

J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 2, Number 3 · November 2003

# A Feasibility Study of On-the-Fly Item Generation in Adaptive Testing

Isaac I. Bejar, René R. Lawless,  
Mary E. Morley, Michael E. Wagner,  
Randy E. Bennett, and Javier Revuelta

[www.jtla.org](http://www.jtla.org)

A publication of the Technology and Assessment Study Collaborative  
Caroline A. & Peter S. Lynch School of Education, Boston College

## **A Feasibility Study of On-the-Fly Item Generation in Adaptive Testing**

Isaac I. Bejar, René R. Lawless, Mary E. Morley, Michael E. Wagner, Randy E. Bennett,  
and Javier Revuelta

Editor: Michael Russell  
russelmh@bc.edu  
Technology and Assessment Study Collaborative  
Lynch School of Education, Boston College  
Chestnut Hill, MA 02467

Copy Editor: Kathleen O'Connor  
Design and Layout: Thomas Hoffmann

JTLa is a free on-line journal, published by the Technology and Assessment Study Collaborative,  
Caroline A. & Peter S. Lynch School of Education, Boston College.

Copyright ©2003 by the Journal of Technology, Learning, and Assessment (ISSN 1540-2525).  
Permission is hereby granted to copy any article provided that the Journal of Technology, Learning,  
and Assessment is credited and copies are not sold.

### **Preferred citation:**

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003).  
A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology,  
Learning, and Assessment*, 2(3). Available from <http://www.jtla.org>

### **Abstract:**

The goal of this study was to assess the feasibility of an approach to adaptive testing using item models based on the quantitative section of the Graduate Record Examination (GRE) test. An item model is a means of generating items that are isomorphic, that is, equivalent in content and equivalent psychometrically. Item models, like items, are calibrated by fitting an IRT response model. The resulting set of parameter estimates is imputed to all the items generated by the model. An on-the-fly adaptive test tailors the test to examinees and presents instances of an item model rather than independently developed items. A simulation study was designed to explore the effect an on-the-fly test design would have on score precision and bias as a function of the level of item model isomorphism. In addition, two types of experimental tests were administered – an experimental, on-the-fly, adaptive quantitative-reasoning test as well as an experimental quantitative-reasoning linear test consisting of items based on item models. Results of the simulation study showed that under different levels of isomorphism, there was no bias, but precision of measurement was eroded at some level. However, the comparison of experimental, on-the-fly adaptive test scores with the GRE test scores closely matched the test-retest correlation observed under operational conditions. Analyses of item functioning on the experimental linear test forms suggested that a high level of isomorphism across items within models was achieved. The current study provides a promising first step toward significant cost reduction and theoretical improvement in test creation methodology for educational assessment.

# A Feasibility Study of On-the-Fly Item Generation in Adaptive Testing

## Introduction

Item modeling (LaDuca, Staples, Templeton, & Holzman, 1986; Bejar, 1996), an example of generative testing (e.g., Bejar, 1993), is an approach to test development that is construct-driven and potentially validity-enhancing while providing many significant practical advantages, including improved cost effectiveness over standard item writing. Item modeling can be thought of as a procedure for instantiating isomorphic items – items that contain comparable content and are exchangeable psychometrically. The feasibility of the approach rests in part on whether the instances of item models, that is, the items produced by an item model, are sufficiently isomorphic. A further feasibility criterion is whether scores based on tests composed of items produced by an item model have an adequate level of score precision. We studied these questions regarding the feasibility of item modeling by means of both a simulation study and an empirical study based on the Graduate Record Examination (GRE).

We view item modeling as a construct-driven and validity-enhancing approach because it entails a more thorough understanding of the goals of the assessment and the application of pertinent psychological research to the design of test content than the current mode of item development. That is, item models set the expectation for the behavior of the instances produced by a given model (e.g., difficulty and discrimination) and those expectations can be verified upon administration of a test consisting of those instances, thus providing an opportunity to refine our understanding of the construct and supporting psychological principles.

In addition to their role as a validity-enhancing mechanism, item models may have practical advantages. In particular, manual item production is a labor-intensive process that treats each item as an isolated entity to be individually authored, reviewed, and formatted, regardless of how similar it may be to other items. An item modeling approach automates many of the details of producing items (instances) once the item model has been formulated and calibrated. Moreover, by representing a class of items abstractly an item model can be leveraged in various ways. For example, an item model can be designed in such a way that instances can be rendered in a language of choice, at least in certain domains like mathematics and deductive reasoning (Bejar, Fairon, & Williamson, 2002). In addition, item models can be extended in such a way that instance-specific feedback or tutoring can be built into the item model.

The roots of item modeling can be traced to criterion-referenced testing (Hively, Patterson, & Page, 1968; Hively, 1974) and computer-assisted instruction (e.g., Uttal, Rogers, Hieronymous, & Pasich, 1969). Hively's work on criterion-referenced testing specifically emphasized automated generation. In his approach, a domain is defined "in terms of operationally stated rules called *item form* rules, which allow for an explicit description of the complete set of items that could be written" (Macready, 1983, p. 149). This early research recognized the need to control both homogeneity and difficulty. At the time, however, accountings of difficulty were rare because the cognitive theories needed to psychometrically model items were not yet available.

During the same period, Uttal et al. (1969) used the term *generative instruction* to describe an alternative to the machine learning efforts of the 1960's, which were based on Skinnerian principles. Skinner (1954, 1958, & 1961) 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 cognitive perspective underlying generative instruction was elaborated by Brown and Burton (1978), among others, into a branch of cognitive science known as *intelligent tutoring* that relies on a detailed, dynamically updated description – or *student model* – as the basis for presenting instruction (e.g., Clancey, 1986; van Lehn, 1988; Martin & van Lehn, 1995; Mislevy, 1995). Student modeling is now an integral part of the evidence-centered design assessment framework (e.g., Mislevy, Steinberg, & Almond, 2002). As such, there is a strong conceptual linkage between item models, assessment, and instruction (e.g., Bejar, 1993).

The cognitive perspective that first started in an instructional context now prevails in psychometric modeling as well. For example, item-difficulty modeling is now a common method for gathering evidence related to what Embretson (1983) has called *construct representation*, a key aspect of test validity concerned with understanding the cognitive mechanisms related to the item solution and item features that call on these mechanisms. The utility and feasibility of this perspective can be judged by the variety of domains in which it has been successfully applied. These domains include ability and achievement testing, as well as the measurement of complex skills, such as troubleshooting, clinical diagnoses, and pedagogical skills. A growing number of projects demonstrate the feasibility of the generative approach (e.g., Bejar, 1990, 1993, 2002; Bejar & Braun, 1999; Bejar & Yocom, 1991; Embretson, 1999; Hornke & Habon, 1986; Irvine, Dunn, & Anderson, 1990; Kenney, 1997; Meisner, Luecht, & Reckase, 1993; Singley & Bennett, 2002; Wolfe & Larson, 1990), as well as the feasibility of the approach in different contexts (Irvine & Kyllonen, 2002). The pioneering work of Bejar (1986), Hornke & Habon (1986), and Irvine et al. (1990) are especially noteworthy because they provided a conceptual and experimental basis for subsequent work.

At some level, item modeling is a simple extension of the current approach to test development where the "item models" are not explicitly stated, that is, they reside in the test developer's head and yield a single "instance". By contrast item models are explicitly represented as a set of variables and yield multiple instances

intended to be isomorphic with respect to difficulty and other parameters. At least two approaches to modeling difficulty within a generative approach seem feasible: *strong theory* versus *weak theory*. Implied in this discussion of strong theory versus weak theory is that the modeling of difficulty needs to be tied to a psychometric response model (e.g., Embretson, 1999). Therefore, feasibility needs to be judged not just by the adequacy of the statistical fit of the psychometric model to the data but also the theoretical fit of the observed data to the predictions entailed by each item model, especially whether the instances generated by a model can be considered exchangeable given the purposes of the assessment. For example, a practice test would be judged differently than a high stakes test because the increased score imprecision that results from lack of isomorphism may be acceptable in a practice test.

Strong theory relies on the psychological principles underlying domain performance to finely control difficulty, either among the models that compose a test or among the instances that a model produces. In the former case, each model is written to generate items that are isomorphic. Psychological principles are used to create variation in difficulty among, rather than within, models and to predict the response parameters for each model (e.g., Embretson's [1993, 1999] work with matrix completion tasks). In the latter case, principles are employed to create a single model that generates calibrated items that vary widely in difficulty. For example, Bejar (1990) relied on the psychology of mental rotation to generate instances and to estimate item parameters. Strong theory works well in narrow domains where cognitive analysis is feasible and where well-developed theory is more likely to exist.

In broader domains, strong theory may not be available. In these domains, weak theory may be applicable. Weak theory begins with a set of calibrated test items that cover the domain of interest in terms of difficulty and content. Each item serves as the basis for an item model. The models themselves are written using best-practice guidelines, as opposed to psychological principles, so that each model generates isomorphs. In this study, we use weak theory to calibrate an item model by means of a 3-parameter item response theory (IRT) model (Lord, 1980). Moreover, we impute the model parameters to all instances of the model. In other words, we estimate the difficulty, discrimination, and a guessing parameter for each model, and apply the estimates to each instance of a model. Therefore, the emphasis is on producing items that are well described by a single set of parameters, that is, isomorphism.

Independent of whether we are operating under strong theory or weak theory, the specifics of parameter estimation for the assumed psychometric response model need to be considered. In particular, it is important to distinguish the case in which a model needs to be calibrated from scratch versus the case in which previously calibrated items can be thought of as instances of a model. In calibrating from scratch one might treat the randomly assigned instances of a model as if they were the same "item." Then, a standard parameter-estimation program could be used to fit the responses for different instances to a single item-characteristic

curve (ICC). The fit of the resulting estimated ICC would depend, in part, on the level of isomorphism – that is, it would depend on the degree of variation among item parameters of the different instances that were treated as if they were a single item. A major shortcoming of this approach is that the variability that may exist among instances of a model is not captured explicitly. Therefore, a more satisfying approach is to formulate a statistical model whereby variability among instances is captured along with “base” parameter estimates that characterize the class of items from a given item model.<sup>1</sup>

In the second case – in which previously calibrated items can be thought of as instances of a model – we need to distinguish whether one or more calibrated items are available. In either case, the goal is to estimate parameters for the model from the available data. In this study we use the expected response function method for the case in which we use the parameter estimates of a single item as the basis for estimating the parameters for the item model. The case in which multiple existing item parameter estimates are available remains to be explored.

In short, an approach to test development based on item modeling can potentially provide a mechanism to enhance the validity of the scores through corroboration of theoretical predictions about the psychometric behavior of instances of an item model. The approach also potentially has many practical advantages. The goal of this investigation was to assess the feasibility of item modeling in the context of the GRE, specifically, studying the feasibility of adaptive testing (see e.g., Wainer, 2000) based on item models. The investigation involved two studies – a simulation study and a field study. The simulation study was conducted to assess theoretically the impact on score precision of lack of isomorphism among item model instances in an adaptive test. The field study involved two components, the first an experimental adaptive test that administered items generated from item models during the administration of the test, that is, on the fly, as well as traditional items. This test was administered to determine if it could produce scores equivalent to those from the GRE General Test. The second component of the field study was an experimental linear test, comprised solely of items generated from item models. The linear test was administered to determine how similarly items from a given model functioned empirically.

## Item Modeling Procedures

In the following section we describe in more detail the nature of item modeling in the context of this study as well as the specific item models used. We then describe the approach to psychometric parameter estimation based on expected response functions that was used both in the simulation study as well as in the field study.

### Item Modeling

Two major tasks involved in the creation of an adaptive test are item pool construction and item calibration. In the current study, we used a specific item pool that was released with PowerPrep®, a test preparation program distributed by Educational Testing Service for the GRE General Test. A subset of 147 items from this 408-item pool was used to develop item models. The items with the highest predicted exposure, that is, those most likely to be administered, were chosen to be converted to item models to ensure that any given student would be responding to instances from as many item models as possible. (See Mills and Steffen, 2000, for a description of the GRE adaptive design.)

Some items were excluded from modeling for several reasons. First, data-interpretation sets were not modeled. One reason why these sets were not modeled is because significant effort would have been required to make the instances appear credible. Also, neither quantitative-comparison items nor problem-solving items that had low predicted exposure were modeled because most likely they would not be chosen by the item selection algorithm. Finally, items that would lead to models that would produce only a few instances were excluded from modeling. But, in order to illustrate the feasibility of producing items on-the-fly with dynamically-generated graphical material, we did model discrete items with figures. Item models were reviewed by experienced test-development staff, who manually generated possible instances to evaluate their equivalence. They evaluated models in terms of the content variability of the instances and their subjectively estimated difficulty. Models that did not strike a balance of some diversity in content variability and evidenced more than a minimal spread in difficulty were excluded.

The quantitative-comparison item in Figure 1 requires the examinee to use the information in the stem to determine whether the quantity in Column A or the quantity in Column B is greater, equal, or if the relationship cannot be determined based on the information given. This item was used to derive an item model allowing for constrained variations. This item model, detailed in Figure 2, identifies integer (numeric) variables by **I** and string variables by **S**. For example, in the stem the integer variable, 30, becomes **I1** which can be substituted with an integer between 30 and 90 in increments of 30 (i.e., 30, 60, 90) and the scale variables, centimeters and kilometers, become **S1.1** and **S1.2**, respectively, and can represent “centimeters or inches” or “kilometers or miles,” respectively. The model allows for the actual distance in Column A to vary, with the digit in the thousands place, indicated as **I2**, having constraints of 2 or 4. The distance in Column A is pre-determined by the variable in the stem – **S12**. The value of Column B (**I3 S1.3**) is

stipulated to be less than the value of Column A (**I2,000 S1.2**) with the Column B value (**I3**) varying dependent on the integer selected for Column A. The integer for Column B (**I3**) is  $\{[(\mathbf{I2} \times 1000 / \mathbf{I1}) / 10] \times 10\}$  – basically, the centimeter equivalent of Column A with the number rounded down to the nearest 10, as a result of integer division, thus, consistently obtaining a value less than that in Column A. The variables **I4** and **I5** do not appear in the problem, but rather are needed to specify constraints on the Column B variable, **I3**. The string variable in Column B, **S1.3**, is simply the plural of the string variable in the stem (**S1.1**). Sample instances resulting from this item model (Figure 2) are shown in Figure 3.

**Figure 1**

On a map drawn to scale, 1 centimeter represents 30 kilometers.

<u>Column A</u>	<u>Column B</u>
The distance on the map between two cities that are actually 2,000 kilometers apart	60 centimeters

- ☐ The quantity in Column A is greater.
- ☐ The quantity in Column B is greater.
- ☐ The two quantities are equal.
- ☐ The relationship cannot be determined from the information given.

Figure 1. Sample textual quantitative-comparison item.



Figure 2

Quantitative-Comparison Model

Stem

On a map drawn to scale, 1 **S1.1** represents **I1** **S1.2**.

Column A value

The distance on the map between two cities that are actually **I2**,000 **S1.2** apart

Column B value

**I3** **S1.3**

Variables

**S1.1** Range: "inch" or "centimeter"  
**S1.2** Range: "miles" or "kilometers"  
**S1.3** Range: "inches" or "centimeters"  
**I1** Value range: 30–90 by 30  
**I2** Value range: 2 or 4  
**I3**  
**I4**  
**I5**

Constraints

**I4** = **I2** \* 1000/**I1**  
**I5** = **I4**/**I0**  
**I3** = **I5** \* 10

Key

A

- 1 String variable **S1.1** varies according to whether the map scale is in inches or centimeters.
- 2 **I1** is a numeric variable constrained to take on integer values between 30 and 90 in increments of 30.
- 3 **S1.2** is a string variable for the units of distance – either miles or kilometers.
- 4 **I2** is a numeric variable constrained to take on integer values 2 or 4.
- 5 **I3** is an integer variable that is calculated to be slightly less than the value of Column A.
- 6 **S1.3** is the plural of **S1.1**.
- 7 **I4** and **I5** are integer variables.

Figure 2. Quantitative-comparison item model for item depicted in Figure 1.

**Figure 3**

1. On a map drawn to scale, 1 centimeter represents 30 kilometers.

The distance on the map between two cities that are actually 4,000 kilometers apart 130 centimeters

2. On a map drawn to scale, 1 inch represents 60 miles.

The distance on the map between two cities that are actually 2,000 miles apart 30 inches

3. On a map drawn to scale, 1 inch represents 30 miles.

The distance on the map between two cities that are actually 2,000 miles apart 60 inches

4. On a map drawn to scale, 1 centimeter represents 90 kilometers.

The distance on the map between two cities that are actually 4,000 kilometers apart 40 centimeters

Figure 3. Sample isomorphs derived from model depicted in Figure 2.

## Calibrating the Item Models Using the Expected Response Function

Because item models are meant to produce isomorphic instances, our approach is to calibrate an item model and then impute the model calibration to all instances of the model. For the present study, we modified existing parameter estimates for the items giving rise to each model, but under assumptions of different levels of lack of isomorphism. The procedure we used for this purpose was the expected response function.

The expected response function (ERF) is based on the work of Charles Lewis, as implemented by Mislevy, Wingersky, and Sheehan (1994). ERF is a procedure for attenuating parameter estimates as a function of the uncertainty in them. It is common for item parameters to be used in estimating ability as if they were known, without any provision for the uncertainty associated with the estimates. Such a practice overstates the precision of ability estimates. The ERF methodology enables us to attenuate parameter estimates as a function of that uncertainty and as a result more accurately measure score precision. The methodology is directly applicable in the present context in which, in addition to the usual uncertainty, instances from a given item model will vary somewhat in their psychometric characteristics.<sup>2</sup>

Assuming a 3-parameter logistic model, computing the ERF is a matter of averaging the ICCs over all instances of the item model. That is, averaging the response probabilities at selected values of  $\theta$ . The resulting vector of averaged probabilities is then fitted to the closest 3PL curve. To the extent that there is lack of isomorphism, the ERF will tend to have a shallower slope than item model instances. A shallower slope translates into a loss in precision of measurement because a shallower slope implies a lower degree of item discrimination. Conversely, to the extent that isomorphism holds, the ICCs will coincide with the slope of the ERF, and there will not be any loss in precision in the ability estimates due to lack of isomorphism.

Computation of the ERF requires, for each item model, estimates of both  $\beta$ , the vector of item parameters corresponding to discrimination, difficulty, and guessing, and  $\Sigma$ , the variance-covariance matrix of item parameter estimates. Using these estimates, the computational procedure performs multiple draws from a multivariate normal distribution with  $\Sigma$  as its covariance matrix and  $\beta$  as the mean vector. (To this end, the  $a$  and  $c$  parameters are transformed to approximate normality.) Such estimates could be obtained by administering instances of an item model to equivalent examinee samples and computing the variance-covariance matrix from the resulting estimates. Because we could not collect the data to derive these estimates empirically for  $\beta$ , we instead used the existing parameter estimates for the 147 items that gave rise to the 147 item models. For  $\Sigma$ , we located repeated calibrations of the same items, the “linking sets” used to scale pretest items. The logic of this choice is that the resulting variability is what would be expected under complete isomorphism – that is, when the item is the same and the only difference is in the calibrations over repeated administrations. For each

of these linking items, we computed the variance-covariance matrix among the parameter estimates of each item. After examining the matrices, we selected one matrix at each of three levels of variability in  $b$ , which we labeled best ( $\Sigma_1$ ), medium ( $\Sigma_2$ ), and worst-case ( $\Sigma_3$ ) scenarios. The matrices, without transforming  $a$  and  $c$  to normality, were selected for purposes of computing the ERF. The diagonal of these matrices shows the variability of these parameters and are as follows:

$$\Sigma_1 = \begin{array}{c} \begin{array}{ccc} & a & b & c \\ a & .003 & .002 & .001 \\ b & & .023 & .011 \\ c & & & .006 \end{array} \end{array}$$

$$\Sigma_2 = \begin{array}{c} \begin{array}{ccc} & a & b & c \\ a & .012 & .051 & .012 \\ b & & .237 & .054 \\ c & & & .014 \end{array} \end{array}$$

$$\Sigma_3 = \begin{array}{c} \begin{array}{ccc} & a & b & c \\ a & .015 & .067 & .016 \\ b & & .337 & .081 \\ c & & & .020 \end{array} \end{array}$$

For each of the 147 item models, we next computed three ERFs, one for each scenario. For any given model,  $\beta$  was set to the values of  $a$ ,  $b$ , and  $c$  associated with the item that gave rise to the model in the first place. However, the same covariance matrix was used for all 147 estimates for any given scenario (i.e., best-, medium-, and worst-case scenarios). In a more operational situation, we would associate a different matrix to each item model. Figure 4 shows the relationship between the original  $a$  parameter estimates and the attenuated estimates – that is, the estimates computed by the ERF procedure, assuming the worst-case scenario. As expected, the estimates are attenuated, indicating that some information will be lost as a result of the lack of isomorphism. The  $c$  estimates were estimated to be higher after the application of the ERF procedure while the  $b$  estimates changed only very slightly as a result of the application of the ERF procedure.

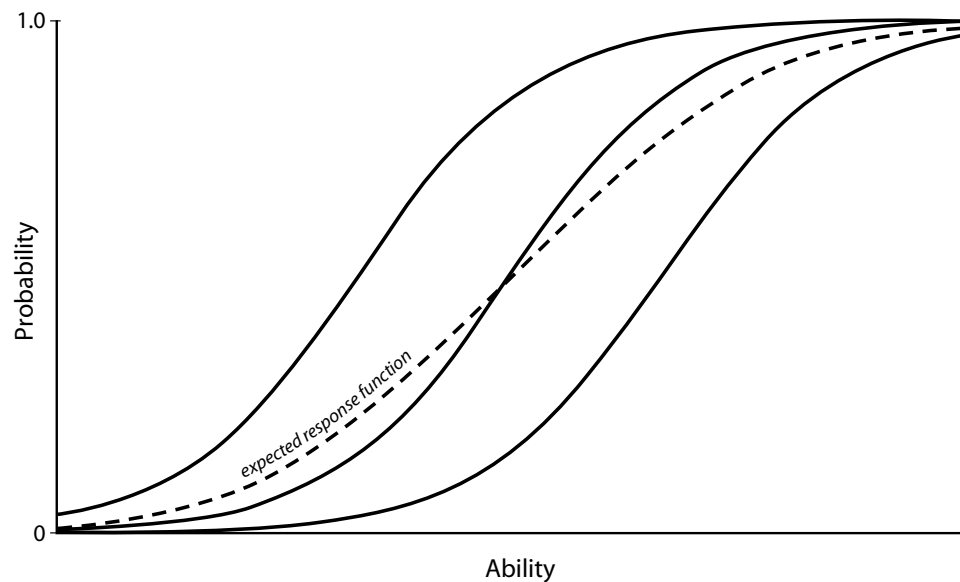
**Figure 4**

Figure 4. Graph of expected response function (dashed curve) against three item characteristic curves at three levels of difficulty.

## Feasibility Studies

This section details the studies conducted to investigate the use of the item models in adaptive testing. The simulation study and its results are presented to evaluate the feasibility of adaptive testing based on item models under controlled conditions. The field study consisting of an experimental, on-the-fly adaptive test and three forms of an experimental linear test is reviewed to evaluate the feasibility of applying the models under more realistic conditions.

### Simulation Study

As noted earlier, the feasibility of an approach based on item models rests in part on the extent to which the instances from the item models are isomorphic. A simulation approach is an expedient procedure to perform what-if analyses of the effect of lack of isomorphism on score reliability under idealized but realistic conditions. Specifically, in this case the simulation generates data for an adaptive test using many of the conditions present in the field study – specifically, using the same adaptive algorithm and the same item parameter estimates. The simulation study was conducted to assess the impact of different levels of isomorphism on score precision. A program was used to simulate the results of administering to a sample of examinees (or simulees) an adaptive test based on the item pool described previously, comprised of 261 items and 147 item models. Each model was calibrated using the ERF and for any model selected an instance was randomly drawn from that model.

The input to the simulation was the same as it was for the ERF program: a mean vector  $\beta$  corresponding to the original parameters and a  $\Sigma$  matrix describing the covariation among  $a$ ,  $b$ , and  $c$  under each of the three scenarios. Conceptually, the simulation procedure was as follows:

For each replicate at a value of theta:

If required item is from an item model, then:

Choose the next item [satisfying relevant constraints and current value of theta].

Draw a set of “true”  $a$ ,  $b$ , and  $c$  parameters from a distribution with mean  $\bar{a}, \bar{b}, \bar{c}$  [set to the PowerPrep parameter estimates] and a common covariance matrix.

Compute probability of correct response for current theta and the  $a$ ,  $b$ ,  $c$  drawn in the previous step.

If above probability > draw from a rectangular [0,1] distribution, response is correct; incorrect otherwise.

Update estimated ability using attenuated item parameter estimates.

Else [required item is a regular item]:

Using PowerPrep  $a$ ,  $b$ ,  $c$  for this item:

Compute probability of correct response for current theta.

If above probability > draw from a rectangular [0,1] distribution, response is correct; incorrect otherwise.

Update estimated ability estimate using PowerPrep item parameter estimates.

Repeat until 28 items are administered.

It is important to note that, in the case of item models, the probability of a correct response is computed based on the “true” item parameters, but ability is updated with the attenuated parameter estimates. In contrast, for items, the probability of a correct response is computed based on the original PowerPrep® item parameters rather than from a set of parameters drawn from a distribution. This difference in procedure means that whether a given examinee gets an item correct or not will depend on “true” item parameters regardless of whether the item is a regular item or an instance from a model.

We conducted four simulations. The “no isomorph” condition can be thought of as the case in which each item model produces instances that are isomorphic – that is, with identical item parameters. Alternatively, we can think of this condition as a case in which there is a single item and we know its true parameters. In either case, the parameters used to compute the response probability and updating theta are the same and, therefore, rather ideal.

For the other three simulations, the procedure creates a discrepancy between the parameters used to compute the response probability and the parameter estimates used to update ability. The magnitude of the discrepancy is determined by the covariance matrix used. Specifically, the higher the variability of the  $b$  estimates, the shallower the slope of the ERF, and therefore, the greater the discrepancy between the ERF and the “true” ICC. The greater this discrepancy is, the less information is contributed by the modeled item to estimation of ability.

## Results of the Simulation Study

For our purposes, the most relevant outcome of the simulation is an assessment of bias and standard error at different levels of ability for each of the four conditions. For historical reasons, ability is expressed on a true-score metric ranging from 0 to 60. Figure 5 shows the standard error for the four conditions. The solid curve plots the conditional standard error of measurement at different true abilities. This standard error is simply the standard deviation of the difference between estimated and true ability at each value of ability. As noted earlier, the curve for the no-isomorph condition might be viewed as unrealistically high because it assumes the item parameters are known rather than estimated. Nevertheless, the best-case scenario closely matches this curve. For the medium- and worst-case scenarios, a loss in precision of measurement is observed. It is not the case, as one might have expected, that the medium-case scenario is between the worst-case and best-case scenarios. Instead, the medium- and worst-case scenarios cluster closely. Therefore, these results are suggestive rather than indicative of the loss of precision we might expect. A possible reason for this clustering is that the variability embodied in the medium-case scenario really impacts measurement precision as severely as the worst-case scenario.

**Figure 5**

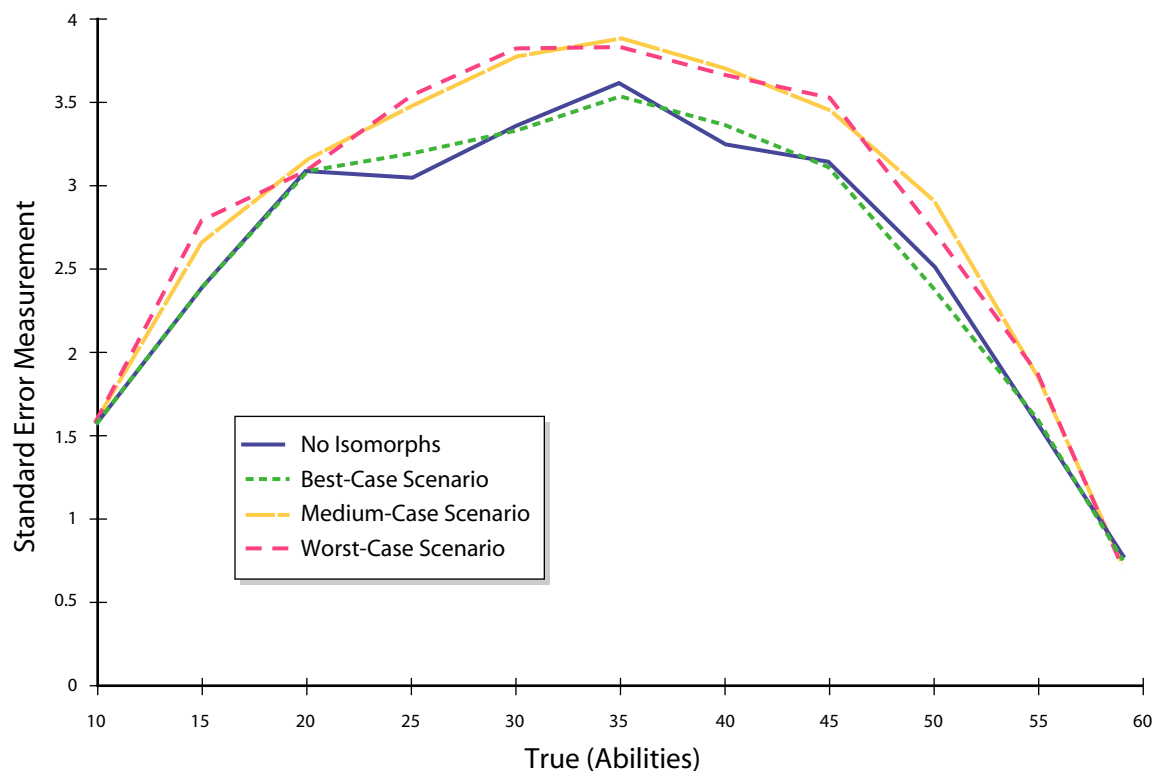


Figure 5. Standard error for the four simulated testing conditions.

Figure 6 shows the results for bias. As can be seen, no bias is observed under any condition. Thus, as has been observed elsewhere (Bejar, 1996; Embretson, 1999), the impact of lack of isomorphism is primarily in measurement precision, at least within limits, and the losses at some levels of ability appear to be minimal. This outcome is fortunate, as a loss of precision can be compensated by lengthening the test, but bias would be more difficult to correct.

**Figure 6**

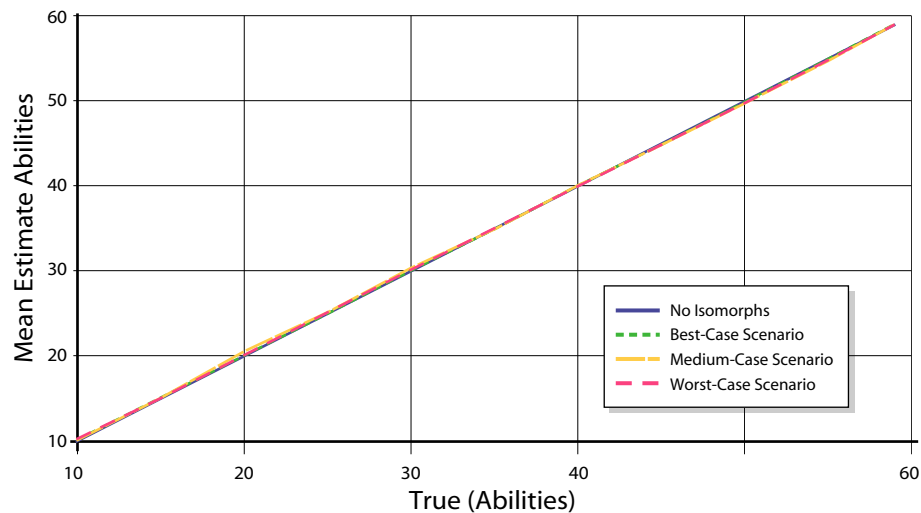


Figure 6. Estimated versus true abilities in the four testing conditions.

## Field Study

The foregoing results, based on a simulation of GRE testing conditions with item models in lieu of items, suggest the psychometric feasibility of an adaptive test consisting entirely of item models provided lack of isomorphism is moderate. In this section we present the results of the empirical study.

### Test Delivery System

The system we used consisted of a test delivery system, an item generation system, and a database of items and item models. The test delivery system was Web-based – meaning that the student interacted with the system through an Internet browser that resided on a local computer that in turn interacted with a server by way of the Internet.<sup>3</sup>

The test delivery system managed the interaction with the examinee, decided which item to administer next, and called the item generation system to instantiate an item model or to retrieve an item from the database. The test delivery system then sent a fully formatted item to the browser for display. The browser, in turn, returned response information. The test delivery system scored and recorded the response and updated the ability estimate. At that point, a new item or item model was selected from the database following an adaptive item-selection algorithm, and the process was repeated until all 28 items were administered.



## Participants

Two hundred eighty-two paid volunteers were recruited through flyers, newspaper advertisements, word of mouth, and other methods from college seniors and first-year graduate students who had taken the GRE General Test between January 1998 and January 2001. Data were collected in computer laboratories at Michigan State University (East Lansing), Fordham University (New York City), and CompUSA training centers (in New York City and Philadelphia). After eliminating records due to data problems, data for both the on-the-fly adaptive test and linear tests remained intact for 243 participants. Data for the linear test alone remained intact for 277 participants.

As shown in Table 1, males comprised 48%–49% of the study sample, as compared to 35% in the GRE test-taking population. However, the most notable difference between the current sample and the GRE population occurred in the ethnicity distribution. In the present study, Asians were overrepresented by 42 percentage points and Whites were underrepresented by 37 percentage points.

**Table 1**      **Demographic Characteristics of Subjects Versus GRE Test-Taking Population**

Attribute	Adaptive test n = 243	Linear test n = 277	GRE operational test* (annually)
<b>Gender</b>			
Male	49%	48%	35%
Female	51%	52%	65%
<b>Ethnicity</b>			
Native American or Alaskan Native	1%	1%	1%
Black or African American	4%	4%	9%
Mexican, Mexican American, Chicano	1%	1%	2%
Asian, Asian American, or Pacific Islander	47%	47%	5%
Puerto Rican	0%	0%	1%
Other Hispanic or Latin American	0%	0%	2%
White (non-Hispanic)	40%	40%	77%
Other	7%	7%	3%
<b>Citizenship Status</b>			
U.S. citizen	50%	50%	75%
Non-U.S. citizen	50%	50%	25%

\* Source: Educational Testing Service. (2000). *Graduate Record Examinations: Sex, race, ethnicity, and performance on the GRE® General Test 2000-2001* (I.N. 989404). Princeton, NJ: Author.

## Instruments and Data Collection

### *Experimental, on-the-fly adaptive test*

The experimental adaptive test was comprised of the same pool (and calibrations) used in the Simulation Study – 261 items and 147 item models. Of the 147 item models, 101 were quantitative-comparison item models and 46 were problem-solving item models. Subjects were administered the computer-adaptive test first and were allowed 45 minutes to complete the test.

### *Experimental linear test*

A linear test was administered to participants, after they completed the experimental, adaptive test. Participants were randomly assigned to one of three computer-administered test forms, each comprised of 30 items. Each form was intended to be parallel, having been generated from the same set of item models, with the item generated by a given model always appearing in the same position across forms. The first 20 items were created from 20 different item models; the first 12 were quantitative-comparison items and the last 8 were problem-solving items. Within these two groupings, the models were sequenced in order of difficulty from easiest to hardest. The last 10 items on each of the linear tests were comprised of new instances generated from 10 of the same 20 item models used for the first part of the test – 5 were randomly chosen quantitative-comparison item models and 5 were randomly chosen problem-solving item models. Subjects were allowed 45 minutes to complete the linear test. The 20 item models used for the linear forms were not administered as part of the on-the-fly adaptive test.

### *GRE scores*

In addition to the experimental scores described above, we also obtained for each subject the GRE score they had obtained previously.

## Results of the Field Study

### Experimental Adaptive Test Results

Our main interest in the field study was to explore the comparability of the experimental scores and GRE quantitative scores previously obtained by each study participant. We first sought to determine whether the scores were on a comparable metric. Second, we sought to determine how well correlated the GRE and experimental scores were.

Table 2 shows the mean scores and standard deviations for study participants on both the GRE and the experimental tests as well as for the overall GRE test-taking population. Comparing the mean GRE score of our sample, 718, to the mean of 565 for the GRE test-taking population, we see that our sample appears to be much more able in quantitative reasoning than the GRE population as a whole. Our subjects are also much more homogeneous. The GRE score standard

deviation for our sample is 88, whereas it is 143 for the GRE test-taking population. Table 2 also indicates that participants' experimental scores are lower than their GRE scores, and that the variability of the experimental scores is somewhat higher. The second aspect of comparability – the relationship between operational and experimental scores – shows a more promising result; the correlation between the two sets of scores was .87. This correlation turns out to be as high as the GRE quantitative section's test-retest correlation (R. Durso, personal communication, January 18, 2000).

**Table 2** Experimental and GRE Scores for Study Participants

	Mean	SD
GRE score of sample	718	88
Experimental score of sample	693	101
GRE score for test-taking population	565	143

Note.  $N = 243$

Recall that the 28-item experimental adaptive test was composed of both items and item models. No subject's test consisted of fewer than 14 models, and some subjects received as many as 21 models. Thus, an adaptive GRE quantitative experimental score in which 50% to 75% of the items were generated by item models was able to order examinees equivalently to the GRE score. Although the high correlation with GRE scores is reassuring, the difference in score scale warrants additional investigation. At one level, the drop is not surprising. First, these were high scoring students; regression to the mean would explain some of the drop. Second, lower motivation in the study context could explain part of the drop as well.

To fully explore the latter idea, one might hypothesize that perseverance on the more difficult items would be lessened under experimental conditions, or that students would not try hard enough in general. Given our sample size, the adaptive nature of the test, and the absence of response-time data for the original PowerPrep® pool, our analytic options were limited. Nevertheless, we examined the responses of students for whom there had been a large change in scores. Differences between GRE and experimental scores – from a drop of 150 points to a gain of 90 points – were examined. Although such score changes also occur in an operational setting, we wanted to examine any study factors that may have had some influence on these differences. For subjects whose experimental scores were more than 50 points lower than their GRE scores (71 subjects), we examined:

- Occurrences of computer abnormalities during the testing session
- Total number of completed items
- Number of models administered to the student
- Number of items completed in less than 10 seconds
- Overall completion time for the adaptive test

We could not detect any patterns from this examination. We also examined the possibility that the drop was the result of using attenuated parameters in estimating ability. To that effect, we recomputed experimental scores with the original PowerPrep® parameter estimates. However, the recomputed scores did not change either the correlation with the GRE score or the mean score.

### Experimental Linear Test Results

Our interest in conducting this analysis was to assess the equivalence of different instances of the same models and their relationship to the difficulty estimates for the items from which they originated. The fact that each of the three linear tests we administered was comprised of different instantiations of the same item models, and that these item models had not been administered as part of the adaptive test, facilitated this investigation.

The estimated difficulties for the three instances of each item model were computed by obtaining the logit of the proportion correct for each instance. Table 3 shows the correlation among the three sets of model instances and with the difficulty estimates from PowerPrep®. Correlations with the estimates range from .77 to .87. Correlations among the difficulties of the model instances range from .80 to .88. Table 4 displays the corresponding means and standard deviations.

**Table 3** Correlation of PowerPrep® Difficulty Estimates With Estimates for Linear Isomorphic Test Forms

	PowerPrep®	Form 1	Form 2	Form 3
PowerPrep®	–	.87	.82	.77
Form 1		–	.81	.88
Form 2			–	.80
Form 3				–

**Table 4** Means and Standard Deviations of Difficulty Estimates for PowerPrep® and Linear Isomorphic Forms

	Mean	SD
PowerPrep®	0.09	1.15
Form 1	-0.65	0.47
Form 2	-0.52	0.36
Form 3	-0.52	0.37

Figure 7 plots the difficulties associated with each experimental linear test form against the difficulty estimates obtained from PowerPrep®. The most salient finding is the different scales of the experimental versus PowerPrep® parameters. This difference is not surprising because our subjects were high scoring compared to the overall GRE test-taking population. As noted earlier, item model instances were placed on the test in order of difficulty (easy-to-hard) based on PowerPrep®

difficulty estimates; the first 12 items were quantitative-comparison items and the next 8 items were problem-solving items. It was these first 20 items in each test form that were evaluated for item difficulty and response time. As can be seen from the graph, item difficulty increases serially up to the twelfth item. It appears that difficulties increase more rapidly for the PowerPrep® items, but in reality, difficulties for the item model instances are on a different metric – that is, difficulties of the item models are logit-based and difficulties of PowerPrep® items are 3PL  $b$  estimates. The same pattern is observed for the 8 problem-solving items (items 13–20). Difficulty estimates obtained for the model instances are closely clustered, as might be expected if the item models were yielding equivalent instances.

**Figure 7**

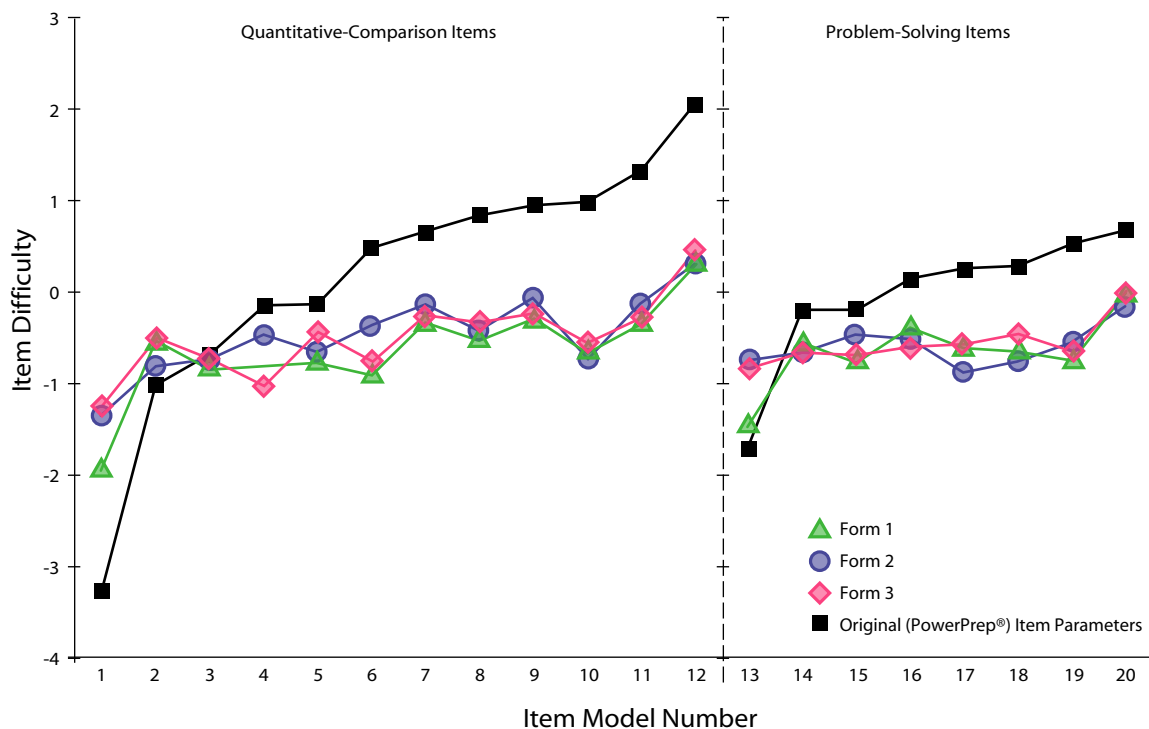


Figure 7. Comparison of difficulty estimates for PowerPrep® and linear forms by position of item on linear form.

This suggestion of isomorphism is reinforced by an analysis of response time. Figure 8 shows the mean response time for the 20 model instances in each of the three experimental linear test forms corresponding to the data shown in Figure 7. (Unfortunately, the mean response time for the PowerPrep® difficulty estimates was not available.) Figure 8 suggests that indeed the item model instances are equivalent because they are tightly clustered together within an item model, while across models there is substantial variability. It is interesting to note that, unlike the case for difficulty, there is no serially increasing trend within item type for response time. In summary, the analyses of difficulty and response time both suggest that the item models indeed produced isomorphic instances.

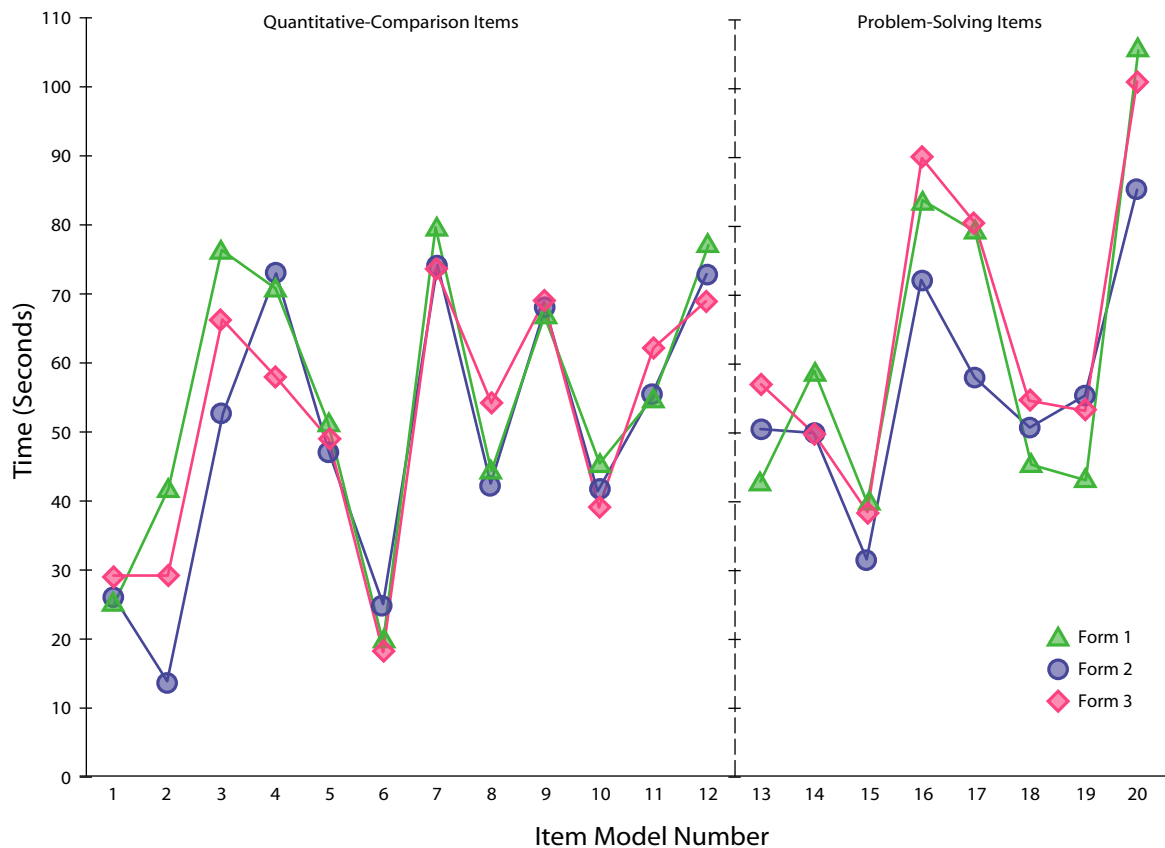
**Figure 8**

Figure 8. Mean response time for items in linear forms by serial position.

## Discussion

The results of this study provide initial evidence that an approach to adaptive measurement of quantitative reasoning based on item models is feasible psychometrically. The simulation study showed some erosion of precision due to lack of isomorphism but the empirical study showed that an experimental quantitative GRE correlated highly with previously obtained GRE scores. Moreover, analyses of the experimental linear tests suggest a high level of isomorphism was achieved. Nevertheless, it would be important in future studies to characterize more precisely the measurement properties of adaptive testing based on item models.

A confirmation of these findings through more extensive studies could open the door to more cost effective approaches to adaptive testing of quantitative reasoning. The cost improvement would not necessarily entail a sacrifice in score precision. The drops in score precision we observed with simulated data, in principle, can be compensated for with a slight lengthening of the test. Moreover, our highly selective student sample, the actual high score correlation with GRE scores, and

the consideration that nearly half of the items were from item models suggest that, in reality, measurement quality need not be severely affected under an adaptive design based on item models. Here we discuss the nature of additional evidence needed to corroborate our main conclusion that an adaptive, generative model is a technically feasible, cost-effective approach to admissions testing.

An obvious practical advantage of item models is that they can yield many pre-calibrated items once the item model has been calibrated. However, in a high stakes environment the availability of a large supply of items does not necessarily reduce cost. The items must be the right items, that is, the item models should be selectable by the adaptive algorithms to, in effect, distribute exposure over the entire repository of item models. Steps toward that goal require changes in the item selection algorithm and the content classification of the item models such that the goal of even and distributed exposure can be achieved. A redesign of the adaptive model together with the use of item models, whereby item models are approximately equally and predictably exposed, would represent a major achievement – this is under active investigation.

A related concern is the similarities that may exist among instances of an item model. Such similarities could be exploited in a number of ways to advantage some students. Therefore, it is essential to create models that produce dissimilar instances while maintaining homogeneity of item parameters. Additionally, models could be created in such a way that the content similarities of the items they produce could not be counted on by test takers to arrive at a correct answer without the skills the item models were designed to tap. Work by Morley, Bridgeman & Lawless (2003) addresses the issue indirectly.

A further issue relates to model calibration and the effects of variation in parameter estimates on examinee scores. Rizavi, Way, Davey, and Herbert (2002) examined the variation in item parameter estimates that occurs over repeated uses of the same GRE items. Such estimates of parameter variability provide a benchmark for judging isomorphism because they represent the variability in estimates when the same item is recalibrated. More empirical results are needed to estimate the variability in item parameter estimates over presumably isomorphic instances of item models. However, the results presented here demonstrate that we were able to attain a high level of isomorphism in this study.

The foregoing issues should be resolved with additional research. However, while practical feasibility is an appropriate concern, it may be equally important that, from a theoretical perspective, item models can enhance validity because by designing a test with item models requires taking advantage of the cognition of the construct under measurement. Once we have incorporated theoretical knowledge into the item model, its use as part of a test represents a test of that knowledge. Specifically, if isomorphism does not hold, an investigation of the reasons is bound to serve as refinement of the underlying theoretical basis. If isomorphism holds, the underlying theoretical basis is further supported. Such ongoing monitoring of theoretical prediction and the resulting enhancement of the knowledge base behind a test should greatly enhance the validity evidence behind scores.

## Summary and Conclusion

The goal of this study was to assess the feasibility of an approach to adaptive testing based on item models. The study was motivated by some of the challenges raised by continuous adaptive testing – most notably the increased need for new items in order to maintain acceptable test and item security. We first presented results from a simulation study designed to explore the effects of item modeling on score precision and bias. The results showed that under different levels of isomorphism, there was no bias, but precision of measurement was eroded, especially in the middle range of the true-score scale. We feel that more extensive simulations need to be done to better understand the impact of item models on score precision.

We next presented results from a field study in which we administered an experimental, on-the-fly, adaptive quantitative-reasoning test as well as an experimental linear test form. Because it was not feasible to calibrate item models as part of this study, we recalibrated existing item parameters assuming the greatest lack of isomorphism used in the simulation. That is, we attenuated the item parameters of 147 item models from their original parameter estimates, assuming a covariance matrix among item parameters with a high variance for difficulty.

The resulting comparisons with GRE scores were extremely reassuring. The correlation of experimentally obtained scores and GRE scores was .87, which is similar to the test-retest correlation observed under operational conditions for test takers choosing to repeat the test. This correlation is especially meaningful because our sample was made up of a highly selective group of subjects and because participants received a large percentage of items from item models. We did find a reduction in mean performance, which we attributed to a combination of regression and, possibly, lower student motivation. We also presented analyses of the functioning of items on linear isomorphic forms – specifically, difficulty and response time. Both analyses suggested a high level of isomorphism across items within models. This high level of isomorphism is likely the reason we obtained a correlation with GRE scores that was indistinguishable from test-retest correlations.

Although, as discussed earlier, a transition to an operational on-the-fly approach presents significant challenges in the areas of exposure control and model parameterization, we conclude that this study provides a promising first step toward what we hope will be significant cost reduction and theoretical improvement in test creation methodology for educational assessment.



## Endnotes

- <sup>1</sup> Janssen, Tuerlinckx, Meulders, and De Boeck (2000) and Wright (2002) have proposed such models. Other Bayesian approaches that can be oriented to generation can be found in Bradlow, Wainer, and Wang (1999) and Fox and Glas (1998). One program, Scoright (Wang, Bradlow, & Wainer, 2000), already exists for the three-parameter and graded-response case, although it was originally developed to model dependencies in sets of questions with a common stimulus and in its present form is not precisely suited to calibrating item models. Glas and van der Linden (2003); Johnson and Sinharay (2002); Sinharay, Johnson, and Williamson (2003); Williamson Johnson, Sinharay, and Bejar (2002a); and Williamson, Johnson, Sinharay, and Bejar (2002b) discuss approaches based on Multi-Chain-Monte Carlo (MCMC) estimation specifically in connection with item models. Applications to educational surveys (e.g., Hombo & Dresher, 2001) are also under investigation.
- <sup>2</sup> We are grateful to Bob Mislevy for suggesting the use of expected response functions in the context of item modeling.
- <sup>3</sup> The technical report on which this article is based contains a more detailed description of the system (<http://www.ets.org/research/dload/RR-02-23.pdf>).

## References

- Bejar, I. I. (1986). *Final report: Adaptive testing of spatial abilities* (ONR 150 531). Princeton, NJ: Educational Testing Service.
- Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, 14(3), 237–245.
- 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: Lawrence Erlbaum.
- Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS Research Report 96–13). Princeton, NJ: Educational Testing Service.
- Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Generating items from cognitive tests: Theory and practice* (pp. 199–217). Mahwah, NJ: Lawrence Erlbaum.
- Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation: Final report to the National Council of Architectural Registration Boards* (Research Memorandum 99-2). Princeton, NJ: Educational Testing Service.
- Bejar, I. I., Fairon, C., & Williamson, D. M. (2002, June). Multilingual item modeling as a mechanism for test adaptation: Applications to open ended and discrete items. In D. Bartram (Chair), *Item & Test Generation*. Symposium conducted at the International Conference on Computer-Based Testing and the Internet: Building Guidelines for Best Practice (ITC Conference 2002), Winchester, England.
- Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement*, 15(2), 129–137.

- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64(2), 153–168.
- Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematics skills. *Cognitive Science*, 2, 155–192.
- Clancey, W. J. (1986). Qualitative student models. *Annual Review of Computer Science*, 1, 381–450.
- Educational Testing Service. (2000). *Graduate Record Examinations: Sex, race, ethnicity, and performance on the GRE General Test 2000–2001* (Identification Number 989404). Princeton, NJ: Author.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Hillsdale, NJ: Lawrence Erlbaum.
- Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407–433.
- Fox, J. P., & Glas, C. E. W. (1998). *Multi-level IRT with measurement error in the predictor space* (Research Report 98–16). Enschede, the Netherlands: University of Twente.
- Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, 27(4), 247–261.
- Hively, W. (1974). Introduction to domain-reference testing. *Educational Technology*, 14(6), 5–10.
- Hively, W., Patterson, H. L., & Page, S. H. (1968). A “universe-defined” system of arithmetic achievement tests. *Journal of Educational Measurement*, 5(4), 275–290.
- Hombo, C. M., & Drescher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Hornke, L., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, 10(4), 369–380.
- Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.
- Irvine, S. H., Dunn, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology*, 81, 173–195.
- Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, 25(3), 285–306.

- Johnson, M. S., & Sinharay, S. (2002, April). *Hierarchical approaches to item model calibration*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Kenney, J. F. (1997). New testing methodologies for the Architect Registration Examination. *CLEAR Exam Review*, 8(2), 23–28.
- LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modeling procedure for constructing content-equivalent multiple-choice questions. *Medical Education*, 20(1), 53–56.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Macready, G. B. (1983). The use of generalizability theory for assessing relations among items within domains in diagnostic testing. *Applied Psychological Measurement*, 1, 149–157.
- Martin, J. D., & 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., & 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: American College Testing Program.
- Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75–99). Dordrecht, the Netherlands: Kluwer Academic Publishers.
- 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: Lawrence Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Roles of task model variables. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 102–106). Mahwah, NJ: Lawrence Erlbaum.
- Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (ETS Research Report 94-28-ONR). Princeton, NJ: Educational Testing Service.
- Morley, M. E., Bridgeman, B., & Lawless, R. R. (2003, April). Impact on the use of item variants on math test performance. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.
- Rizavi, S., Way, W. D., Davey, T., & Herbert, E. (2002, April). *Tolerable variation in item parameter estimation*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

- Sinharay, S., Johnson M. S., & Williamson, D. M. (2003). An application of a Bayesian hierarchical model for item family calibration (ETS Research Report 03-04). Princeton, NJ: Educational Testing Service.
- Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Lawrence Erlbaum.
- Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, 24, 86–97.
- Skinner, B. F. (1958). Teaching machines. *Science*, 128, 969–977.
- Skinner, B. F. (1961). Why we need teaching machines. *Harvard Educational Review*, 31, 377–398.
- Uttal, W. R., Rogers, M., Hieronymous, R., & Pasich, T. (1969). *Generative CAI in analytic geometry*. Ann Arbor, MI: University of Michigan.
- van Lehn, K. (1988). Student modeling. In J. J. Richardson & M. C. Polson (Eds.), *Intelligent tutoring systems* (pp. 55–78). Hillsdale, NJ: Lawrence Erlbaum.
- Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Wang, X., Bradlow, E. T., & Wainer, H. (2000). *A general Bayesian model for testlets: Theory and applications* (GRE Board Professional Report No. 98-01). Princeton, NJ: Educational Testing Service.
- Williamson, D. M., & Bejar, I. I. (2002, April). *Using testlet response theory to evaluate the equivalence of automatically generated multiple-choice items*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Williamson, D. M., Johnson, M. S., Sinharay, S., & Bejar, I. (2002a, April). *Hierarchical IRT examination of isomorphic equivalence of complex constructed response tasks*. Paper presented at the American Educational Research Association, New Orleans, LA.
- Williamson, D. M., Johnson, M. S., Sinharay, S., & Bejar, I. (2002b, April). *Applying Hierarchical model calibration to automatically generated items*. Paper presented at the American Educational Research Association, New Orleans, LA.
- Wolfe, J. H., & Larson, G. E. (1990). *Generative adaptive testing with digit span items*. San Diego, CA: Testing Systems Department, Navy Personnel Research and Development Center.
- Wright, D. (2002). Scoring tests when items have been generated. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 277–286). Mahwah, NJ: Lawrence Erlbaum.

## Author Biographies

Isaac I. Bejar is Director of the Assessment Design and Scoring Center at ETS Research and Development Division. He has been involved in research in adaptive testing since the 1970's and more recently in the development of generative assessment methods and automated scoring. He is currently co-editing a volume on automated scoring. Dr. Bejar earned a Masters and Ph. D from the Department of Psychology at the University of Minnesota.

René R. Lawless is a senior research associate in the Center for Assessment Design and Scoring in the Research & Development Division at Educational Testing Service. Since 1999, she has been involved in projects investigating the feasibility of the use of item generation in mathematics test items and understanding how students transfer information about mathematics test items. Ms. Lawless earned her Ed.M in educational statistics and measurement from the Graduate School of Education at Rutgers University. She received her B.A. in sociology from Douglass College.

Dr. Mary Morley is a Mathematician with 12 years experience in teaching Mathematics. She has taught mathematics courses at University of Chicago, University of Utah, Rutgers University, Purdue University, and Temple University. Since joining ETS in 1992, she has worked in both research and test development. She has also conducted research in the areas of Numerical Analysis and computer mathematics. She is the author of several publications on numerical solutions of partial differential equations. In 2003, Dr. Morley joined the staff of the College Board.

Michael Wagner is a Director of Product Development at Educational Testing Service. He has been with ETS for the past 12 years, working in both Research and Product Development. His current interests are in technological support for the test development process, collection and scoring of constructed response mathematics, and new test delivery mechanisms for complex tasks. Mr. Wagner has an M.S. degree in Computer Science from the University of Iowa.

Randy Elliot Bennett is Distinguished Presidential Appointee in ETS Research & Development. Since the 1980's, he has conducted research on the applications of technology to testing, on new forms of assessment, and on the assessment of students with disabilities. Dr. Bennett's work on the use of new technology to improve assessment has included research on presenting and scoring open-ended test items on the computer, on multimedia and simulation in testing, and on generating test items automatically. He is currently co-directing the Technology Based Assessment Project, a series of studies designed to explore computerized testing in the U.S. National Assessment of Educational Progress.

Javier Revuelta is currently Assistant Professor at the Autonoma University of Madrid, Spain, where he also earned a Ph.D. His research interests include item response modeling, Bayesian inference and statistical simulation for psychometric models, and computerized adaptive testing, in particular item exposure control.



## The Journal of Technology, Learning, and Assessment

### Editorial Board

**Michael Russell, Editor**  
Boston College

**Allan Collins**  
Northwestern University

**Cathleen Norris**  
University of North Texas

**Edys S. Quellmalz**  
SRI International

**Elliot Soloway**  
University of Michigan

**George Madaus**  
Boston College

**Gerald A. Tindal**  
University of Oregon

**James Pellegrino**  
University of Illinois at Chicago

**Katerine Bielaczyc**  
Harvard University

**Larry Cuban**  
Stanford University

**Lawrence M. Rudner**  
University of Maryland

**Mark R. Wilson**  
UC Berkeley

**Marshall S. Smith**  
Stanford University

**Paul Holland**  
ETS

**Randy Elliot Bennett**  
ETS

**Robert J. Mislevy**  
University of Maryland

**Ronald H. Stevens**  
UCLA

**Seymour A. Papert**  
MIT

**Terry P. Vendlinski**  
UCLA

**Walt Haney**  
Boston College

**Walter F. Heinecke**  
University of Virginia

[www.jtla.org](http://www.jtla.org)