scores.

Acknowledgments

I'd like to acknowledge Sean Whalen's contribution to this report by writing the program implementing the algorithm in Figure 5. I'm also grateful to Bob Mislevy and John Fremmer, two anonymous reviewers, and the editor for comments that helped improve the manuscript. Finally, Susan Miller was instrumental in improving the readability of the manuscript.

References

Ackerman, T.A., & Smith, P.L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12, 117-128.

Baron, J. B. (1991). Performance assessment: Blurring the edges of assessment, curriculum, and instruction. In G. Kulm & S. M. Malcolm (Eds.), Science assessment in the service of reform (pp. 247-266). Washington, DC: American Association for the Advancement of Science.

Bejar, I. I. (1983). Achievement testing: Recent advances. Beverly Hills, CA: Sage.

Bejar, I. I. (1988). A sentence-based automated approach to the assessment of writing: A feasibility study. Machine-Mediated Learning, 2, 321-332.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. Applied Psychological Measurement, 14,(3), 237-245.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. Journal of Applied Psychology, 76, 522-532.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 323-359). Hillsdale, NJ: Erlbaum.

Bejar, I. I., Stabler, E. P., Jr., & Camp, R. (1987). Syntactic complexity and psychometric difficulty: A preliminary investigation (Research Report No. RR-87-25). Princeton, NJ: Educational Testing Service.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of hidden-figure items. Applied Psychological Measurement, 15,(2), 129-137.

Bennett, R. E. (1993). Constructed-response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 99-124). Hillsdale, NJ: Erlbaum.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27, 93-108.

Breland, H. M., & Lytle, E. G. (1990). Computer-assisted writing assessment using WordMAP(TM). Paper presented at the annual conference of the National Council on Measurement in Education, Boston, MA.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Charniak, E. (1993). Statistical language learning. Cambridge, MA: MIT Press.

Clancey, W. J. (1986). Qualitative student models. Annual Review of Computer Science, 1, 381-450.

Dwyer, C. A. (1994). Innovation and Reform: Examples from teacher assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp. 265-289). Hillsdale, NJ: Erlbaum.

Flower, L., & Hayes, J. R. (1981). The pregnant pause: An inquiry into the nature of planning. Research in the teaching of English, 15, 229-243.

Frederiksen, N. (1986). Construct validity and construct similarity: Methods for use in test development a test validation. Multivariate Behavioral Research, 21, 3-28.

Freedle, R. (1988). A semi-automatic procedure for scoring protocols resulting from a free response sentence-combining writing task. Machine Mediated Learning, 2, 309-319.

Fremer, J., & Anastasio, E. J. (1969). Computer-assisted item writing--I (Spelling items). Journal of Educational Measurement, 6,(2), 69-74.

Green, B. F. (1964). Intelligence and computer simulation. Transactions of the New York Academy of Sciences, 27,(1), 55-63.

Guttman, L. (1969). Integration of test design and analysis. Paper presented at Educational Testing Service Proceedings of the 1969 Invitational Conference on Testing Problems, Princeton, NJ.

Guttman, L. (1980). Integration of test design and analysis: Status in 1979: Measuring achievement. In W. B. Schrader (Ed.), Proceedings of the 1979 Educational Testing Service Invitational Conference. San Francisco: Jossey-Bass.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Nijhoff.

Hively, W. (1974). Introduction to domain-reference testing. Educational Technology, 14, 5-9.

Hornke, L. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. Applied Psychological Measurement, 10, (4), 369-380.

Irvine, S. H., Dunn, P. L., & Anderson, J. D. (1989). Towards a theory of algorithm-determined cognitive test construction. Devon, UK: Polytechnic South West.

Jones, D. H., Jin, Z. (1994). Optimal Sequential Designs for On-Line Item Estimation. Psychometrika, 59, (1), 59-75.

Kaplan, R. M. (1992). Using a trainable pattern-directed computer program to score natural language item responses (Research Report No. 91-31). Princeton, NJ: Educational Testing Service.

Kaplan, R. M., & Burstein, J. C. (Eds.). (1994). Proceedings of the Educational Testing Service conference on natural language processing techniques and technology in assessment and education. Princeton, NJ: Educational Testing Service.

Kyllonen, P. C. (1990, April). Taxonomies of cognitive abilities. Paper presented at the American Educational Research Association Meeting, Boston, MA.

Lewis, C. (1985). Estimating item abilities with imperfectly known item response functions. Paper presented at the Annual Meeting of the Psychometric Society, Nashville, TN.

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

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lytle, E. G. (1986). WordMAP(TM)-1 reference manual. Panaca, NV: Linguistic Technologies.

Martin, J. T., & van Lehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), Cognitively diagnostic assessment (pp. 141-165). Hillsdale, NJ: Erlbaum.

Meisner, R. M., Luecht, R. M., & Reckase, M. D. (1993). The comparability of the statistical characteristics of test items generated by computer algorithms (ACT Research Report Series No. 93-9). Iowa City, IA: The American College Testing Program.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). New York: American Council on Education.

Mislevy, R. J. (1992). The variance of Rasch ability estimates for partially-known item parameters (Research Report No. 92-9-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), Cognitively diagnostic assessment (pp. 43-71). Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. Journal of Educational Measurement, 30,(1), 55-78.

Mislevy, R. J., Wingersky, M, Sheehan, K. M. (1994). Dealing with uncertainty about item parameters: Expected response functions. (Research Report No. RR-94-28-ONR). Princeton, NJ: Educational Testing Service.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. Review of Educational Research, 64,(4), 575-603.

Nitko, A. J., & Lane, S. (1990, August). Solving problems is not enough: Assessing and diagnosing the ways in which students organize statistical concepts. Paper presented at the Third International Conference on Teaching Statistics, Dunedin, New Zealand.

Oltman, P. K., Bejar, I. I., & Kim, S. H. (1993). An approach to automated scoring of architectural designs. In U. Flemming & S. van Wyk (Eds.), CAAD Futures 93 (pp. 215-224). New York: North-Holland.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. Educational Measurement: Issues and Practice, 8,(3), 11-15.

Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. New York: Academic.

Sheehan, K. M. (1996). A tree-based approach to proficiency scaling. Princeton, NJ: Educational Testing Service. Manuscript in preparation.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. Educational Researcher, 20,(7), 2-16.

Snow, R. E., & Mandinach, E. B. (1991). Integrating assessment and instruction: A research and development agenda (Research Report No. 91-8). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K., & Tatsuoka, M. M. (1992). A psychometrically sound cognitive diagnostic model: Effect of remediation as empirical validity (Research Report No. RR-92-38-ONR). Princeton, NJ: Educational Testing Service.

Uttal, W. R., Rogers, M., Hieronymous, R., & Pasich, T. (1970). Generative computer-assisted instruction in analytical geometry. Newburyport, MA: Entelek.

van Lehn, K. (1988). Student modeling. In J. J. Richardson & M. C. Polson (Eds.), Intelligent tutoring systems (pp. 55-78). Hillsdale, NJ: Erlbaum.

Vapnik, V. N. (1995). The nature of statistical learning theory. New York: Springer-Verlag.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. Journal of Educational Statistics, 12, 339-368.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Erlbaum.

Ward, W. C. (1984). Using microcomputers to administer tests. Educational Measurement: Issues and Practice, 3,(2),16-20.

Weiss, D. J. (1978). Proceedings of the 1977 Computerized Adaptive Testing Conference. In D. J. Weiss (Ed.), Comperized Adaptive Testing Conference. University of Minnesota.

Wolfe, J. H., & Larson, G. E. (1990). Generative adaptive testing with digit span items. Testing Systems Department, Navy Personnel Research and Development Center, San Diego, CA.

TABLE 1

Regression results for TSWE data

| Model | Rating | Depth | Depth of error | Intercept | $R^2$ | $R^2_{adj}$ |
|---|---|---|---|---|---|---|
| Delta = F(Rating) | .45 | | | 4.93 | .09 | .07 |
| Delta = f(Rating, Depth) | .34 | .45 | | 1.68 | .21 | .17 |
| Delta = (Depth) | | .51 | | 4.35 | .16 | .14 |
| Delta = f(Rating, Depth, Depth of Error) | .50 | .58 | -.39 | .77 | .31 | .25 |

TABLE 2

Design for Monte Carlo study

| | Number of items in adaptive test | |
|---|---|---|
| Error level | 20 | 30 |
| .50 | 20 items in the adaptive test and level of error at .50 | 30 items in the adaptive test and level of error at .50 |
| 1.00 | 20 items in the adaptive test and level of error at 1.0 | 30 items in the adaptive test and level of error at 1.0 |

TABLE 3A

Results under no error in the item parameters

| $\theta$ | Mean($\hat{\theta}$) | SEM Mean($1/\sqrt{I(\hat{\theta})}$) | Info Mean $I(\hat{\theta})$ | OInfo $1/\mathrm{Var}(\hat{\theta})$ | Bias Mean($\hat{\theta}-\theta$) | Oinfo/ Info | Ratio of Oinfo to observed information under no error |
|---|---|---|---|---|---|---|---|
| -3 | -2.986 | 0.321 | 11.736 | 9.683 | -0.015 | 0.825 | 1.0 |
| -2 | -2.013 | 0.285 | 12.345 | 12.336 | 0.013 | 0.999 | 1.0 |
| -1 | -0.996 | 0.284 | 12.680 | 12.401 | -0.004 | 0.978 | 1.0 |
| 0 | 0.006 | 0.283 | 12.994 | 12.527 | -0.006 | 0.964 | 1.0 |
| 1 | 0.992 | 0.265 | 12.723 | 14.229 | 0.008 | 1.118 | 1.0 |
| 2 | 2.027 | 0.328 | 12.252 | 9.300 | -0.027 | 0.759 | 1.0 |
| 3 | 3.000 | 0.291 | 11.838 | 11.827 | 0.001 | 0.999 | 1.0 |

TABLE 3B

Results under the condition with 30 items in the adaptive test and a .50 level of error in the difficulty item parameter

| $\theta$ | Mean($\hat{\theta}$) | SEM Mean($1/\sqrt{I(\hat{\theta})}$) | Info Mean $I(\hat{\theta})$ | OInfo $1/\mathrm{Var}(\hat{\theta})$ | Bias Mean($\hat{\theta}-\theta$) | Oinfo/ Info | Ratio of Oinfo to observed information under no error |
|---|---|---|---|---|---|---|---|
| -3 | -2.979 | 0.295 | 18.725 | 11.485 | -0.021 | 0.613 | 1.186 |
| -2 | -2.007 | 0.267 | 19.043 | 14.043 | 0.006 | 0.737 | 1.138 |
| -1 | -1.040 | 0.260 | 19.616 | 14.753 | 0.040 | 0.752 | 1.190 |
| 0 | -0.046 | 0.251 | 19.799 | 15.844 | 0.046 | 0.800 | 1.265 |
| 1 | 1.014 | 0.250 | 19.784 | 15.976 | -0.014 | 0.808 | 1.123 |
| 2 | 1.967 | 0.272 | 18.984 | 13.540 | 0.033 | 0.713 | 1.456 |
| 3 | 3.015 | 0.280 | 18.519 | 12.742 | -0.015 | 0.688 | 1.077 |

TABLE 3C

Results under the condition with 20 items in the adaptive test and a .50 level of error in the difficulty item parameter

| $\theta$ | Mean($\hat{\theta}$) | SEM Mean($1/\sqrt{I(\hat{\theta})}$) | Info Mean $I(\hat{\theta})$ | OInfo $1/\mathrm{Var}(\hat{\theta})$ | Bias Mean($\hat{\theta}-\theta$) | Oinfo/ Info | Ratio of Oinfo to observed information under no error |
|---|---|---|---|---|---|---|---|
| -3 | -2.974 | 0.347 | 11.595 | 8.329 | -0.026 | 0.718 | 0.860 |
| -2 | -2.033 | 0.335 | 12.269 | 8.895 | 0.033 | 0.725 | 0.721 |
| -1 | -1.027 | 0.307 | 12.696 | 10.589 | 0.027 | 0.834 | 0.854 |
| 0 | 0.054 | 0.297 | 12.866 | 11.337 | -0.054 | 0.881 | 0.905 |
| 1 | 0.919 | 0.325 | 12.808 | 9.497 | 0.081 | 0.741 | 0.667 |
| 2 | 1.957 | 0.329 | 12.232 | 9.216 | 0.043 | 0.753 | 0.991 |
| 3 | 3.011 | 0.318 | 11.792 | 9.914 | -0.011 | 0.841 | 0.838 |

TABLE 3D

Results under the condition with 30 items in the adaptive test and a 1.00 level of error in the difficulty item parameter

| $\theta$ | Mean($\hat{\theta}$) | SEM Mean($1/\sqrt{I(\hat{\theta})}$) | Info Mean $I(\hat{\theta})$ | OInfo $1/\mathrm{Var}(\hat{\theta})$ | Bias Mean($\hat{\theta}-\theta$) | Oinfo/ Info | Ratio of Oinfo to observed information under no error |
|---|---|---|---|---|---|---|---|
| -3 | -2.892 | 0.343 | 18.397 | 8.486 | -0.108 | 0.461 | 0.876 |
| -2 | -1.951 | 0.309 | 19.031 | 10.465 | -0.050 | 0.550 | 0.848 |
| -1 | -0.963 | 0.363 | 19.336 | 7.609 | -0.037 | 0.394 | 0.614 |
| 0 | -0.025 | 0.305 | 19.867 | 10.738 | 0.025 | 0.540 | 0.857 |
| 1 | 0.928 | 0.354 | 19.517 | 8.002 | 0.073 | 0.410 | 0.562 |
| 2 | 1.954 | 0.330 | 19.148 | 9.192 | 0.047 | 0.480 | 0.988 |
| 3 | 2.914 | 0.325 | 18.399 | 9.497 | 0.087 | 0.516 | 0.803 |

TABLE 3E

Results under the condition with 20 items in the adaptive test and a 1.00 level of error in the difficulty item parameter

| $\theta$ | Mean($\hat{\theta}$) | SEM Mean($1/\sqrt{I(\hat{\theta})}$) | Info Mean $I(\hat{\theta})$ | OInfo $1/\mathrm{Var}(\hat{\theta})$ | Bias Mean($\hat{\theta}-\theta$) | Oinfo/ Info | Ratio of Oinfo to observed information under no error |
|---|---|---|---|---|---|---|---|
| -3 | -2.865 | 0.346 | 11.536 | 8.329 | -0.136 | 0.722 | 0.860 |
| -2 | -1.919 | 0.478 | 12.148 | 4.385 | -0.081 | 0.361 | 0.355 |
| -1 | -0.985 | 0.408 | 12.609 | 6.012 | -0.016 | 0.477 | 0.485 |
| 0 | -0.015 | 0.415 | 12.711 | 5.795 | 0.015 | 0.456 | 0.463 |
| 1 | 0.903 | 0.434 | 12.658 | 5.305 | 0.097 | 0.419 | 0.373 |
| 2 | 1.914 | 0.444 | 12.084 | 5.062 | 0.086 | 0.419 | 0.544 |
| 3 | 2.932 | 0.420 | 11.601 | 5.666 | 0.068 | 0.488 | 0.479 |

## Figure Captions

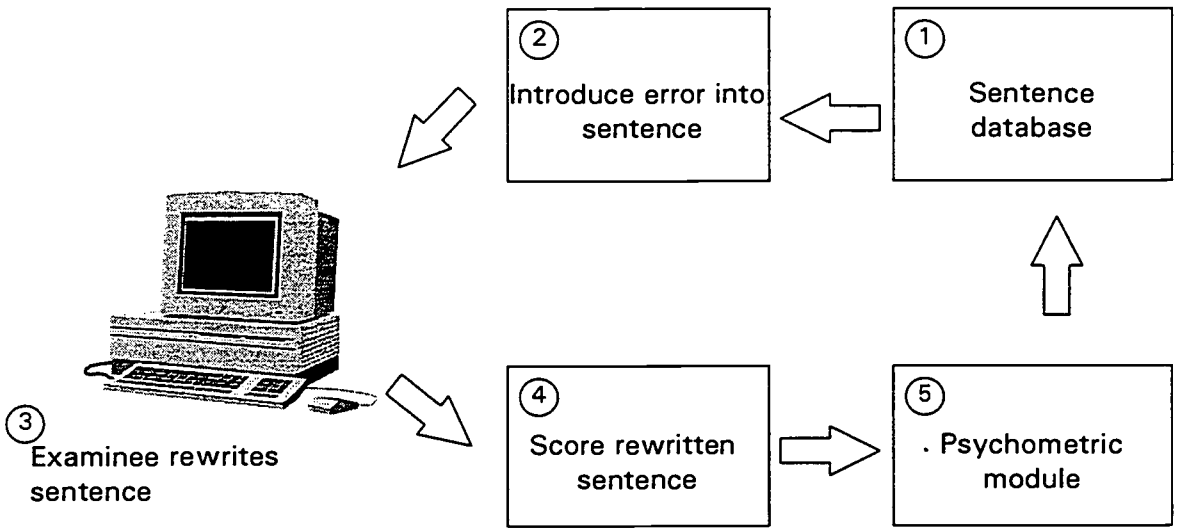
Figure 1. Components of a free-response sentence rewriting test.

Figure 2. Possible outcomes in a free-response sentence rewriting test.

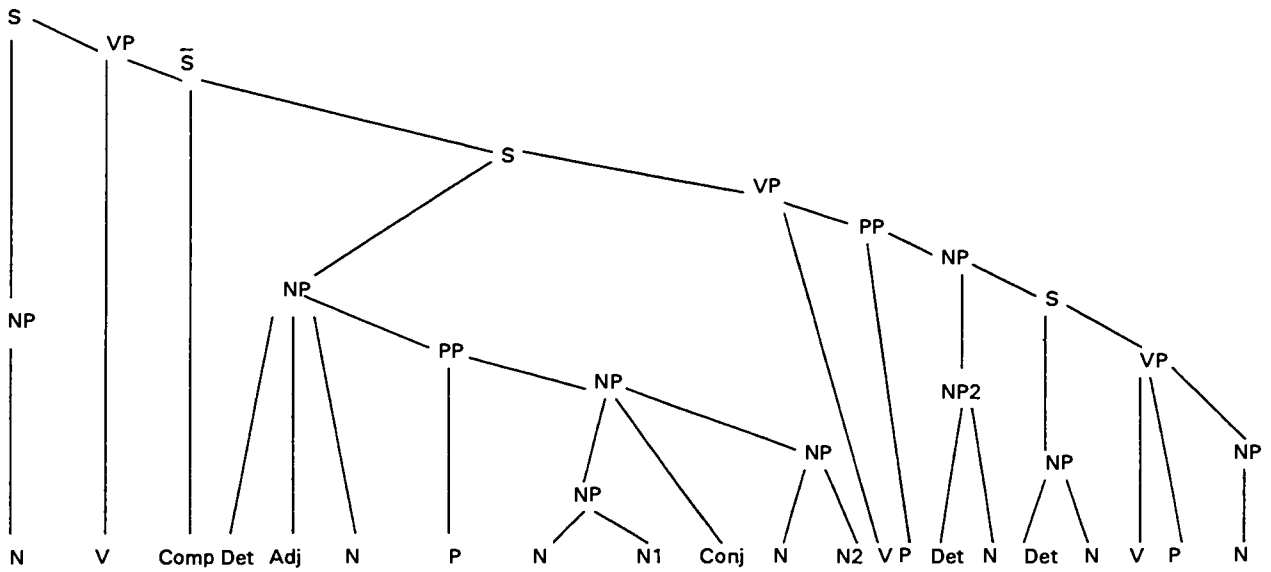
Figure 3. Sample TSWE Sentence of medium difficulty.

Figure 4. Sample TSWE sentence of low difficulty.

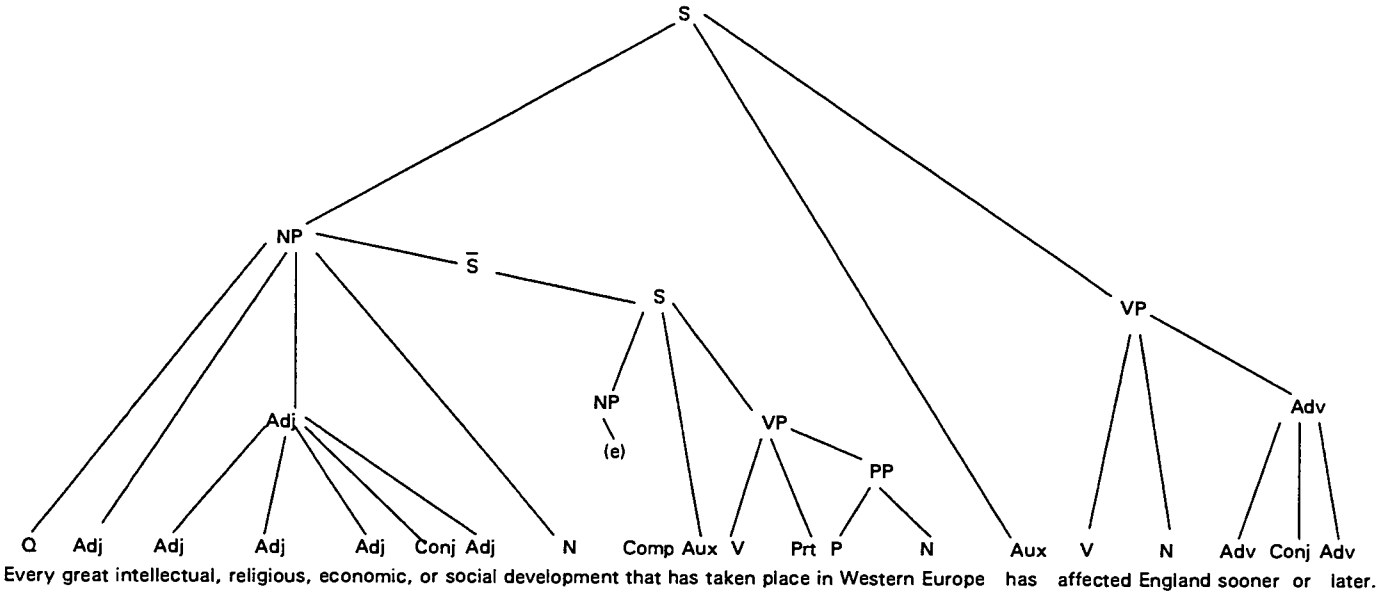
Figure 5. Monte Carlo Design for simulation to assess effect of error in item parameters on ability estimates.


2
Introduce error into sentence
1
Sentence database
3
Examinee rewrites sentence
4
Score rewritten sentence
5
Psychometric module

| Examinee rewrites? | Original sentence: without error | Original sentence: with error |
|---|---|---|
| yes | Is rewritten sentence semantically equivalent? | Maximum credit given if error correctly removed and none introduced |
| no | Examinee given maximum credit | Note type of error not recognized by examinee |




Mark suggested that the major difference between Hogarth's etchings and Rembrandt's is in the view each artist had of humanity.


S
NP
S̄
S
VP
NP
(e)
Adj
VP
Adv
PP
Q Adj Adj Adj Adj Conj Adj N Comp Aux V Prt P N Aux V N Adv Conj Adv
Every great intellectual, religious, economic, or social development that has taken place in Western Europe has affected England sooner or later.

Monte Carlo procedure

For each θ from -3.0 to 3.0 by increment of 1.0
For each replicate from 1 to 100
Present item of presumed difficulty 0.0
*(Where the true difficulty is (0.0 - error) where error is normally distributed with mean 0 and standard deviation controlled by the experimenter.)*
Repeat
Compute probability of correct answer using the true difficulty and theta
If correct, next item is higher in difficulty by .2
else next item's difficulty is lower by a constant, -.2
Until a mixed response pattern is observed
*(That is, a response vector with a mixture of correct and incorrect responses is obtained so that it is possible to compute the maximum likelihood estimate.)*
Compute MLE ability estimate $\hat{\theta}$ *(based on presumed item difficulty estimates)*

*(Now that we have a preliminary estimate of ability proceed to administer the rest of the test adaptively. Instead of increasing or decreasing the difficulty of the next item by a constant as above the next items difficulty is set to the current estimate of ability, which presupposes an infinite item pool, as indicated in the text.)*
Repeat
Administer an item of presumed difficulty = $\hat{\theta}$
Compute true difficulty as ($\hat{\theta}$ - error)
Compute probability of correct answer using the true difficulty
Compute revised $\hat{\theta}$
Until a total of n items have been given
Next replicate
Next θ

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

# NOTICE

## REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").