GRE®
RESEARCH

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
Javier Revuelta

October 2002

GRE Board Professional Report No. 98-12P

ETS Research Report 02-23

ETS® Educational Testing Service

Princeton, NJ 08541

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

Isaac I. Bejar, René R. Lawless, Mary E. Morley, Michael E. Wagner, and Randy E. Bennett
Educational Testing Service

Javier Revuelta
Universidad Autónoma de Madrid, Madrid, Spain

GRE Board Report No. 98-12P

October 2002

This report presents the findings of a
research project funded by and carried
out under the auspices of the
Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Educational Testing Service
Princeton, NJ 08541

Abstract

The goal of this study was to assess the feasibility of an approach to adaptive testing based on item models. A simulation study was designed to explore the affects of item modeling on score precision and bias, and two experimental tests were administered — an experimental, on-the-fly, adaptive quantitative-reasoning test as well as a linear test. Results of the simulation study showed that under different levels of isomorphicity, there was no bias, but precision of measurement was eroded, especially in the middle range of the true-score scale. However, the comparison of adaptive test scores with operational Graduate Record Examinations (GRE) test scores matched the test-retest correlation observed under operational conditions. Analyses of item functioning on linear forms suggested a high level of isomorphicity across items within models. The current study provides a promising first step toward significant cost and theoretical improvement in test creation methodology for educational assessment.

Keywords: Adaptive testing, CAT, item response theory, expected response function, automated item generation, quantitative reasoning

Acknowledgements

We would like to acknowledge the involvement of many individuals, without whom this project simply would have not been feasible. Bob Mislevy has been helpful in this and previous efforts as a source of advice. In connection with this project, he and Marilyn Wingersky devised a procedure to enhance the expected response function program. Martha Stocking modified the simulation program used to evaluate Graduate Record Examinations (GRE®) item pools to enable it to use item models. Without her participation we would not have been able to conduct the simulation study reported here. Len Swanson was the source of the adaptive engine we used in the test delivery program. He graciously modified it as necessary to deal with item models. Bob Smith was the source of the estimated parameters for the linking items from which we derived the covariance matrices used to attenuate parameters. Tim Davey reviewed an earlier draft of this report. Rob Durso provided information on GRE quantitative test-retest performance from a specially conducted study. Steve Szyszkiewicz lent valuable assistance in data management and system programming. Many test developers reviewed and edited item models: Beth Brownstein, Gloria Dion, Daryl Ezzo, Michael Grinfeld, Jutta Levin, Steve Penrice, Judy Smith, Esther Tesar, Sheng Wang, and Barbara Wiener. Their role in this project was critical, and without their assistance and insight, this project would not have been possible. Lisa Hemat was invaluable with logistical assistance: proofreading item models, plastering Philadelphia with recruitment flyers, and designing and placing newspaper advertisements.

A special thanks goes to our collaborators Ed Wolfe of Michigan State University and Mitch Rabinowitz of Fordham University. Without their help, our data collection efforts would have been unattainable. Their assistance proved critical in obtaining computer laboratory space and identifying unusually interested and conscientious student proctors.

We would like to thank our proctors Chandra Donnell and Michelle Carlson from Michigan State University. Kevin Moloney and Teresa (Tracey) Hogan from Fordham University were wonderful proctors, and their assistance with contacting college newspapers was a big help. We would like to say a special “thank you” to the following proctors who went above and beyond the call of duty: Bing (Sabrina) Shi from the University of Pennsylvania for her assistance in reaching the college communities of Philadelphia and for proctoring; Lixiong Gu and Somvung (Ford) Vongpunsawad from Michigan State University, who dedicated much of their free time to recruiting, beta-testing, proctoring, and assisting us with troubleshooting; and

Elena Kokkinofta, who not only recruited subjects, reconfirmed their appointments, and arranged special test sessions for individuals, but who also traveled throughout New York City to ensure that our study was publicized at city colleges and universities.

Table of Contents

List of Tables

List of Figures

Introduction

After the introduction of computer-based testing in the early 1990s, it became obvious that continuous testing presented different challenges than paper-and-pencil testing. In particular, the cost of content production increased significantly. The preference of certain items by the item selection algorithm, coupled with the relative unpredictability of item statistics and the time consuming nature of the pretesting process needed to compensate for that unpredictability, conspired to increase costs in order to maintain acceptable security. A generative approach to creating test content can potentially alleviate some of these problems. In this report we present the results of an investigation into the feasibility of measuring quantitative reasoning in a generative *and* adaptive mode.

Generative testing (e.g., Bejar, 1993) is a construct-driven approach to assessment that may have, in addition to its measurement benefits, many significant practical advantages. In this report, we have adopted the term *item modeling* to refer this approach. An item model (Bejar, 1996) can be thought of as a procedure for instantiating isomorphic items — items that contain comparable content and are exchangeable psychometrically. We view item modeling as construct-driven because it entails an understanding of the goals of the assessment and the application of pertinent psychological research. That is, item models set the expectation for the behavior of the instances produced by a given model, and those expectations can be verified upon administration of the isomorphs, thus providing an opportunity to refine our understanding of the construct and supporting psychological principles. In addition to their role as a validity-enabling approach, 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 reviewed and formatted, regardless of how similar it may be to other items. An item modeling approach automates many of the details of producing instances once the item model has been formulated.

In this report, we explore the feasibility of item modeling in conjunction with adaptive testing. That is, the "item pool" consists of a combination of item models and items. In this approach, instances of a model are presented at delivery time — that is, they are generated *on-the-fly* from an item model. Otherwise, the same adaptive engine that is used operationally is employed.

The goal of the present study was to compare Graduate Record Examinations (GRE®)

General Test scores obtained operationally with scores obtained by way of a *generative adaptive* examination. However, it is important to state at the outset that we see item modeling only as a partial solution to the challenges presented by continuous, computer-based admissions testing. For example, a specific challenge is controlling the exposure of items (e.g., Stocking & Lewis, 2000). Item models, like items, could be over-exposed unless the appropriate precautions are taken. Therefore, a complete solution to the challenges of continuous adaptive testing would require, among other things, a test specification that satisfies difficulty, content, and exposure constraints in a manner that is consistent with the construct we wish to measure.

Item Modeling

Generative assessment is concerned with the systematic creation of items based on principles. It is appealing to create the items by automated means, but the mode of generation is not critical. Item models could also be instantiated manually by test developers using the models as a prescription for authoring. This approach was used successfully in the development of a computer-based licensing examination for architects (Kenney, 1997; Bejar & Braun, 1999; Bejar, 2002). Generative assessment has its roots in computer-assisted instruction (e.g., Uttal, Rogers, Hieronymous, & Pasich, 1969) and in criterion-referenced testing (Hively, 1974). Hively's work emphasized automated generation. In Hively's 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 1960s, which were based on Skinnerian principles. Skinner (1954, 1958, and 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,* which 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; Bejar & Braun, 1997; Bejar & Yocom, 1991; Embretson, 1999; Hornke & Habon, 1986; Irvine, Dunn, & Anderson, 1990; Meisner, Luecht, & Reckase, 1993; Wolfe & Larson, 1990). A recent conference held at Educational Testing Service (ETS®; Irvine & Kyllonen, 2002) gave further examples of the feasibility of the approach in different domains. 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 much subsequent work.

A specific approach to generative modeling is based on *item modeling*, a term used by LaDuca, Staples, Templeton, and Holzman (1986). The term *task model* has also been proposed in the context of evidence-centered assessment design (Mislevy et al., 2002). *Task* model subsumes *item* model in several respects. First, task models are applicable to large tasks, including complex simulations. As a result, task models include explicit and detailed connections to other aspects of assessment design, such as scoring. Item models, which have been most frequently used with multiple-choice items, have simpler working connections with other test components. For example, as part of adaptive testing, an item model supplies items that are calibrated through item response theory (Lord, 1980). Based on those calibrations, an estimate of ability is obtained and the next item is chosen, in part, based on that estimate. The task models employed under evidence-centered assessment design could also be calibrated in this way, but in addition, through a much broader range of psychometric characterization. Thus, our use of *item model* is consistent with both the original meaning in LaDuca et al. (1986) as well as task

modeling in evidence-centered design. However, we do emphasize one aspect in the present work that is new — namely, the totally automated generation of instances of an item model as part of an adaptive testing procedure.

Automated item generation raises the question of the calibration of item models and their instances. The estimation approach would seem to depend in turn on our understanding of difficulty. At least two approaches to modeling difficulty within a generative approach seem feasible: strong theory versus weak theory (Bennett, in preparation). 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 variants 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 *between*, 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 widely vary 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 and impute the parameters to all instances of the model*. Therefore, the emphasis is on producing items that are well described by a single set of parameters.

Independent of whether we are operating under weak or strong theory, the specifics of parameter estimation 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 the first case, 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, 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. Janssen, Tuerlinckx, Meulders, and De Boeck (2000) and Wright (2001) 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. Applications to educational surveys (e.g., Hombo & Dresher, 2001) are also under investigation.

In the second case — in which 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. As we shall see below, 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.

Report Overview

As noted earlier, the goal of this investigation was to assess the feasibility of an approach to adaptive testing based on item models. The investigation involved three components:

1. Operational GRE General Test scores were compared with those from an experimental adaptive test that included both item models and traditional items.

2. A simulation study based on the same item pool was conducted to theoretically assess the impact of lack of isomorphism among item model instances.

3. Specific item models that were administered in a linear test following the administration of the adaptive test were analyzed to empirically assess the level of isomorphicity yielded by the item models.

In the next section, we describe the procedures for each of these components. Analyses and findings are presented next, followed by discussion and conclusion.

Procedures

Section Overview

In this section, we first describe the adaptive testing procedure that was used to collect data from subjects who previously took the GRE. The purpose of the adaptive testing was to assess the psychometric feasibility of on-the-fly testing. In other words, we sought to determine whether test scores based on items generated on-the-fly are equivalent to test scores based on items created by conventional methods, and whether these test scores can be considered equivalent to operational test scores.

Next, to corroborate our adaptive data collection procedures, we describe our simulation study, which we completed to help us understand the behavior of the scores under idealized conditions. Third, we relate the system we used for data collection — including some of the details of the Web-based delivery system we used to collect the data. Fourth, we detail the linear tests that we administered after the adaptive tests to further study the functioning of the item models. We conclude the section with a discussion of our recruitment procedures.

Adaptive Test Procedures

Modeling the Item Pool

A significant investment has been made to make continuous testing a reality after many years of research (see Wang et al., 2000, for a summary of research related to adaptive testing). With this in mind, our goal was to build upon that research foundation by utilizing the existing psychometric infrastructure as much as possible. Specifically, our implementation of on-the-fly adaptive testing relied on the GRE adaptive-testing model.

A key component of an adaptive test is the construction of one or more item pools. Our starting point in developing an on-the-fly adaptive exam was selecting an existing pool of items. In theory, the choice of the next item in an adaptive test is driven by the goal of maximizing the precision of measurement. When there are no other considerations at play, the process of choosing a next, maximally-informative item is simple. In practice, however, many practical considerations have to be factored into this decision. In particular, limits need to be imposed on which items are presented so that the set of items a student receives provides an adequate sampling of the content domain and that no one item is presented so frequently that its security is compromised.

Pools are constructed according to a complex procedure that aims to satisfy these many

constraints (e.g., Mills & Steffen, 2000). In the current study, we used a specific item pool known as CAT Pool 2, which is one of the pools released with PowerPrep®, a test preparation program distributed by ETS for the GRE General Test. A subset of 147 items from this 408-item pool was converted to item models. That is, each of the items in the subset was taken to be an instance of an item model that was created from that original item. The intent of item modeling is to be able to generate the specific instance that motivated the model, as well as many more psychometrically-equivalent instances. Those items with the highest predicted exposure 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. (Predicted exposure rate is calculated for a given pool through simulations, as part of the process of configuring new pools. For this pool the predicted exposure and observed exposure rates were found to be very similar). Table 1 displays counts of the models and items that comprise the study pool.

Table 1

Description of Item Pool by Item Type and Number of Item Models

	Type			
	Quantitative comparison	Problem solving	Data interpretation	Total
Models	100	47	0	147
Items	48	65	148	261
Pool	148	112	148	408

Some items were excluded from modeling for several reasons. First, we did not model any data-interpretation sets. One reason why data-interpretation sets were not modeled is because significant effort would have been required to make the items appear credible. Also, we did not model any quantitative-comparison or problem-solving items that had an extremely low exposure rate. Specifically, if the probability of a student receiving an item was less than 0.02, we did not model it because it would not be likely to be presented. Finally, we did not model items that did not have enough surface features to vary or that would produce only a few instances. We did model discrete items with figures to illustrate the feasibility of producing items on-the-fly with dynamically-generated graphical material.

Item models were reviewed by experienced test-development staff, who generated instances to evaluate the equivalence of the items. They evaluated models in terms of the surface variability of the instances and their subjectively estimated difficulty. Models that did not strike a balance of some diversity in surface variability and little spread in difficulty were excluded.

Figure 1 shows a quantitative-comparison item taken from the original PowerPrep pool, while Figure 2 shows the item model that was derived from it. In Figure 2, the components of the item model are identified in bold font. The section labeled "stem" identifies variables **S1.1** and **S1.2**, which represent string variables corresponding to "centimeter" and "kilometer," respectively; **I1** refers to an integer variable. Columns A and B display additional variables. All variables are further defined in the section labeled "variables," which lists the range of values the variables can take on. The variables **I4** and **I5** do not appear in the problem, but rather are needed to specify constraints on **I3**. Figure 3 shows sample instances of the item model depicted in Figure 2.

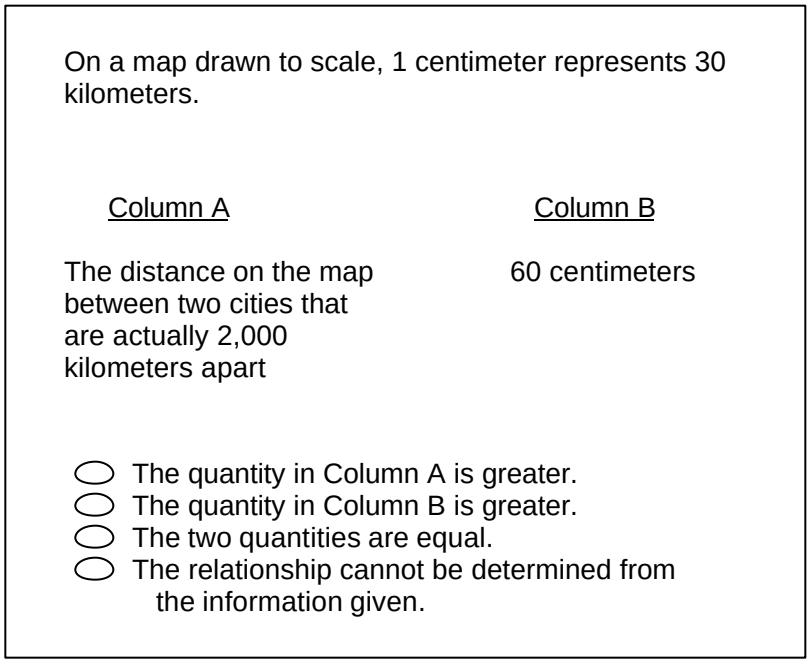

Figure 1. **Sample textual quantitative-comparison item.**

Quantitative-Comparison Model

Stem

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

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

Column B	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: 2-4
I3
I4
I5

Constraints

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

Key

A

Notes:

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.
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**.

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

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.**

Models With Dynamic Figures

For purposes of this investigation, it was important to demonstrate the feasibility of item models that were figural in nature. To accomplish this, the figures in these item models were generated automatically with different instances of the model. Nine item models were chosen to represent a cross section of the types of graphical items that are found in the quantitative section of the GRE General Test. The graphics included tables, number lines, geometric figures, and pie charts. Tables were included because in the current test-creation system, items with tables are produced using labor-intensive artwork. The nine item models included three with tables, two with number lines, three with geometric figures, and one with a pie chart. For example, the Figure 4 shows a base item containing a triangle with two sides labeled.

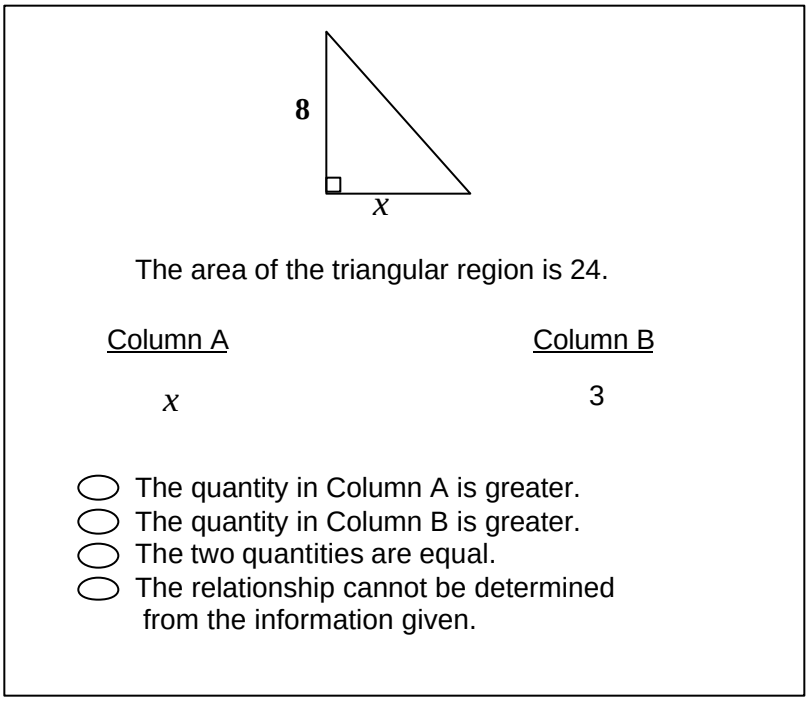

Figure 4. **Sample quantitative-comparison item with figure.**

In the model for this item, the sides of the triangle had varying lengths. A triangle with sides containing the correct ratio of lengths was drawn on-the-fly as part of the process of generating item instances. Figure 5 shows the item model for this item. It should be noted that within the approach we are taking to representing item models and their instantiations, there is little difference between textual and figural items. In particular, as Figure 5 shows, the geometric attributes of the figure are represented as set of variables. Moreover, constraints can operate on those variables. Thus, the rendering of the figure is very much like the rendering of text.

Quantitative-Comparison Model

Stem

The area of the triangular region is **I3**.

Column A	Column A value
S1	

Column B	Column B value
I6	

Variables

I2 Range from 5 - 10 by 1: Length of vertical leg of the triangle
I6 Range from 5 - 10 by 1: ½ of the length of horizontal leg of the triangle
S1 String x, y, w, v, z: Label shown on graph

Constraints

I1 = 2 * **I6**: Length of the horizontal leg of triangle
I3 = **I2** * **I6**: Area of triangle

Key

A

Notes:
1 **I3** is a numeric variable constrained to take on integer values between 25 and 100 and is the area of the triangle.
2 **S1** is a string variable that is the label of the horizontal leg of the triangle.
3 **I6** is a numeric variable constrained to take on integer values between 5 and 10 and equals ½ of the length of the horizontal leg of the triangle.
4 **I2** is an integer value that equals the length of the vertical leg of the triangle.
5 **I1** is an integer that equals the length of the horizontal leg of the triangle.

Figure 5. **Item model for figural quantitative-comparison item depicted in Figure 4.**

Calibration of Item Models: The Expected Response Function

For a generative approach to increase the efficiency of test production, it must allow all instances of an item model to be treated as a single item. In addition to modeling the "content parameters" so that many instances are derived from a single model, we need to model the instances *psychometrically* as well. That is, item models need to be calibrated just as items are. 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, however, it was not feasible to estimate model parameters from scratch. Instead, we modified existing parameter estimates for the items giving rise to each model, but under assumptions of different levels of lack of isomorphicism. The procedure we used for this purpose was the *expected response function*.

Expected response function. The expected response function (ERF) is derived from 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. That is, item parameters are 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. The methodology is directly applicable[1] 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.

Suppose that a given item is calibrated using the three parameter logistic (3PL) model. And let $\beta = (a, b, c)$, the vector of the item parameters corresponding to discrimination, difficulty, and guessing, respectively. Then, two key elements of the ERF approach are the probability response function and the uncertainty distribution concerning the parameters. The probability response function, $P(r \mid \theta, \beta)$, indicates the probability of the observed response, r, conditional on the item parameters (the 3PL) and the subject's ability, θ. The uncertainty distribution about β is $P(\beta \mid \Sigma)$, where Σ is the variance-covariance matrix among item parameter estimates. Using both probabilities, the joint density of r and β is:

$$P(r, ß \mid ?) = P(r \mid ?, ß) P(ß \mid \Sigma) \qquad (1)$$

Finally, the marginal distribution of r can be computed to remove the dependence on the unknown β:

$$P(r \mid ?) = \int P(r \mid ?, ß)\ P(ß \mid S)\ dß \qquad (2)$$

The expression (2) is the ERF. In practice $P(r \mid ?)$ is not evaluated according to the definition (2), but approximated by the closest 3PL curve.

The applicability of the approach for estimating the parameters for an item model is

illustrated in Figure 6. The figure shows several ICCs that vary in difficulty, with each ICC corresponding to an instance from the same hypothetical item model. Computing the ERF is a matter of averaging the ICC over all instances of the item model. That is, averaging the response probabilities at selected values of ?. The resulting vector of averaged probabilities is then fitted to the closest 3PL curve. To the extent that there is lack of isomorphicity, the ERF will tend to have a shallower slope than item model instances. A shallower slope translates into a loss in precision of measurement. Conversely, to the extent that isomorphicism holds, the ICCs will coincide with the slope of the ERF, and there will not be any loss in precision.

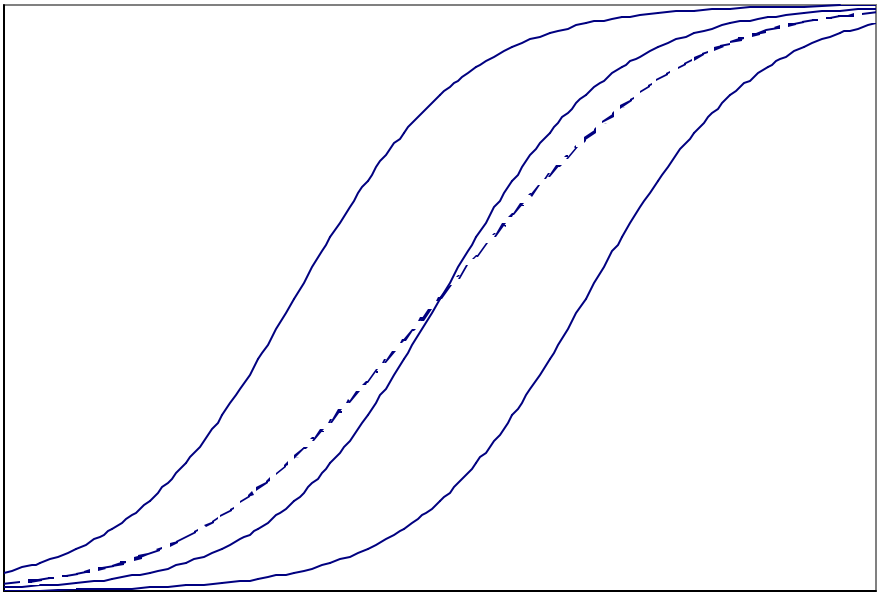

Figure 6. **Graph of expected response function (dashed curve) against three item characteristic curves at three levels of difficulty.**

As noted above, the computation of ERF requires estimates of both β and Σ for each item model. Using these estimates, the computational procedure performs multiple draws from a multivariate normal distribution with S as its covariance matrix and ß 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 couldn't collect the data to derive these estimates empirically for β, we instead used the existing parameter estimates for the 147 items that gave rise to the 147 item models. For S, we located

repeated calibrations *of the same items* from GRE program files from a "linking set" used to scale pretest items.[2] The logic of this choice is that the resulting variability is what would be expected under complete isomorphicity— that is, when the item is the same from time to time. Because such sets are used multiple times, they are recalibrated each time and put on a common scale. For each of these linking items, we computed the variance-covariance matrix among the item 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 (S_1), medium (S_2), and worst-case (S_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:

$$\sum_1 = \begin{array}{c|ccc} & a & b & c \\ a & .003 & .002 & .001 \\ b & & .023 & .011 \\ c & & & .006 \end{array}$$

$$\sum_2 = \begin{array}{c|ccc} & a & b & c \\ a & .012 & .051 & .012 \\ b & & .237 & .054 \\ c & & & .014 \end{array}$$

$$\sum_3 = \begin{array}{c|ccc} & a & b & c \\ a & .015 & .067 & .016 \\ b & & .339 & .081 \\ c & & & .020 \end{array}$$

For each of the 147 item models, we next computed three ERFs, one for each scenario. For any given scenario (e.g., worst case), the *same* covariance matrix was used for all 147 estimates. In a more operational situation, we would associate a different matrix to each item model. However, for any given model, ß was set to the values of *a*, *b*, and *c* associated with the item that gave rise to the model in the first place.

Figures 7 through 9 show the relationship between the original parameter estimates and the attenuated estimates — that is, the estimates computed by the ERF procedure, assuming the worst-case scenario. In Figure 7, we see that, as expected, the *a* estimates are attenuated greatly, indicating that some information will be lost as a result of lack of isomorphicity. An opposite

effect is seen with the *c* estimates in Figure 9: Here, the originally very low *c* estimates are estimated higher after attenuation. Finally, the *b* estimates change very slightly as a result of the application of the ERF procedure.

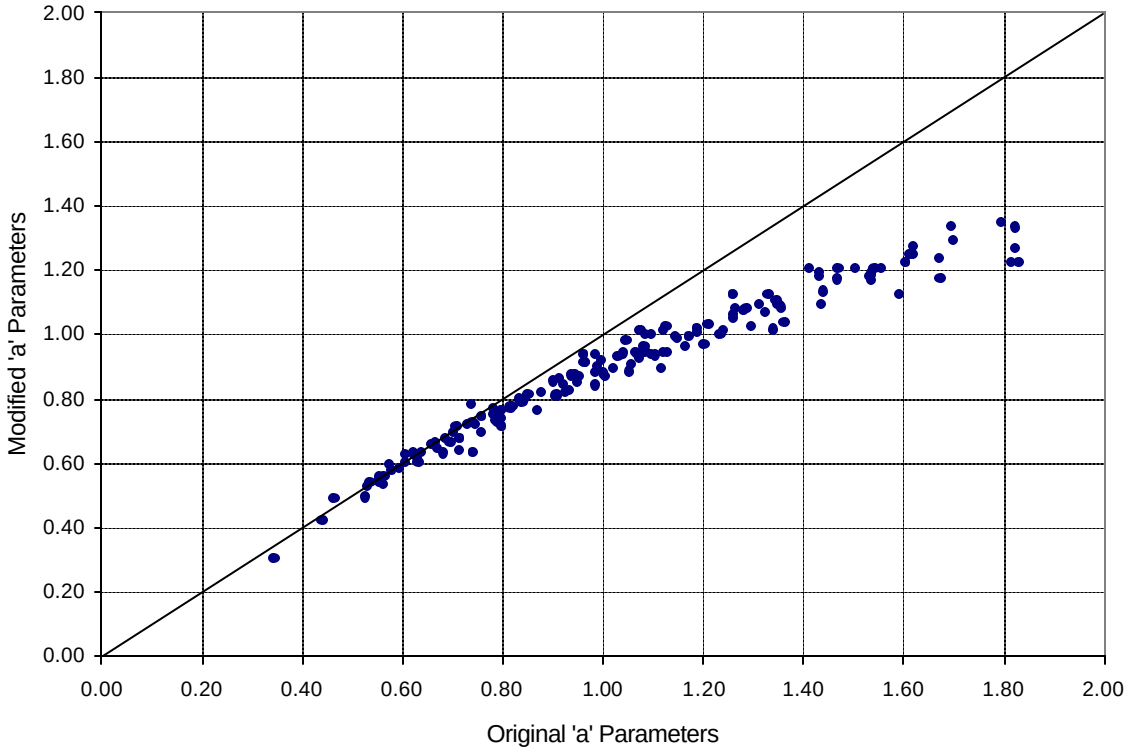

Figure 7. **Plot of original and attenuated *a* parameter estimates based on the worst-case-scenario matrix.**

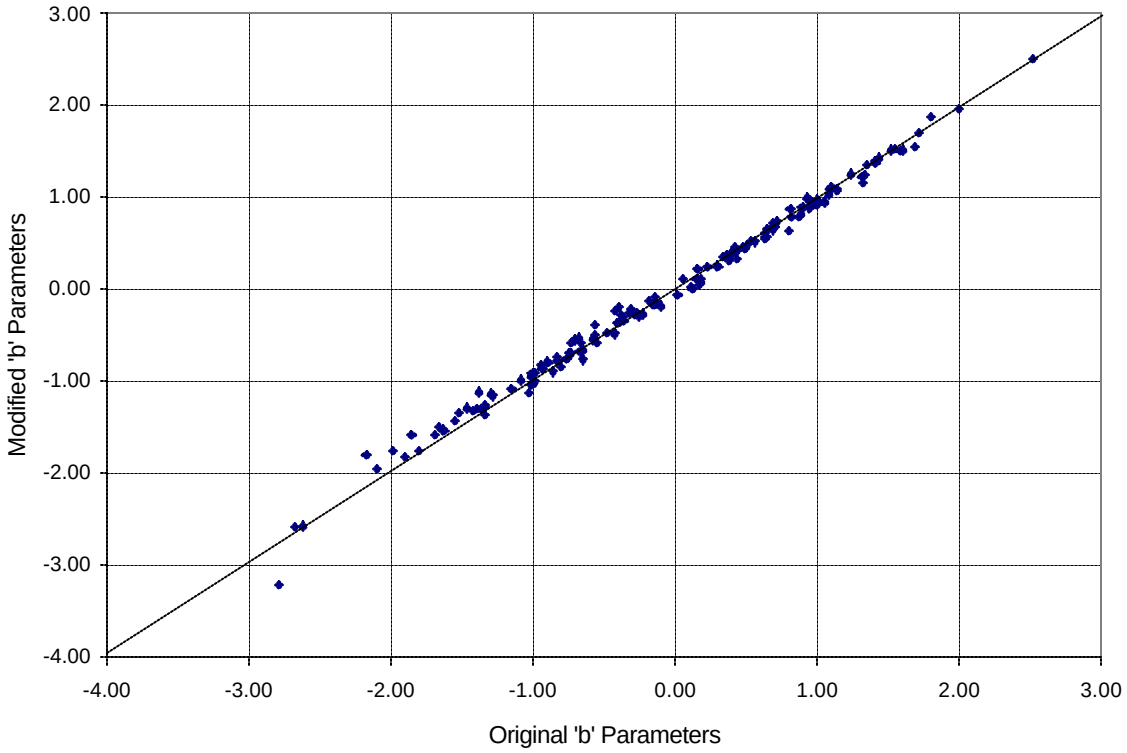

Figure 8. **Plot of original and attenuated *b* parameter estimates based on the worst-case-scenario matrix.**

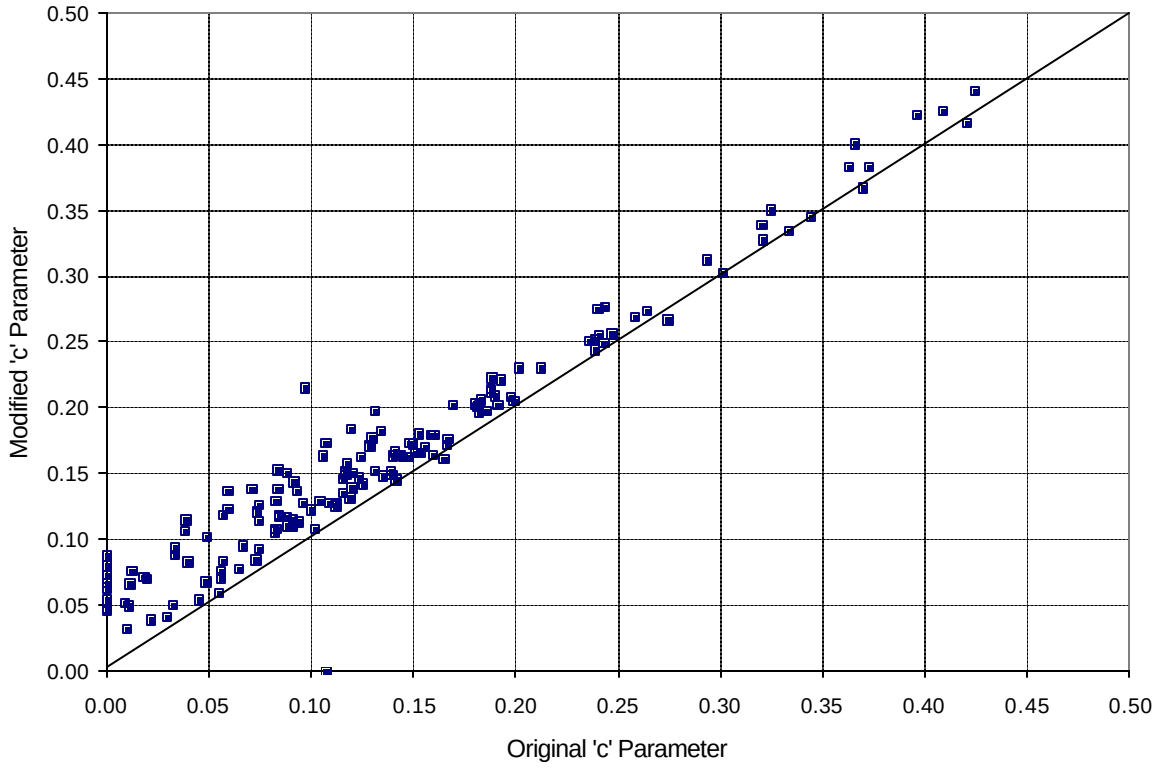

Figure 9. **Plot of original and attenuated *c* parameter estimates based on the worst-case-scenario matrix.**

Simulation Study

To assess the impact of different levels of isomorphicity on score precision, we conducted a simulation study. The program that is used operationally at ETS in the preparation of item pools was modified for this purpose.[3] The modification needed was to simulate the process of randomly assigning an instance of item models to a simulee. This was done in a manner consistent with the code in the ERF program. The procedure, as originally programmed, yielded draws with unrealistically high *c* estimates, so it was therefore modified to pull these down.[4] The modified code was used by the simulation program to draw a single instance for a given item model. The input to the simulation was, thus, the same as it was for the ERF program: a mean vector ß corresponding to the original parameters and a S matrix describing the covariation among *a*, *b*, and *c* under each of the three scenarios. Conceptually, the simulation is described in Figure 10.

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 $\overline{a}, \overline{b}, \overline{c}$ [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.

Figure 10. **Description of simulation procedure with item models and items.**

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 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 static 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 will

be, and therefore, the greater the discrepancy between the ERF and the "true" ICC will be. The greater this discrepancy is, the less information is contributed by the modeled item to theta.

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, and we do so here as well. Figure 11 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 over replicates. 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.

Figure 12 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 isomorphicism is primarily in measurement precision, although the losses at some levels of ability appear to be minimal. This outcome is fortunate, as a loss of precision can be compensated, but bias would be more difficult to correct.

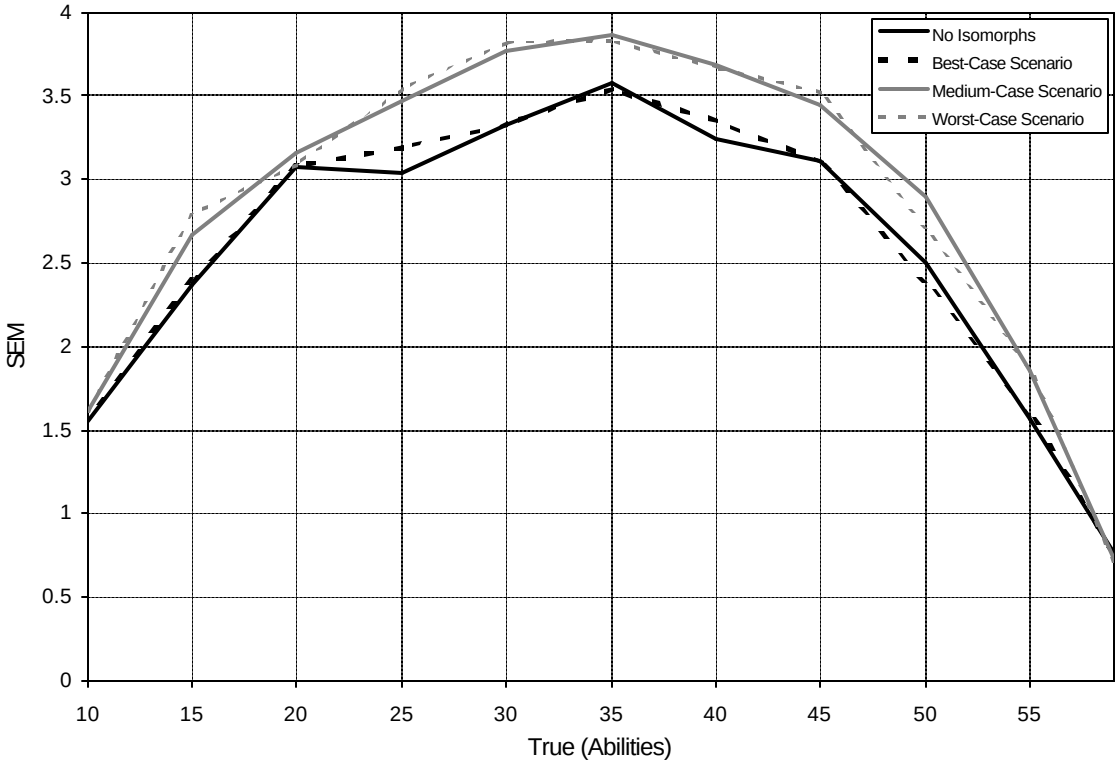

Figure 11. **Standard error for the four testing conditions.**

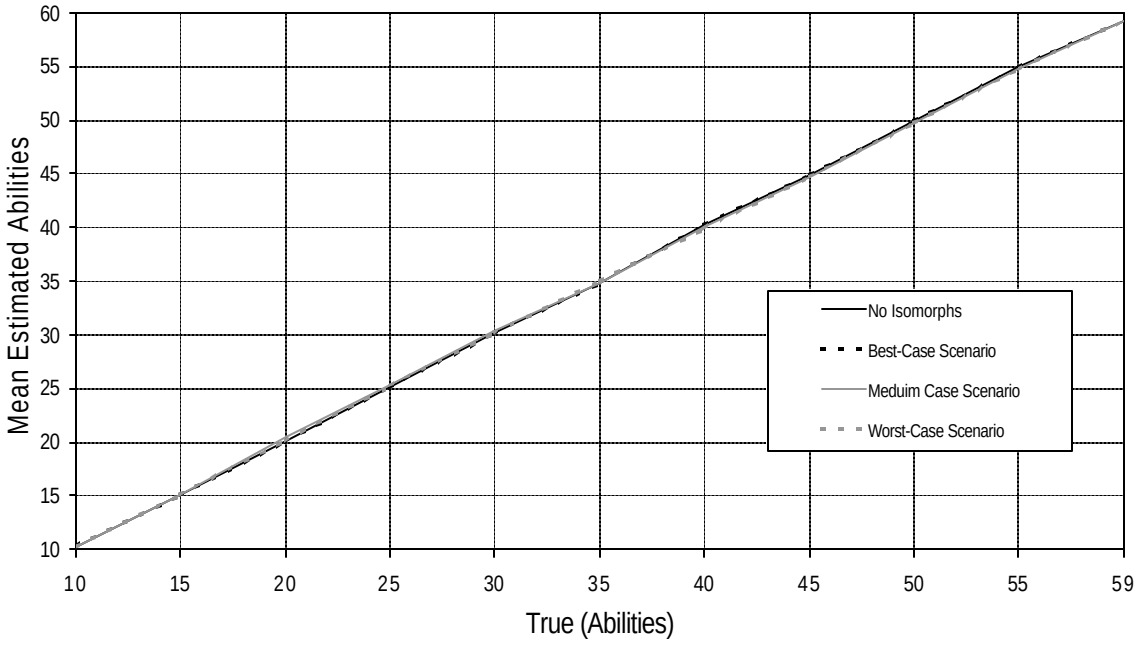

Figure 12. **Estimated versus true ability in four testing conditions.**

Systems Design: From Item Generation to Test Delivery

In this section, we describe the software system we used for data collection. The system is especially important because it is designed to deliver items generated from models as the test is given, or on-the-fly. The key technological contribution of the system was the use of extensible markup language (XML) as a means of representing items and models.

The system we used consists of a test delivery system, an item generation system, and a database of items and item models. The system is Web-based — meaning that the student interacts with the system through an Internet browser that resides on a local computer that in turn interacts with a server by way of the Internet. However, very little computation occurs at the local level. The remote server contains the test delivery and item generation systems, as well as the database of items and models.

The test delivery system manages the interaction with the examinee, decides which item to administer next, and calls the item generation system to instantiate an item model or to retrieve an item from the database. The test delivery system sends a fully formatted item to the browser for display. The browser, in turn, returns response information. The test delivery system scores and records the response and updates the ability estimate. At that point, a new item or item model is selected from the database following an adaptive item-selection algorithm, and the process is repeated until all 28 items have been administered.

Database of Items and Item Models

As noted, the item models, as well as the generated items, are represented in XML.[5] This representation specifies the content of the models and items (e.g., stems, choices, constraints), but *not* how the content should be formatted. Formatting for screen display is done by the test delivery system. Mathematical expressions in the models are represented in mathematics mark-up language (MathML), while graphics are represented by scalable vector graphics (SVG). Because both MathML and SVG are based on XML, the item generation system is able to treat mathematical symbols and graphics as it does text when substituting bound variables. Item models were authored in the mathematics Test Creation Assistant (TCA; Singley & Bennett, 2001), which was extended to export the models into our XML format.[6]

Test Delivery System

The test delivery system we used for this study is a modification of a system that was developed at ETS. To this existing system, we added generation capabilities as well as the capability to deliver adaptive tests. We used the same "adaptive engine" used operationally by ETS. The other elements of the delivery system are a plug-in "scoring engine," an examinee-performance-record database, and a user-interface that very closely mimics the standard, computer-based GRE General Test.

The same code libraries that are used to implement the estimation of ability by maximum likelihood estimation and automatic item selection (AIS) systems in the operational GRE exam were added to the test delivery system. One slight modification was made to these libraries to support using estimated item parameters for items generated from models. Items are assigned unique identifiers called accession numbers. These accession numbers are returned by the AIS system to tell the test delivery system which item to deliver next. Also, they are used by the ability estimation system to retrieve item parameters when determining the current ability estimate. We assigned the same accession numbers to item models that were assigned to the items on which the models were based. This made it possible for the AIS system to function transparently with either items or item models. A further modification was to allow the delivery system to substitute estimated item parameters for an accession number that represented an item model.

In addition to reusing existing systems, it was desirable to reuse items (that were not being replaced by item models in the pool) with as little recoding as possible. A program was written to convert the existing rich-text-format items to XML. This process was only semi-automated and required some manual work. Also, many of the original static items used graphics, including mathematical equations. These graphics were reused untouched, and font faces and sizes were selected to be as close as possible to those used in these graphics.

Item Generation

As explained earlier, the item generation system produces instances from an item model. Also, as shown earlier, item models include variables used within the model, constraints that specify limits of values for some of those variables, and a template for the item into which bound variables are substituted. Variables can be divided into two groups: independent variables, the

values of which are generated at random, and dependent variables, the values of which are derived, by way of constraints, from the values of other variables. Models may specify limits on the permissible values that may be assigned to an independent variable, such as stating that a particular value must be an integer and a multiple of 2 between 0 and 20.

The system we used first generates values for a given model's independent variables. These values are produced by a random number generator, which is seeded with an initial value that is optionally given as a parameter to the model instantiation process. The process guarantees that item instances are reproducible from a given seed value. This is important because if there are problems with a particular item instance, it can be investigated later. This reproducibility method allows the test delivery system to store only the seed value rather than the entire item.

The next step is to derive values for dependent variables. This is accomplished by solving the constraints in the model. An iterative process is used: first, solving constraints that only depend on independent variables, then solving constraints that depend on already bound variables, and so on. The constraints allowed are equalities (a = b + c) and simple inequalities (a ? b + c). In order to simplify the process, all equality constraints are solved first, and then the inequalities are checked to see if they are satisfied. If any are not satisfied, the whole process is attempted again, including independent variable generation.

The final step is instantiating the item template — that is, going from the XML representation to the HTML representation, which is the representation that a browser can understand. First, variables with assigned values are substituted into the item template. This substitution is recursive, so one pass of substitutions can create further targets for substitution on the next pass. Next, the item instance is formatted. For example, the distractors are sorted to meet presentation requirements (usually least to greatest for numeric choices). Figure 13 shows the complete system.

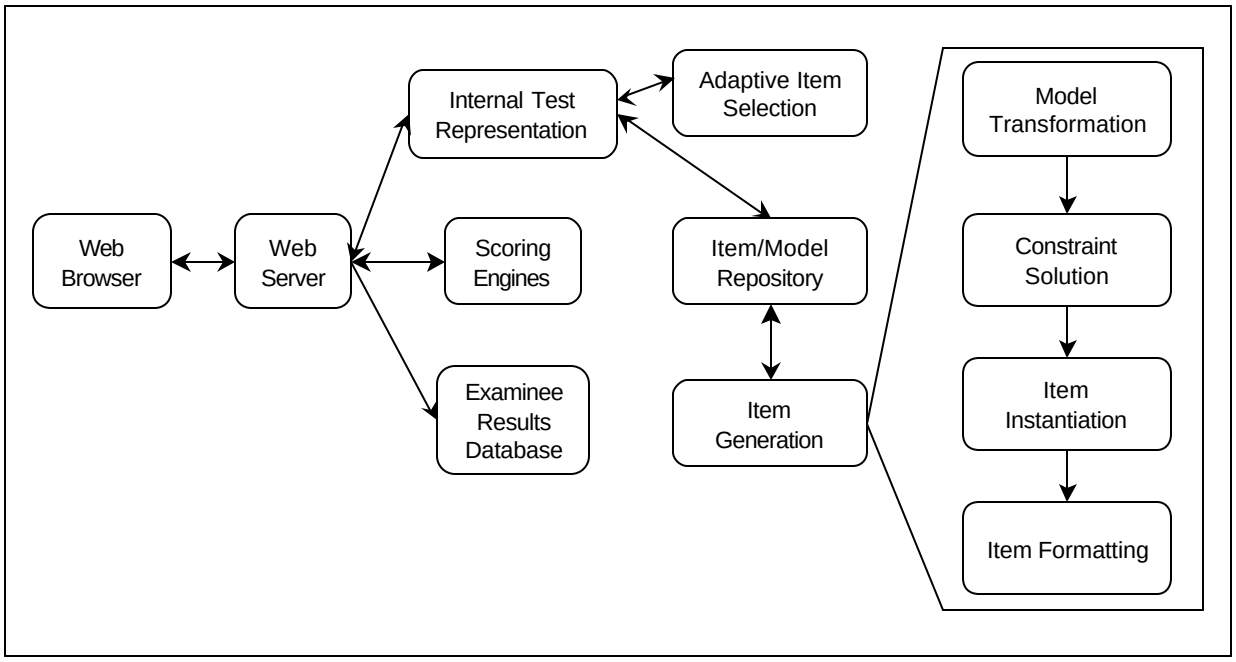

Figure13. **Components of the test delivery system.**

Linear Test Procedures

Linear Test Forms

In addition to our adaptive, on-the-fly version of the quantitative section of the GRE General Test, subjects were administered a 30-item, computer-delivered, linear test consisting entirely of item models as part of the study. The purpose of administering these forms was to assess the level of isomorphicity yielded by item models. Three linear test forms were generated, each comprised of different instances of the same item-models, all identically sequenced. The first 20 items were created from 20 different models, the first 12 of which were quantitative-comparison items, and the last eight of which were problem-solving items. Within each of these groupings, the models were sequenced in order of difficulty from easiest to hardest. The final 10 items were instances generated from 10 of the 20 models — five randomly chosen quantitative-comparison items and five randomly chosen problem-solving items.

Linear test forms were randomly assigned to subjects by imbedding form numbers (that is, Form 1, Form 2, and Form 3) into the subjects' test identification numbers listed on roster sheets. The first test form was assigned to the first test identification number on the roster, the second test form was assigned to the second test identification number, the third test form was

assigned to the third test identification number, then the assignment was repeated for each subsequent grouping of three test identification numbers. When subjects signed in for their test administration, test forms were therefore already randomized. By assigning forms randomly to different subjects, we were able to evaluate the equivalence of different instances of an item model.

Subjects were allowed 45 minutes to complete a linear test. The 20 item models used for the linear forms were not administered as part of the adaptive test. Although these original base items were included in the adaptive pool, they had been flagged to ensure that they would not be administered to an examinee as part of the adaptive test.

Participants and Data

Target population. Two hundred eighty-two subjects were recruited. The target population was comprised of college seniors and first-year graduate students who had taken the GRE General Test between January 1998 and January 2001. Students who had used PowerPrep software for test preparation were initially excluded from participation because the experimental adaptive and linear tests were created from items in the PowerPrep pool. However, this criterion was dropped to maximize the sample size. Of the 282 subjects, 78 (28%) indicated they had used PowerPrep for test preparation.

Recruitment and data collection. We used a variety of methods to recruit participants. Nineteen percent of recruits responded to flyers, and 19% responded to college newspaper advertisements. An additional 26% expressed interest in the study after hearing about it from a friend. The decision to place an advertisement for participants on the GRE Web site (www.GRE.org) proved to be an effective strategy, as 34% of all recruits responded through this method.

All advertising directed interested students to an ETS Web site that displayed general information about the study. Those who visited this Web site were routed to a survey that was linked to a database. Questions on the survey were designed to qualify participants for the study, to secure personal information needed to obtain their prior GRE scores, to request permission to contact them using e-mail, to identify which testing sites they would use, and to obtain contact information. The database was used to send mass e-mails to students who met the participation criteria.

Testing sites and subject payment.[7] Data were collected at Michigan State University (East Lansing), Fordham University (New York City), and CompUSA training centers (in New York City and Philadelphia). We tested 138 subjects in East Lansing, 42 in New York City, and 63 in Philadelphia. At each location, we employed existing computer laboratories. All test administrations were supervised by trained proctors and conducted in computer laboratories reserved solely for this study. Computer laboratories at Michigan State University and Fordham University each contained 20 computers that were available at pre-arranged testing times and dates; CompUSA's training rooms contained 12 computers. In exchange for their participation, students were paid $50 in the form of an Internet gift certificate, redeemable at more than 700 local and Internet merchants. This method of payment proved to be both economical and efficient.

Data elimination. Although we tested 282 students, some data were lost due to unrecoverable computer delivery problems. In some cases, these errors corrupted the scoring records for both the adaptive and linear tests. Proctors recorded these errors on the subject rosters at the testing sites. Examinee performance records (EPRs) were then reviewed to ascertain whether a given subject's adaptive or linear EPR was usable. Inspection of the roster documentation revealed that in many cases linear test scores were intact. In all, only six subjects were eliminated from both test analyses.

Some EPRs were eliminated from the adaptive test analyses because previous operational GRE scores could not be located in the GRE program database. A total of 39 EPRs were disqualified from the adaptive test analyses for this reason, and another five EPRs were disqualified from the linear test analyses.

Demographics. After eliminating subjects due to data problems, data for both the adaptive and linear tests remained intact for 243 participants, and data for the linear test alone remained intact for 277 participants. In tabulating the demographic distribution of our sample, the total number of participants is based on the linear test sample. All other analyses are based on a sample size of 243 for the adaptive test and 277 for the linear test.

Table 2 describes the sample. Males comprised 48% of the study sample, as compared to 35% in the GRE test-taking population. However, the most notable difference between the current sample and 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 2

Demographic Characteristics of Subjects Versus GRE Test-Taking Population

Attribute	Adaptive test n = 243	Linear test n = 277	GRE operational test* (annually)
Gender			
Male	49%	48%	35%
Female	51%	52%	65%
Ethnicity			
Native American or Alaskan Native	1%	1%	1%
Black or African American	4%	4%	9%
Mexican, Mexican American, or 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%
Citizenship Status			
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.

Analysis and Results

Adaptive Test

Our main interest in the simulation study was the comparability of experimental and operational GRE quantitative scores. Comparability is in part a matter of scale: We first sought to determine whether the scores are on a comparable metric. Second, comparability is concerned with the relationship, or ordering, between operational and experimental scores. Thus, we sought to determine how well correlated the operational and experimental scores were.

Table 3 shows the mean scores and standard deviations for study participants on both the operational and adaptive tests as well as for the overall GRE test-taking population. Comparing the mean operational score of our sample, 718, to the mean of 565 for the GRE test-taking

population,[8] 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 operational-score standard deviation for our sample is 88, whereas it is 143 for the GRE test-taking population. Table 3 also indicates that participants' adaptive, on-the-fly GRE scores are lower than their GRE operational scores, and that the variability of the experimental scores is somewhat higher. We expand on this difference in the *Further Analysis of Adaptive Scores* section, below.

Table 3

Operational and Experimental GRE Scores for Study Participants

	Mean	SD
Operational GRE score of sample	718	88
On-the-fly adaptive score of sample	693	101
Operational GRE score for test-taking population	565	143

Note. N = 243

Figure 14 shows the central finding in the study — the relationship between operational and experimental scores. The diagonal line drawn on the figure represents equivalence. If data points were evenly and tightly clustered around this line, it would mean that adaptive scores and operational GRE scores were equivalent. But as the figure shows, and as we know from Table 3, experimental scores were lower. However, the second aspect of comparability — relationship — 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).

Recall that the 28-item experimental adaptive test was composed of both items and item models. Figure 15 shows the number of item models that were administered to subjects. As the figure indicates, no subject's test consisted of fewer than 14 models, and some subjects received as many as 21 models. Thus, an adaptive GRE quantitative section of between 50% and 75% item models was able to order examinees equivalently to the operational test.

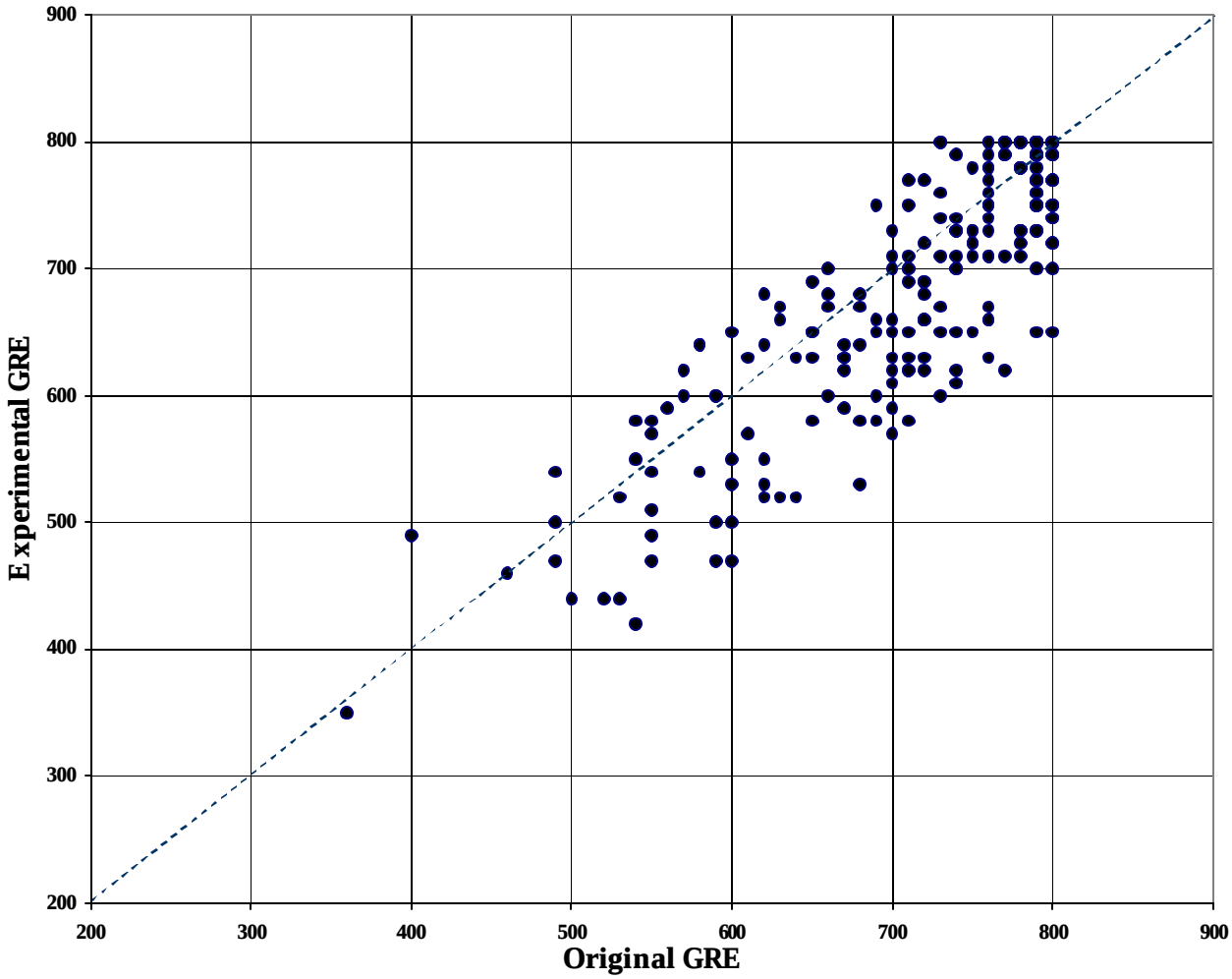

Figure 14. **Plot of operational GRE scores by experimental GRE scores.**

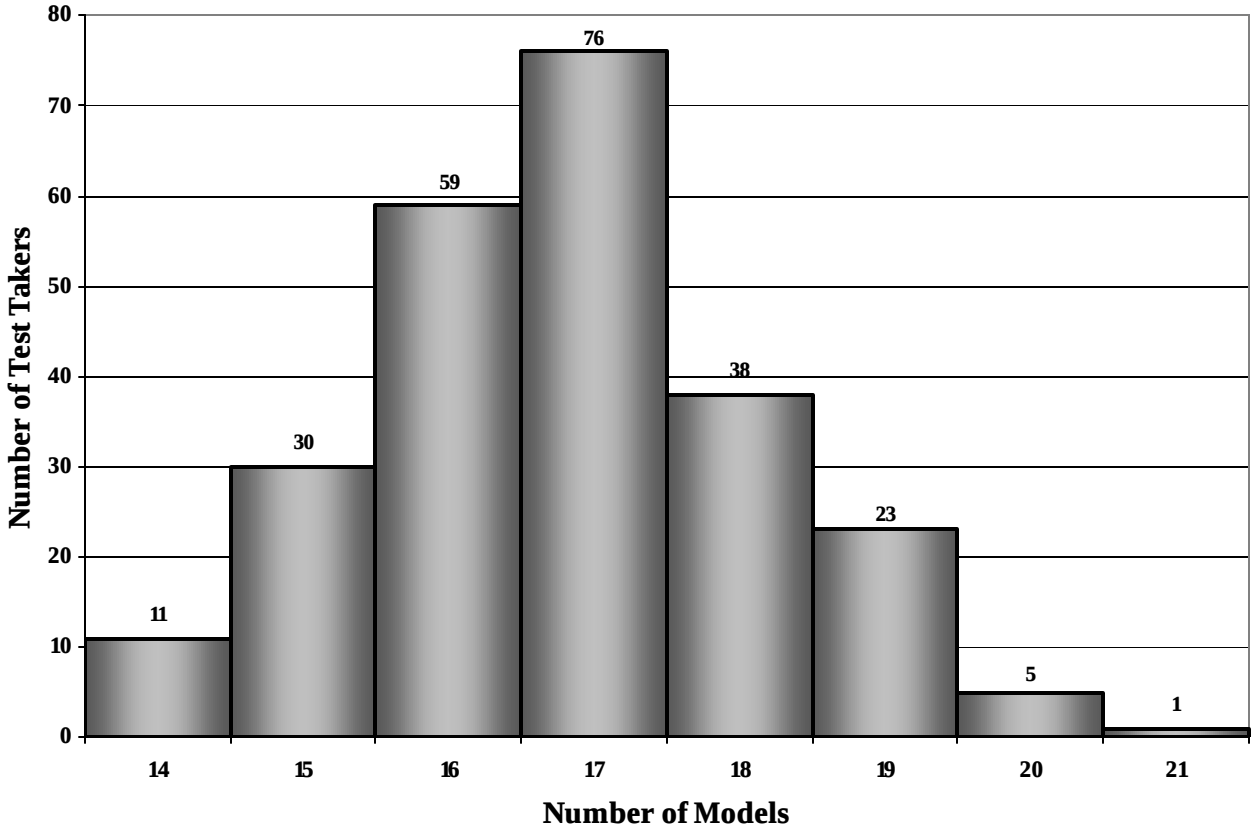

Figure 15. **Number of item models administered to subjects in experimental adaptive test.**

Further Analysis of Adaptive Scores

Although the high correlation with operational 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 very limited. Nevertheless, we examined the responses of students for whom there had been a large change in scores. Differences between operational 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 those scores with changes in excess of 50 points (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 operational GRE score or the mean score.

In summary, the correlation between the operational and experimental scores is as high as the test-retest correlation. The drop in experimental score with respect to the operational score has no obvious or artifactual basis. We believe it is a regression effect — possibly combined with a subtle motivation effect that we have not been able to pinpoint, but that nevertheless could be present.

Linear Test

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 4 shows the correlation among the three sets of model instances and with the operational difficulty estimates from PowerPrep. Correlations with the operational estimates range from .77 to .87; the correlations among the difficulties of the model instances range from .80 to .88. Table 5 displays the corresponding means and standard deviations.

Table 4

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 5

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 16 plots the difficulties associated with each linear test form against the operational difficulty estimates obtained from PowerPrep. The most salient finding is the different scales of the experimental versus operational 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 involved quantitative comparisons and the remaining eight were problem solving items. 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 operational 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 operational items are 3PL *b* estimates. The same pattern is observed for the last eight items. Difficulty estimates obtained for the model instances are closely clustered, as might be expected if the item models were yielding equivalent instances.

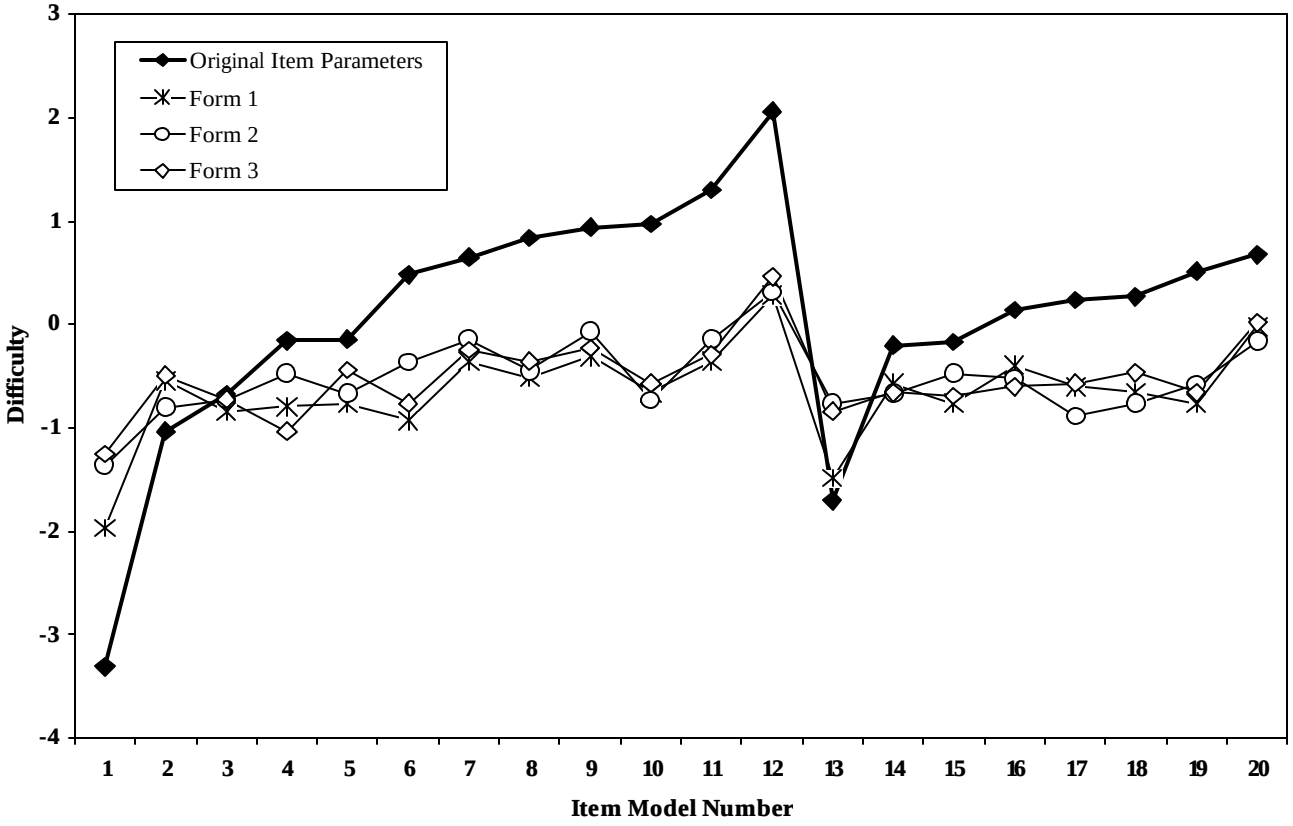

Figure 16. **Comparison of difficulty estimates for PowerPrep and linear forms by position of item on linear form.**

This suggestion of isomorphicism is reinforced by an analysis of response time. Figure 17 shows the mean response time for the 20 model instances in each linear form corresponding to the data shown in Figure 16. (Unfortunately the mean response time for the operational difficulty estimates was not available.) Figure 17 suggests that indeed the 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 analysis of difficulty and response time both suggest that the item models indeed produced isomorphic instances.

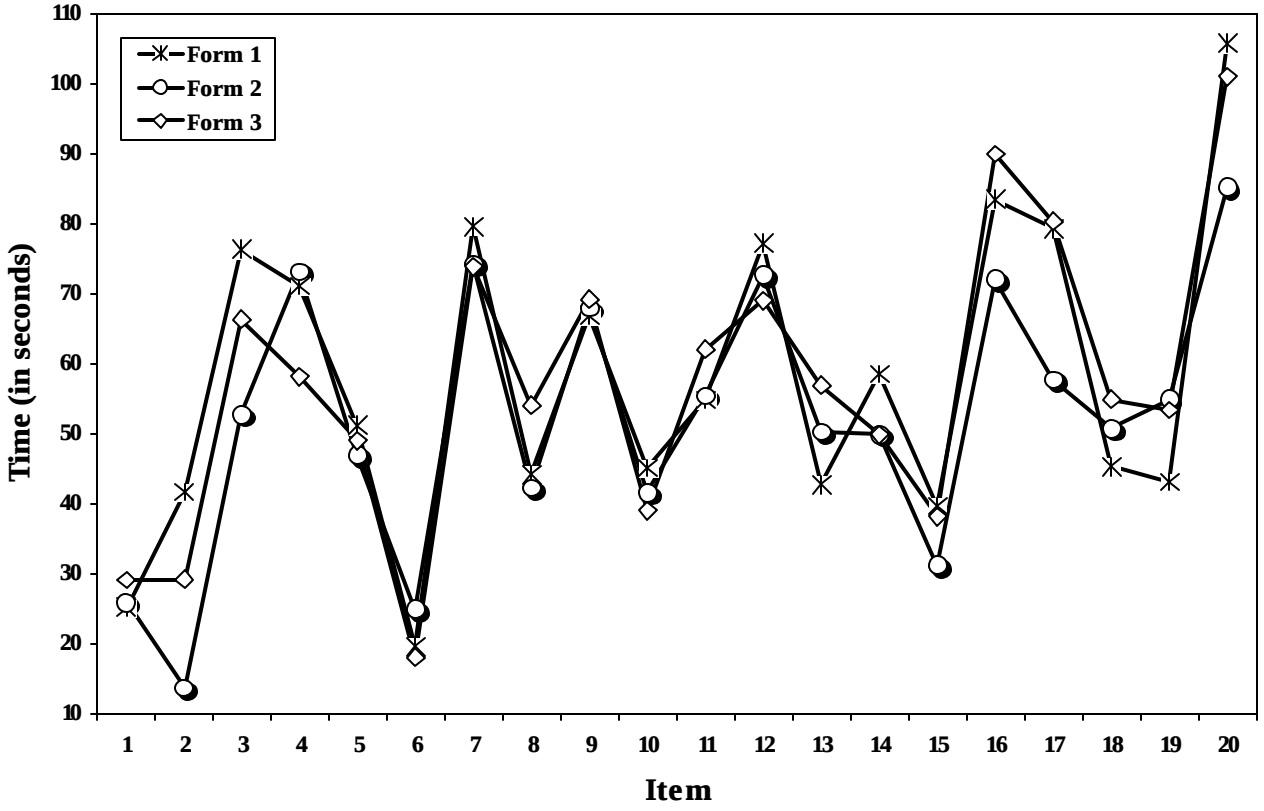

Figure 17. **Mean response time for items in linear forms by serial position.**

Discussion

The results of this study provide initial evidence that an approach to adaptive measurement of quantitative reasoning based on item models is feasible and could prove more efficient and economical than current standard procedures. The cost improvement would not necessarily entail a sacrifice in score precision. The drops in score precision we observed with simulated data can, in principle, be compensated with a slight lengthening of the test. Moreover, our highly selective sample, the actual high score correlation with operational GRE scores, and the consideration that nearly half of the items were from item models suggest that, in reality, measurement quality is not negatively affected under an adaptive, generative model. Here we discuss the nature of additional evidence needed to corroborate our main conclusion that an adaptive, generative model is a technically feasible and cost-effective approach to admissions testing.

Item models by themselves are not likely to resolve the challenges facing computer-based testing. A fundamental problem is the content of the vat, or repository, of all test content for a given measure. In the ideal case, all elements of the vat, whether they are items or item models, are exposed in an even fashion. Studies are underway to study the feasibility of reconfiguring the current item vat to achieve the "equal exposure" ideal. If it is concluded that the current vat can, for the most part, be reconfigured as item models that would be exposed evenly, existing item parameter estimates can be used to bootstrap the calibration of those item models. This result would make a generative, adaptive approach to assessing quantitative reasoning as part of the GRE General Test more practical.

A related concern is the similarity of instances that are generated from a model. In an operational program, large numbers of highly similar items pose a clear security risk. If examinees regularly see items on practice tests and then encounter their isomorphs on the operational exam, item parameters may change and validity may be compromised. This risk increases to the extent that items are more similar within than across models and to the extent that the number of models is small. The former condition will hold, since it is the similarity of items within models that allows generation with calibration. The latter condition is modifiable: We can create many models, but we must be careful not to create so many that the costs of model creation exceed the costs we now incur in writing items individually.

Several studies are underway to address the foregoing concerns. One study considers the extent to which examinee scores are impacted by the interleaving of surface features and deep structure of a given item model. Another project explores different structures for the GRE quantitative reasoning vat, and still another investigates how vat management and item selection procedures might need to change to accommodate the item modeling approach. A fourth study is attempting to create methods based on cognitive principles for calibrating items that vary widely in their mathematics and surface features.

A further major issue relates to model calibration and the effects of variation in parameter estimates on examinee scores. We described one approach to calibration — ERF — and explored its impact on scores with very promising results. However, ERF is only one potential approach, and our exploration of it was restricted to a single item pool. As a means of providing a baseline for considering the amount of error tolerable in model calibration, Rizavi, Way, Davey, and Herbert (2002) are examining the variation in item parameter estimates that occurs over repeated

uses of the same GRE items. Yu, Sclan, and Way (2002) are examining the psychometric basis for ERF as a model calibration method. Similarly, Johnson (2002) is exploring the use of hierarchical methods for model calibration. Finally, Williamson and Bejar (2002) are investigating the use of testlet theory to evaluate the equivalence of automatically generated multiple-choice items. These methods are promising because unlike ERF, they formally capture variation in the parameters of instances from a model.

A final issue concerns the tools available for item modeling. We have in the existing Math Test Creation Assistant a tool capable of generating a wide variety of items from models. To be used operationally on a large scale, the tool will need to be more closely integrated with the evolving ETS production system. Consequently, we are a) recasting the Test Creation Assistant as a set of components that are compatible with the production authoring and delivery environment, b) replacing the existing constraint solver with a more robust generation engine, and c) incorporating linguistic capabilities for generating syntactically correct items.

While practical feasibility is an appropriate concern, it may be equally important that, from a theoretical perspective, item models enhance the validity argument in support of test scores. By designing a test with item models, we are helping to build validity into the scores. The design of item models encourages taking advantage of the cognition of the construct under measurement. Once we have incorporated theoretical knowledge into the item model, its use represents a test of that knowledge. Specifically, if isomorphicity does not hold, an investigation of the reasons is bound to serve as refinement of the underlying theoretical basis. If isomorphicity holds, the underlying theoretical basis is further supported.

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 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 isomorphicity, there was no bias, but precision of measurement was eroded, especially in the middle range of the true-score scale. We feel that much more extensive simulations need to be done to better understand the impact of item models.

We next presented results from a field study in which we administered an experimental, on-the-fly, adaptive quantitative-reasoning test as well as a 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 isomorphicity 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 operational GRE scores were extremely reassuring. The correlation of experimentally obtained scores and operational GRE scores was .87 — as high as can be expected because it matches the test-retest correlation observed under operational conditions. 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 isomorphicity across items within models. This high level of isomorphicity is likely the reason we obtained a correlation with operational scores that was indistinguishable from operational test-retest correlations.

As discussed earlier, some of the work that remains to be done to transition to an operational on-the-fly approach presents significant challenges that do not seem insurmountable. We conclude that the current GRE-funded study provides a promising first step toward what we hope will be significant cost and theoretical improvement in test creation methodology for educational assessment.

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

Bennett, R. E. (in preparation). *Automatic item generation: An overview*. Princeton, NJ: Educational Testing Service.

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.

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.

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., & Dresher, 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. (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.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

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: The 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, 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. (2000). *Leverage points for improving educational assessment*. Princeton, NJ: Educational Testing Service.

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.

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.

Schaeffer, G., Bridgeman, B., Golub-Smith, M., Lewis, C., Potenza, M., & Steffen, M. (1997). *Comparability of paper-and-pencil and computer adaptive test scores on the GRE General Test* (GRE Board Professional Report No. 95-08). 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. (1935). The generic nature of the concepts of stimulus and response. *Journal of General Psychology, 12,* 40-65.

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.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 163-182). Dordrecht, Netherlands: Kluwer Academic Publishers.

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. (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.

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 develop*ment (pp. 277-286). Mahwah, NJ: Lawrence Erlbaum.

Yu, F., Sclan, A., & Way, W. D. (2002, April). *Item calibration using expected response functions*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Notes

[1] We are grateful to Robert Mislevy for pointing out the applicability of this methodology to item modeling.

[2] We are grateful to Robert Smith for providing the data and to Marilyn Wingersky for enhancing the original program for this project.

[3] We are grateful to Martha Stocking for modifying the operational simulation program to include item modeling.

[4] We are grateful to Robert Mislevy and Marilyn Wingersky for devising the procedure.

[5] Information about extensible markup language can be found at: www.w3.org/XML/

[6] The Test Creation Assistant (TCA) was developed prior to this project with the purpose of enabling test developers to produce "variants" from what is, in effect, an item model. However, in operational use, the models in TCA were meant to produce instances rather than having an existence of their own. In this project we used TCA as a means of fine-tuning an item model, at which point it is exported to be used as an autonomous producer of items. The model, rather than the instances of the model, are the focus of interest.

[7] We are grateful to Mitch Rabinowitz of Fordham University for arranging the use of the lab at Fordham and for supervising several proctors. We are also grateful to Ed Wolfe of Michigan State University for arranging data collection and supervising proctors at Michigan State University.

[8] 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.

54020-009051 • Y112E.2 • Printed in U.S.A.