ACT Research Report Series 97-4

Realistic Simulation of Item Response Data

Tim Davey

Michael L. Nering

Tony Thompson

ACT

July 1997

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Realistic Simulation of Item Response Data

Tim Davey
Michael L. Nering
Tony Thompson

Abstract

Potential solutions to applied measurement problems are commonly evaluated using simulated data, largely because it is often impossible, difficult or prohibitively expensive to supply the needed volumes of real data. However, answers proved by simulation studies generalize to real-world problems only to the extent that the simulated data resembles the real thing. We have discovered that in the case of generating item responses using traditional undimensional, logistic response models this resemblance is less than complete. While gross features of the data such as item passing rates, score distributions and test reliabilities are modeled adequately, subtler characteristics of real data are often absent in simulation. Accordingly, we have developed high-dimensional simulation procedures that preserves these finer characteristics, and these procedures will be described along with several empirical examples.

Realistic Simulation Procedures for Item Response Data

1. Motivation

An important feature of latent trait models is the convenient process they offer for generating simulated item response data. As a consequence, research studies based on artificial data have proliferated in the psychometric literature, particularly as computers have become powerful, inexpensive, and widely available. The advantages of simulation studies are substantial and firmly established (see Harwell, Stone, Hsu, & Kirisci, 1996; Spence, 1983). Chief among these is that the true latent characteristics of simulated examinees and test items are known to the researcher. After all, it is the researcher who has specified them. This allows estimated item and person parameters to be compared to their true values, which are by definition unknown and unobservable from real data.

Simulation studies also permit theoretical results to be confirmed in practice. Knowing that an estimation procedure is asymptotically efficient is of limited value in a world full of finite test lengths and examinee samples. Showing that the procedure is effective long before asymptotic results apply is therefore of considerable practical importance. Simulation also makes possible certain types of studies that are otherwise difficult to conduct. For example, the memory of a simulated examinee can be easily erased to eliminate the contaminating effects of a pretest on a postest. Comparable studies with college students would almost certainly run afoul of university human subjects guidelines. Finally, it is faster, cheaper and easier to simulate data than to collect it from live examinees. This allows researchers to evaluate new psychometric models and procedures almost immediately and practically for free.

The hazards inherent in simulation studies are equally well known but tend to be less well publicized. The implicit disclaimer that should be attached to each study would state that the

reported results generalize only to the extent that the simulation procedures produce data that are similar to the actual responses of actual examinees to actual test items. We argue here that in many cases the resemblance of simulated data to the genuine article is far short of complete, and that subtle, complex response behaviors characteristic of real data are not well reproduced by typical simulation methods. Our purpose is to describe and evaluate alternative procedures for generating more realistic item response data.

2. A Brief History of Simulation

Simulation studies predate the advent of high-speed computing, although early efforts were somewhat rudimentary. Pioneers of the method drew numbered balls or slips of paper from jars, or built analog computing devices. Most of the work concentrated either on confirming theoretical results, or on producing instructional demonstrations. An example of the former was W. S. Gosset's use of random sampling to help establish the distribution of the correlation coefficient (Student, 1908). Gosset later used similar procedures to bolster a somewhat shaky theoretical derivation of the t-distribution. Instructional examples include the building of Galton's "quincunx", a device used to randomly generate binomial distributions. Balls dropped into the top of this device were deflected by a series of pins on their trip through to the bottom where they settled into a series of bins.

Simulation methods were formalized in the 1940s by members of the Manhattan Project working to develop the atomic bomb. They used simulations and random sampling to model physical processes, approximate intractable integrations, and refine experimental designs. These methods were extended to a wide variety of fields throughout the 1950s and 1960s as digital computers became more capable and available. Problems in statistical mechanics, radiation,

transportation, logistics, and economics were all approached through what were by then termed *monte carlo* methods. For a time it appeared as though the very popularity of the approach would prove its downfall, as researchers reported discouraging attempts to apply monte carlo methods to problems for which it was ill-suited, inefficient, or just plain inaccurate (Hammersley & Handscomb, 1964). However, in time, the strengths and limitations of simulation methods were identified and the viability of the approach was insured.

Simulation studies began appearing with some regularity in the psychometric literature in the 1960s. Researchers examined the recovery of underlying structure by competing factor extraction and rotation methods, investigated the adequacy and efficiency of estimation procedures, and checked the robustness of various statistical tests to violations of underlying assumptions. The best of these studies were careful to generate data by realistic processes. A notable example was a study by Tucker, Koopman, and Linn (1969) that investigated the recovery of underlying factor structures from simulated correlation matrices. The sampled matrices were not simply drawn from the cleanly defined and parsimonious common-factor structure designed to underlie the correlations (the "formal model"). Instead, dozens of minor common and unique factors influenced the sampling as well (the "simulation model"). What resulted were correlation matrices that contained a realistic level of "noise" that served to partially mask the underlying structure. Although here termed noise, lack of fit to a simple underlying structure does not necessarily indicate that purely random process are at work. Rather, noisy data result from a myriad of subtle, complex, and unmodeled human behaviors.

Simulation of item responses, rather than correlation matrices, became more prevalent as latent trait models came into common use. Some of the earliest latent-trait based simulations

were conducted by Hambleton (1969) and Panchapakesan (1969). Both of these authors investigated the recovery of generating item parameters by the estimation methods available at the time. The Hambleton study paralleled Tucker, Koopman, and Linn (1969) in that data were simulated by models other than those fit during estimation. This was done to examine the robustness of item parameter estimates to serious violations of model assumptions.

In 1991 the Office of Naval Research sponsored a study to investigate procedures for calibrating pretest items embedded in adaptive tests (Holland & Wingersky, 1991). This simulation was termed "high fidelity", with item responses generated from nonlogistic response functions obtained by fitting nonparametric latent trait models to large samples of real data. Data were generated from nonlogistic functions in an attempt to emulate the subtle and complex response behaviors evident in real data. However, the simulation methods remained highly idealized in that responses were still based on a unidimensional underlying model.

Harwell et al. (1996) recently summarized and evaluated the uses of simulation methods in psychometric research. Their conclusions, many of which are echoed here, include the following:

1. Simulations should be the last, rather than the first resort. All viable analytical or empirical options should be exhausted before simulation studies are even considered. Simulations should be used only if a research question cannot reasonably be answered by any other means.

2. Good simulation procedures will not compensate for faulty experimental design, sloppy execution or inept data analysis or interpretation (see Spence, 1983). The same basic rules that apply to laboratory experimentation apply with equal force to simulation studies.

3. Simulations are useful only to the extent that they reflect reality. While it is necessary that simulation models make reasonable assumptions and generate data through plausible processes, this is not sufficient to guarantee realistic outcomes. Simulation models must also be based on realistic underlying parameters. Even the best simulation models are only as good as the parameters that form their foundation.

The remainder of this paper will focus on the third point above.

3. Typical IRT Simulation Procedures

The typical latent trait-based simulation study requires that three basic decisions be made:

1. Specify the form of the item response model. Item responses are most typically generated from unidimensional, conditionally independent, logistic or normal ogive models. Common choices are the one- two- and three-parameter logistic models, where probabilities of correct response are given as a function of both examinee ability and the item's operating characteristics.

2. Specify the parameters of the item response model. Item response models are flexible enough to model test items with widely varying operating characteristics. The measurement properties of a simulated test are therefore not fixed until the parameters defining the test's items are specified. This is usually done in either of two ways: 1) item parameters are randomly drawn from probability distributions, or 2) parameters are selected from those estimated from actual items administered to actual examinees.

3. Specify the form of the examinee ability distribution. Examinee abilities are usually simulated by random draws from a specified population distribution, most often the standard normal distribution.

Item responses are simulated by first randomly drawing an examinee ability parameter from the population distribution. The item response model and the specified item parameters are then used to compute the probability of an examinee with the sampled ability responding correctly to each test item. Each item response probability is next compared to a random deviate drawn from a uniform distribution supported on the unit interval, (0,1). The response is then coded as correct if the uniform deviate is less than the item response probability; the response is scored as incorrect if the deviate exceeds the probability.

When data are simulated from unidimensional, logistic item response models, three important assumptions are inherent in the generation process. Each can be, and often is, untenable to some degree. First, the regressions of item scores on ability are assumed truly logistic. As a

consequence, response functions are constrained to increase monotonically with increasing ability. However, items with nonmonotonic response functions are routinely found with actual data (Levine, 1984; Samejima, 1979). A classic example is a multiple-choice item that is difficult mainly due to the presence of an attractive but incorrect response alternative. Very low ability examinees will lack the knowledge necessary to find any of the alternatives especially appealing, and will therefore respond essentially at random, answering correctly occasionally by chance. However, moderately able examinees might know just enough that they find an attractive but wrong alternative all but irresistible. These examinees will then respond correctly at much lower than chance rates, causing the response function to decline briefly before reascending at higher ability levels where examinees are able to properly reject the distracting alternative.

A second major assumption of typical simulation models is that item responses are fully determined by the examinee's standing on the same, single ability dimension. The simulated data are therefore strictly unidimensional, having a factor structure that features a single common factor. This assumption is also nearly universally violated with real data, as strict unidimensionality is the rare exception rather than the rule. While analyses of most well constructed cognitive tests usually reveal a single, dominant common factor or ability dimension, it typically accounts for well less than half of the score variance. The remainder is attributed to unique factors and to numerous minor factors or traits that influence small clusters of items. Although unidimensional item response models provide a useful approximation to the structure of most tests, they should not be considered as representations of reality.

The third assumption inherent in all simulations is that the item parameters used to generate the response data resemble those likely to be found with an actual test. While this assumption

is certainly reasonable when the item parameters have been estimated from real data, it is less defensible when the item parameters have been randomly generated. For example, given that most real tests are constructed very systematically, it is unclear what sort of test is being simulated when item difficulty parameters are specified by random draws from a standard normal or uniform distribution.

Despite these often unrealistic assumptions, data simulated by unidimensional models reproduce the observed characteristics of actual data in several important ways. Item passing rates, item-test score correlations, number-right score distributions, and test reliabilities all generally correspond closely when simulated data are compared with the real data on which simulating parameters were based. However, unidimensional simulation procedures may prove less satisfactory when examined in greater detail. Interitem correlations found in real data may not be well approximated by a unidimensional simulation. Similarly, the factor structure of data simulated unidimensionally may be much cleaner than is typical of real data. Complex interactions between items and examinees may not be present. The danger is that these differences may lead to certain procedures evaluated as effective with simulated data performing much less well when applied to real data.

4. Realistic Simulation Procedures

Our design for realistic simulation incorporates the same reasoning that motivated Tucker, Koopman, and Linn (1969). The simulation model therefore includes not only the major dimension or dimensions that provide the basic structure for the test, but also includes the numerous minor dimensions that are characteristic of actual data. The process begins by fitting

a multidimensional latent trait (MIRT) model to a large sample of actual data. This can be done with any of several software packages (e.g., Fraser, 1986; Wilson, Wood & Gibbons, 1991). No attempt is made to interpret the resulting solution by rotation to simple structure. The fitted model is simply treated as a template from which new data can be generated.

The estimated multidimensional item response functions are used to generate data by procedures directly analogous to those used with unidimensional simulations. The important difference is that several ability parameters are generated for each simulated examinee, all of which influence each item response. The item and ability parameters combine to produce probabilities of correct responses, just as they do with simpler unidimensional models. Data are generated by comparing these probabilities to random draws from a uniform distribution.

Although multidimensional latent trait models make the same sort of assumptions made by unidimensional models, they are far less restrictive and hence less likely to be seriously violated by real data. For example, while item response surfaces are assumed to be monotone along each of multiple ability dimensions, discrimination parameters are not constrained to be positive. The regressions of item responses on any unidimensional ability composite consequently are not required to be either logistic or monotonically increasing. Multidimensional models therefore are able to approximate closely the nonlogistic unidimensional response functions commonly found in real data. Multidimensional models are also by definition well suited to characterizing the multidimensionality, local item dependence, and general "noisiness" that typifies real data.

5. Empirical Examples

The results of simulating item response data using multidimensional latent trait models are demonstrated below with two empirical examples. In both cases, data were also simulated in the more typical way from unidimensional models in order to provide a basis of comparison and to illustrate the drawbacks of the simpler approach. All simulations, multidimensional and unidimensional, were based on model parameters estimated from samples of 10,000 actual examinees responding to actual test items. Two different tests served as simulation templates, the first measuring mathematics and the second English usage.

The Mathematics test has 60 items designed to assess the skills that students have typically acquired in courses taken up to the beginning of grade 12. The test consists of multiple-choice items that require students to use their reasoning skills to solve practical problems in mathematics. The problems assume knowledge of basic formulas and computational skills, but do not require complex formulas and extensive computation. Material covered on the test emphasizes the major content areas that are prerequisite to successful performance in entry-level courses in college mathematics. Items therefore range from basic operations with integers all the way up through application of trigonometric identities. The test is carefully constructed according to detailed content and statistical specifications. Previous research has found that while the test effectively measures a unitary trait, it does contain distinct clusters of items measuring various narrower content domains (Miller & Hirsch, 1992).

The English has 75 items that measures understanding of the conventions of standard written English (punctuation, grammar and usage, and sentence structure) and of rhetorical skills (strategy, organization, and style). The test consists of five prose passages, each attached to a

series of multiple-choice items. As with Math, detailed analyses of the English Test have shown it to have a dominant primary dimension, but to also include numerous clusters of more tightly focused items. The passage structure of the test is also generally clearly revealed by dimensional analyses.

Simulation Methods

The first step in simulating realistic response data was to fit high-dimensional latent trait models to large samples of real data. This was done using the calibration program NOHARM (Fraser, 1986), which estimates a multidimensional normal ogive model. The fitted model defines the probability of a correct response to item i by examinee j with vector of abilities $\underline{\theta}_j = \langle\theta_{j1}, \theta_{j2},...,\theta_{jm}\rangle$ by the item response function:

$$Prob\left(u_{ij}=1 \mid \underline{\theta}_j, \underline{a}_i, d_i, c_i\right) = P_i(\underline{\theta}_j) = c_i + (1-c_i)\,\Phi\left(\underline{a}_i^T \underline{\theta}_j + d_i\right) \quad (1)$$

where $\Phi(\bullet)$ is the normal distribution function, d_i indexes item difficulty, the vectors $\underline{a}_i$ and $\underline{\theta}_j$ characterize item discrimination and examinee ability on each of the m latent dimensions, and c_i is a scalar representing the lower asymptote, or the probability of a correct response to the item by an examinee with extremely low ability. Note that the unidimensional normal ogive response function is obtained as a special case of Equation 1 where both a_i and θ_j are scalars. NOHARM item parameters are scaled relative to standardized, uncorrelated, normally distributed ability dimensions, that is, relative to a $N(\underline{0},\mathbf{I})$ distribution of ability.

NOHARM does not estimate c parameters, so these must be obtained elsewhere and input to the program. Because c parameters are theoretically unaffected by the dimensionality of the fitted model, unidimensional calibration programs such as BILOG (Mislevy & Bock, 1989) are

a convenient source of *c* parameter estimates. As a first step, separate BILOG calibrations were therefore obtained from large examinee samples from both the Math and English tests. In addition to the *c* estimates needed for NOHARM, these calibrations also provided the model parameters from which unidimensional data would subsequently be simulated.

High-dimensional simulations were based on ten-dimensional NOHARM solutions fit to the same large data sets used with BILOG. That ten dimensions were fit rather than two, five or twenty was largely an arbitrary decision. Determining the most "proper" dimensionality of the fitted MIRT model remains a fairly subjective process. Decades of research on this problem in its factor analytic context have produced dozens of procedures, indices and rules-of-thumb, none of which are universally approved of or followed.

Fortunately, our situation is a bit simpler. Because we are not concerned with interpretation of dimensions, the normal laws of parsimony do not directly apply. Furthermore, given the complexity inherent in real data, we can afford to fit fairly high dimensional models with only minor concerns of overfitting or "capitalizing on chance." There will almost always be more replicable structure in the data than can be reasonably modeled. Experience has so far taught us that, by and large, more dimensions seem preferable to fewer. The greater danger is in underfitting and failing to account for numerous subtle but important underlying characteristics of real data.

Practical concerns aside, the degree to which simulated data resemble real tends to increase with increasing dimensionality. Concerns include such things as calibration program limits, convergence problems, and the general unwieldiness of very high dimensional models. Ten dimensions simply seemed a reasonable compromise between too few dimensions and too many

complications. What seems a promising, if no more objective, approach to determining a proper dimensionality is presented in the conclusion.

Once the unidimensional and multidimensional simulation models were established, data generation proceeded similarly under both approaches. Ability parameters were sampled from the standard normal distribution in the unidimensional case, and from a 10-variate $N(\underline{0},\mathbf{I})$ in the multidimensional case. The probabilities of simulated examinees answering each item correctly were computed by the appropriate form of Equation 1, either unidimensionally or multidimensionally, and compared with random draws from a uniform distribution supported on the unit (0,1) interval. Item responses were coded as correct if the response probability exceeded the uniform random deviate and incorrect otherwise.

These procedures were used to generate a total of four artificial data sets of 10,000 examinees each: unidimensional and ten-dimensional simulations for both for the Math and English tests. Real datasets, also with 10,000 examinees were available on both tests. The quality of the simulation procedures was next assessed by comparing observable features of each artificial data set to those of the appropriate real dataset. The following sections describe the features that were compared and the results of that comparison.

Comparison of Simulated and Real Datasets

Simulated data sets were compared to real data according to four criteria: (1) classical item analysis statistics, (2) the eigenvalue series from the interitem correlation matrices, (3) a cluster analysis-based examination of dimensional structure, and (4) conditional variances of number correct scores. Each of these criteria are described in turn.

Item analysis statistics. Classical item analysis statistics were computed from each data set. These included item passing rates, or *p*-values, item-test score biserial correlations, product-moment interitem correlations, and internal consistency test reliability coefficients (α). Root mean squared differences (RMSD) were calculated between the *p*-values, biserial correlations, and interitem correlations obtained from the simulated and real data. The maximum discrepancy between simulated and real interitem correlations was also recorded. These differences were taken as measures of how well these observable characteristics were preserved by the simulation methods.

Eigenvalue series. The eigenvalue series from each product-moment interitem correlation matrix was also computed. The graph of an eigenvalue series, often called a scree plot, is commonly considered as providing at least rudimentary evidence of a test's dimensional structure. Our intention was not to use it in this way, as there exists a long and ever lengthening list of procedures better suited to this purpose. Rather, we simply view the eigenvalue series as one more observable feature of data that a proper simulation should replicate. Accordingly, scree plots were produced comparing the unidimensional and high-dimensional data sets both to each other and to the appropriate real data set.

Eigenvalues from tetrachoric as well as product moment correlation matrices were initially considered. It was thought that the tetrachoric scree plots would more clearly reveal the strict unidimensionality of the unidimensionally simulated data. However, this was found not to be the case, even when the tetrachorics were corrected for guessing. This is perhaps due to what McDonald and Ahlawat (1974) termed "nonlinearity", masking the pure one-dimensional structure

of the unidimensional data. The result was that the relative shapes of scree plots based on tetrachorics differed little from those based on product moment correlations.

A second type of plot produced shows the difference between each real-data eigenvalue and the corresponding value from both the unidimensional and high-dimensional simulations. Differences between the simulated and real data eigenvalues are shown much more clearly on these plots, mainly because both the Math and English tests were predominantly unidimensional and the first eigenvalue dwarfed all others. This set the scale of the scree plots in such a way as to hide relatively important differences later in the series.

Scree plots and eigenvalue difference are identically interpreted. Greater similarity of simulated to real eigenvales shows better recovery of the underlying content structure of the test.

Cluster analysis. Methods described by Miller and Hirsch (1992) yield a more refined examination of the dimensional structure of the real and simulated data sets. These procedures start by fitting a high dimensional latent trait model to the data and obtaining item parameter estimates. Each item's parameters define a vector in multidimensional space, with direction cosines from origin given by:

$$\alpha_k = \cos \frac{a_k}{\left(\sum_{l=1}^{m} a_l^2 \right)^{\frac{1}{2}}} \qquad (2)$$

for $k = 1,2,...m$.

Each item's vector is interpreted as pointing in the direction in the ability space along which the item measures most strongly. Items with vectors pointing in similar directions are concluded to be measuring similar composites of traits. A similarity measure for each pair of items can therefore be obtained as the direction cosine or angle between the items' vectors. The interitem similarity matrix defined by this measure is submitted to a complete linkage cluster analysis and the resulting cluster tree diagram inspected.

Cluster trees were found for both real and simulated data sets. The unidimensional and high-dimensional trees were compared to the real data trees and the degree of similarity evaluated. Again, the simulation model producing the cluster tree most like that of the real data would be considered as better recovering the subtle underlying characteristics of actual item responses.

Conditional variances. Consider splitting a test's items into two subtests, with items assigned to subsets by any rule. Even numbered items could be in one group, odd in the other; first half of the test in one group, second in the other, etc. The accuracy with which scores on one subtest could predict scores on the other is an indication of the test's reliability. One measure of this accuracy is to compute the variances of the scores on one subtest conditional on scores on the other. For example, choose the first subtest as the base. Assign examinees to groups according to their scores on the base test. Then compute the variance of the second subtest scores within each of these groups. This conditional variance approach has been used within the context of a variety of measurement related research problems (see Woodruff, 1990, 1991).

Conditional variances indicate not just a test's reliability, but also the extent to which the test departs from unidimensionality. One definition of test unidimensionality or homogeniety is that no matter how a test's items were assigned, the resulting subtests would be congeneric. Conditional variances would therefore not change acoording to how subtests were formed. This is not the case with multidimensional tests. Consider a test comprised of two equal sets of items measuring very different traits. Conditional variances would depend strongly on how these items were distributed across the two subtests. Variances would be much smaller if the subtests were made parallel by assigning the two types of items in equal proportions to the two subtests. Variances would be much larger if the subtests each consisted of only items of the same type. Subtle features of a test's dimensional structure are therefore revealed by defining subtests in different ways and observing the effects on conditional variances. Comparing conditional variances derived from real data with those derived from simulated data provides one more measure of the influence of the simulating dimensionality on the realism of the resulting data.

Conditional variances were calculated differently for the English and Math tests. For the English test, the total score on the last two passages (i.e., items 46 through 75) were considered the base subtest and used to sort the total scores on the first three passages (items 1 through 45). The variances of the total scores on the first three passages were plotted at each different total score level observed on the last two passages.

Because the Math test is not passage based, items were sorted into equal-sized subtests according to the cluster analysis conducted on the real Math dataset. One subtest was arbitrarily taken as the base and used to assign examinees to score-level groups. Variances of the second subtest's scores were then computed within each group and the results plotted. The goal was to

identify which simulation method produced conditional variance plots that were most similar to those found in the real data.

Results

Item analysis statistics. Review of the statistics presented in Table 1 might lead to the conclusion that unidimensional and high-dimensional simulation procedures produce very similar item response data. For example, α coefficients for the English test under the unidimensional and 10 dimensional simulations were both approximately 0.93. The mean interitem correlations were both approximately 0.15, very similar to the real data. Small RMSDs indicate that item *p*-values and biserial correlations were also very similar across simulation models and quite close to the true values. The results for Math closely parallel those for English in these aspects. Only when the interitem correlations are looked at closely is there any evidence that the dimensionality of the simulation model is important. The RMSD between simulated and real correlations dropped from .030 to .016 for English and from .026 to .017 for Math as the simulation dimensionality increased from one to ten. Comparison of the largest interitem correlation differences is even more dramatic. These decreased with increasing dimensionality from .258 to .085 for English and from .191 to .070 for Math. One interpretation of this finding, supported by additional evidence offered below, is that unidimensional simulation models are capable of reproducing gross features of real data but fail when examined at a more detailed level.

Eigenvalues. The eigenvalues obtained from the three English datasets are plotted in Figure 1. As shown, the first root in each dataset was approximately 12.5, dropping to between 2.5 and

1.5 for the second root. The scree plot for the high-dimensional simulation falls closer to the real data than does that of the unidimensional simulation. This result is highlighted in Figure 2, where the differences between the eigenvalues obtained from the simulated data and real data are plotted. This figure shows that the differences for the high- dimensional simulated data were consistently much smaller than those of the unidimensional data.

The scree plots and eigenvalue differences for the Math datasets are shown in Figures 3 and 4. These figures concur with those from the English test, and suggest that the structure found in the high-dimensional simulated data sets was more similar to that found in real data.

Cluster Analysis. The cluster tree for the real English dataset is presented in Figure 5. Five significant and substantively interpretable clusters were identified, these labeled one through five in lefthand margin of the diagram. Trees from the simulated data were examined to see whether these same clusters were reproduced.

The tree obtained from the unidimensional simulation is in Figure 6. It is fairly clear that the content structure of the real data was not preserved, as the items that cluster together tightly in the real data are scattered across the tree in an apparently random fashion.

The results from the high-dimensional simulation are considerably more encouraging. While the cluster trees are far from identical, there are important areas of agreement. The five major clusters found in the real data each appear in the simulated data as well. This indicates that the basic dimensional structure of the real data was reproduced by the simulation.

The cluster trees for the Math data sets are presented in Figures 8 through 10. Four substantively interpretable item clusters were identified in the real data (Figure 8). As with the

English test, the unidimensional simulation failed to reproduce any of these clusters, with their items spread seemingly randomly across the cluster tree (Figure 9).

Considerable agreement was again found between the real data and high-dimensional simulation cluster trees. All four item clusters were faithfully reproduced in the simulated data. Compared to the unidimensional tree, any discrepancies noted between the real data and high-dimensional trees appear relatively minor. The cluster analysis again suggests that complex structure of the real data was maintained in the high-dimensional simulation.

Conditional variances. The conditional variances for the English and Math tests are shown in Figures 11 and 12, respectively. It is immediately clear from both figures that the unidimensional simulation greatly underestimated the true or real-data conditional variances. As might be expected, unidimensional simulations produce data that are unrealistically homogeneous and one-dimensional. Any subtle or complex structure present in the real data is lost. Conversely, the conditional variances from the high-dimensional simulation track those of the real data fairly closely. No evidence of underestimation is observed.

Conclusions

Like most observable phenomena, item response data yield different impressions when examined at different levels of analysis. When viewed at a highly summarized or aggregated level, the differences between unidimensional and high-dimensional simulated data appear fairly trivial. Statistics obtained by summing across items and/or examinees are sensitive only to the most basic features of the data, which are well reproduced by both simulation models. The item

analysis statistics in Table 1 show this clearly. Only when the data are analyzed at a more detailed level are the benefits of high-dimensional simulation apparent. While the eigenvalue and conditional variance plots support this assertion, the most compelling evidence is given by the cluster analysis. There it was plain that items in the high-dimensional simulations tended to cluster in ways very similar to what was found in the real dataset. It was also clear that this was not at all true with the unidimensional simulations, where items that clustered in real data were unsystematically distributed across the cluster tree.

Correctly reproducing even minor characteristics of real data may be important in many applications of simulation procedures. For example, consider evaluating several methods of content balancing in computerized testing. Content balancing is the practice of selecting items in a CAT not simply with respect to their measurement or statistical properties, but with regard to some content blueprint as well (Stocking & Swanson, 1993). Implicit in the design of all content balancing schemes is the assumption that examinees respond to items measuring different content areas in different ways. Unfortunately, these differences cannot be solely attributed to differences on a unidimensional underlying ability scale. Instead they are believed to arise from different educational or training backgrounds, different work experiences, different geographical locations, etc. A student who had an outstanding trigonometry teacher and a poor algebra teacher may find trigonometry items relatively easier than algebra items. However, other students find otherwise and respond accordingly. It is due to differences of this sort that most knowledge domains are fit only approximately by strict unidimensional models. These differences are also the reason that most successful tests are built in accord with extensive and detailed content blueprints that stipulate exact numbers of items of each content type.

When item responses are simulated according to unidimensional models all of the reasons for employing content balancing are absent from the data. Under unidimensional models item responses are completely driven by the examinee's standing on the single ability continuum. This is far too simple. Characterizing all aspects of examinee performance by only a single parameter leaves little room for modeling minor differences in educational background. All examinees at the same level on the single ability scale are required to interact identically with all items. The model does not allow some examinees to find items measuring a particular content relatively easier and those measuring a different content relatively more difficult. Item content cannot affect examinee performance. When data are simulated unidimensionally, all content balancing algorithms would be evaluated as performing identically because none was necessary. If content does not matter, the way that it's balanced *certainly* does not matter.

6. Discussion

Generating item response data from realistic multidimensional models is far from a novel idea. The methods have roots in the work of Tucker, Koopman & Linn (1969), and have been used extensively over years to study the robustness of various unidimensional parameter estimates to the inevitable violation of that model's assumptions. We have simply outlined workable procedures for multidimensional simulations and advocate their use more generally.

Researchers have often acknowledged, at least implicitly, that typical simulation procedures are unrealistic. However, these concerns have largely been limited to cases where the entire point of a study is the performance of procedures with data that fail to fit the presumed model exactly. The example of robustness studies mentioned above is one such case. Differential item

functioning studies are another. Like item content, DIF is another cause of different examinees responding to items in different ways. Unfortunately, model fit concerns seem to dissipate when equating procedures, CAT item selection algorithms, or computerized test assembly systems are compared. We argue that evaluation of testing methodology in a realistic context is *always* important.

High-dimensional simulation can be criticized as potentially capitalizing on chance and modeling characteristics that are idiosyncratic or confined to the particular examinee sample that provided the simulation model parameters. Of course, this criticism is true of any simulation design. It is always somewhat of a reach to generalize from the particular test or examinee population on which a simulation was based to tests and examinee populations as a whole. We would argue that generating data realistically expands rather than limits the extent to which a given simulation's results can be inferred to hold more generally. The particular problem of overfitting can be addressed by estimating model parameters from several, independent examinee samples. The proper simulation dimensionality would then be determined as that which reproduces only data characteristics that are observed repeatedly across different samples of real data.

Additional work is needed to determine if using high-dimensional as opposed to unidimensional simulations will lead to different conclusions when comparing different psychometric methods or procedures. Finding such cases will dramatically emphasize the importance of realistic simulation procedures, particularly if the high-dimensional conclusions could be bolstered by evidence from real-data studies.

7. References

Fraser, C. (1986). *NOHARM: An IBM PC Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory* [Computer Program]. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Hambleton, R. K. (1969). *An empirical investigation of the Rasch test theory model.* Unpublished doctoral dissertation, University of Toronto.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo Methods*, Methuen, London.

Harwell, M. Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20,* 101-125.

Holland, P. & Wingersky, M. (1991). *A monte carlo comparison of four approaches to on-line calibration of a CAT.* Paper presented at the Office of Naval Research contractors annual meeting, Princeton, NJ.

Levine, M. V. (1984). *An introduction to multilnear formula score theory.* Model-Based Measurement Laboratory Report 84-4. University of Illinois. Urbana.

McDonald, R. P. & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of mathematical and statistical Psychology, 27,* 82-99.

Miller, T.R. & Hirsch, T.M. (1992). Cluster analysis of angular data in applications of item response theory. *Applied Measurement in Education, 5(3),* 193-211.

Mislevy, R.J. & Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models.* [Computer program]. Mooresville, IN: Scientific Software.

Panchapakesan, E. (1969). *The simple logistic model and mental measurement.* Unpublished doctoral dissertation, University of Chicago.

Samejima, F. (1979). *A new family for multiple choice items* (Research Rep. No. 79-4). Knoxville: Univeristy of Tennessee, Department of Psychology.

Spence, I. (1993). Monte carlo simulation studies. *Applied Psychological Measurement, 7,* 405-425.

Stocking, M.L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292.

Student. (1908). Probable error of the correlation coefficient, *Biometrika*, **6**, 302.

Tucker, L. R., Koopman, R. F. & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34*, 421-459.

Wilson, D. T., Wood, R. & Gibbons, R. (1991) *TESTFACT: Test Scoring, Item Statistics, and Item Factor Analysis*. [Computer program]. Scientific Software. Chicago.

Woodruff, D. (1990). Conditional stamdard error of measurement in prediction. *Journal of Educational Measurement, 27*, 191-208.

Woodruff, D. (1991). Stepping up test score conditional variances. *Journal of Educational Measurement, 28*, 191-196.

Table 1

Comparative Statistics Calculated on Real and Simulated Datasets

Dataset	α	Mean r_{ij}	RMSD			Maximum r_{ij} Difference
			P	Biserial	r_{ij}	
English						
Real	0.928	0.147	n/a	n/a	n/a	n/a
Unidimensional	0.929	0.149	0.004	0.015	0.030	0.258
Multidimensional	0.929	0.150	0.005	0.014	0.016	0.085
Math						
Real	0.929	0.205	n/a	n/a	n/a	n/a
Unidimensional	0.940	0.207	0.007	0.018	0.026	0.191
Multidimensional	0.941	0.209	0.008	0.017	0.017	0.070

Figure 1
Eigenvalue Series for English Datasets

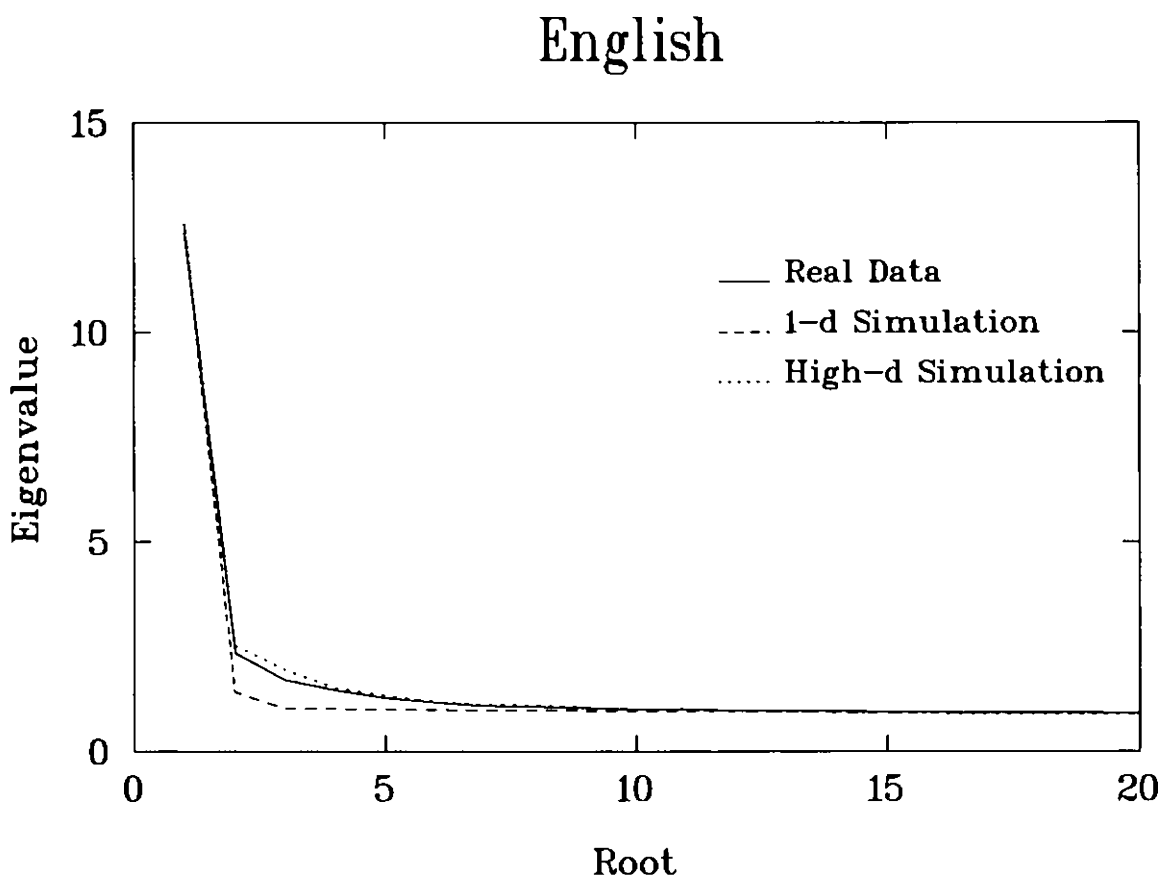

Figure 2
Eigenvalue Differences for English Datasets

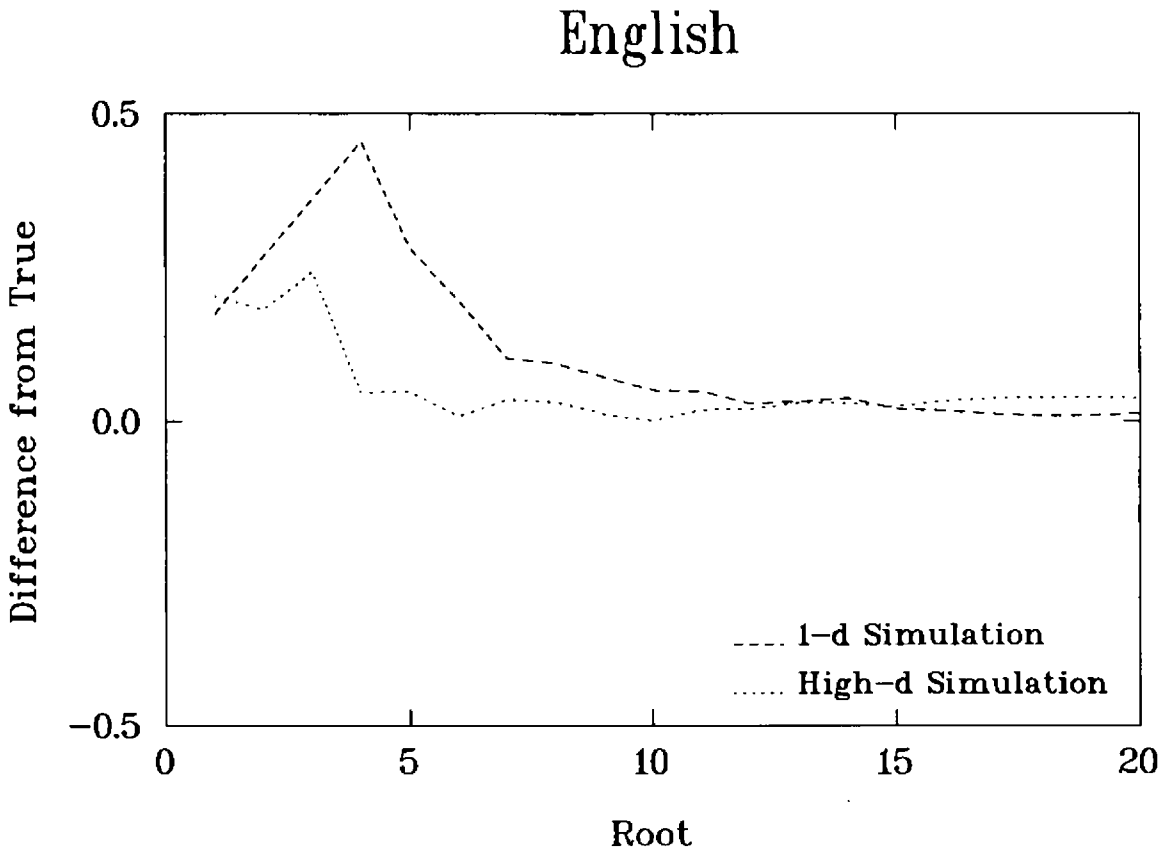

Figure 3
Eigenvalue Series for English Datasets

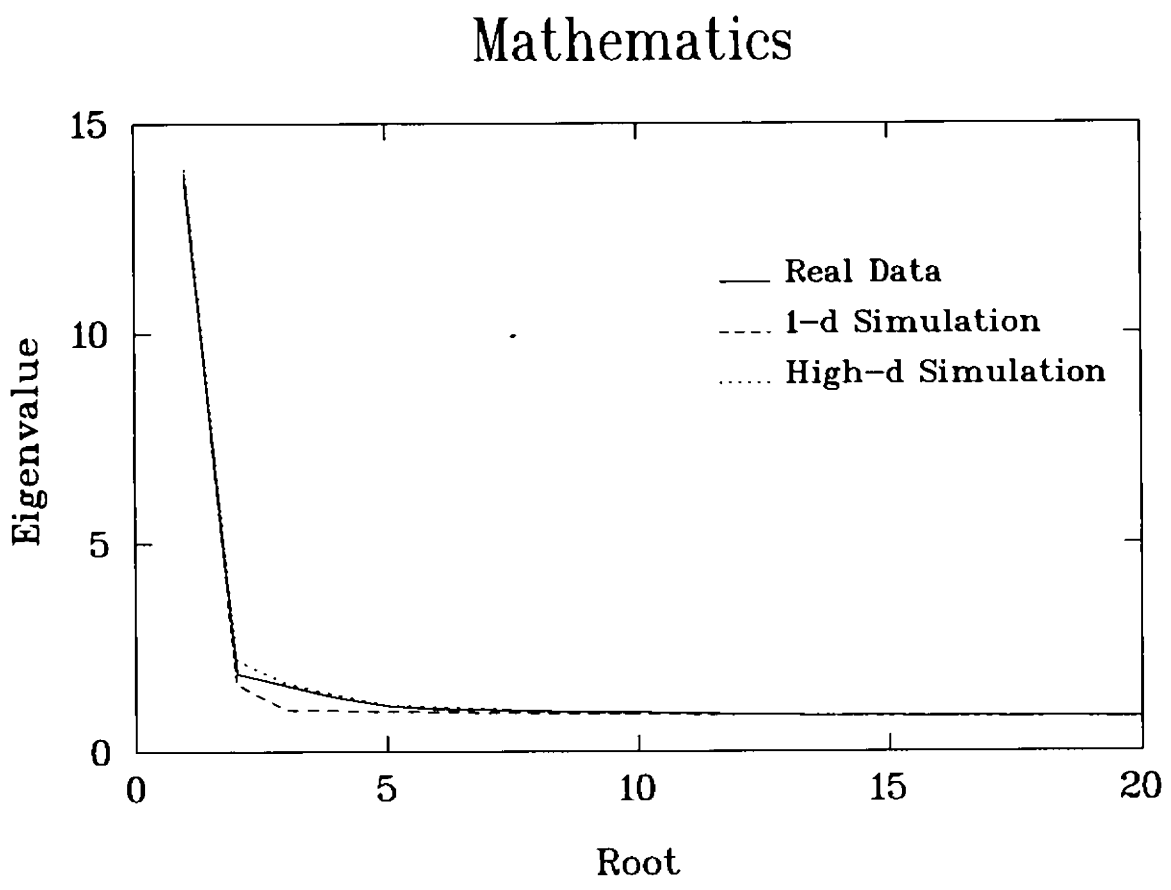

Figure 4
Eigenvalue Differences for English Datasets

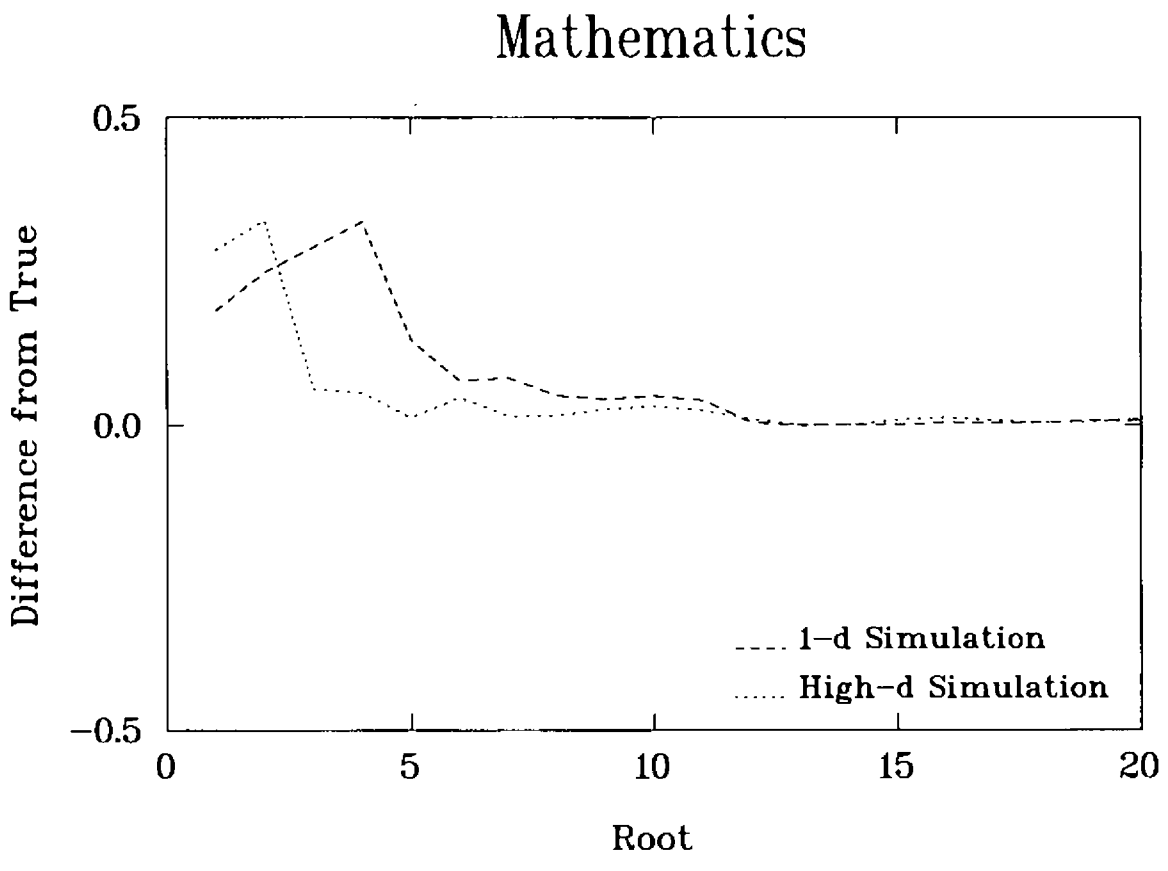

Figure 5
Cluster Diagram from Real English Dataset

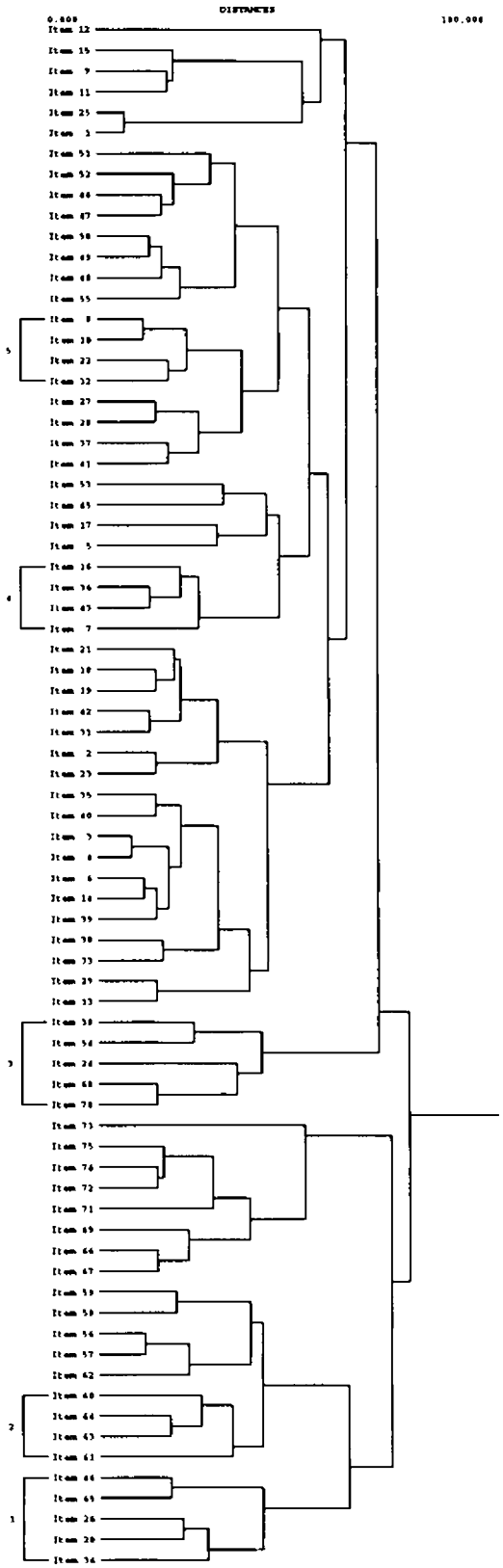

Figure 6
Cluster Diagram from Unidimensionally Simulated English Dataset

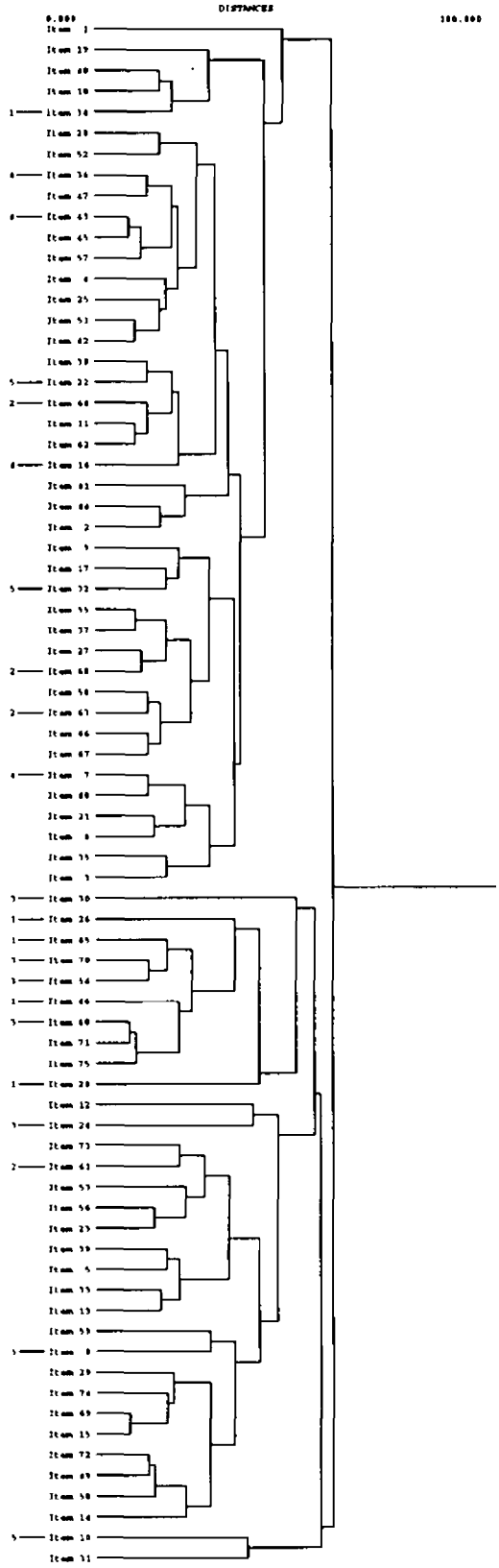

Figure 7
Cluster Diagram from Multidimensionally Simulated English Dataset

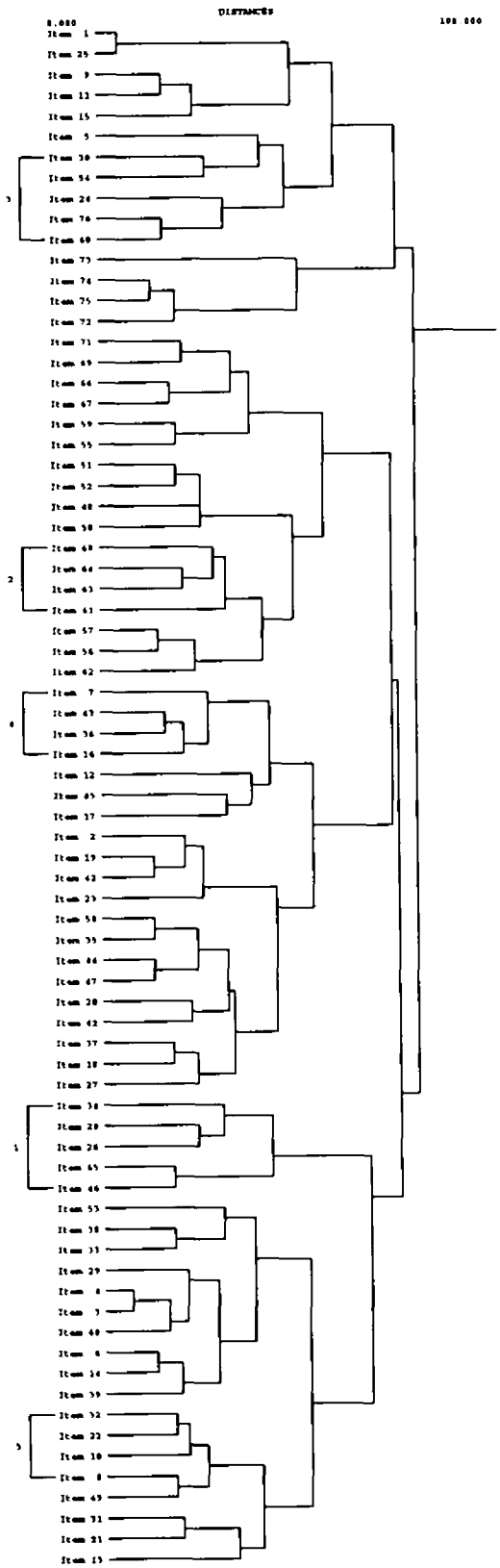

Figure 8
Cluster Diagram from Real Math Dataset

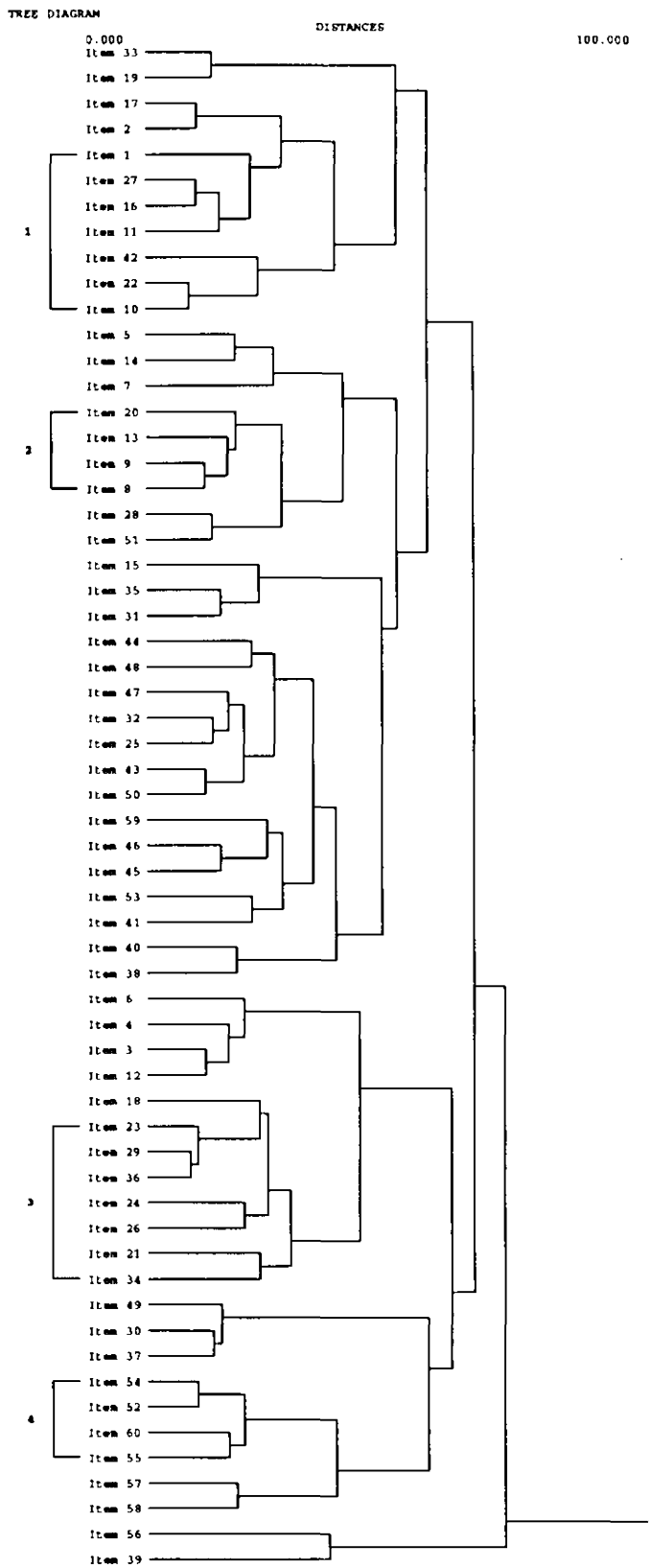

Figure 9
Cluster Diagram from Unidimensionally Simulated Math Dataset

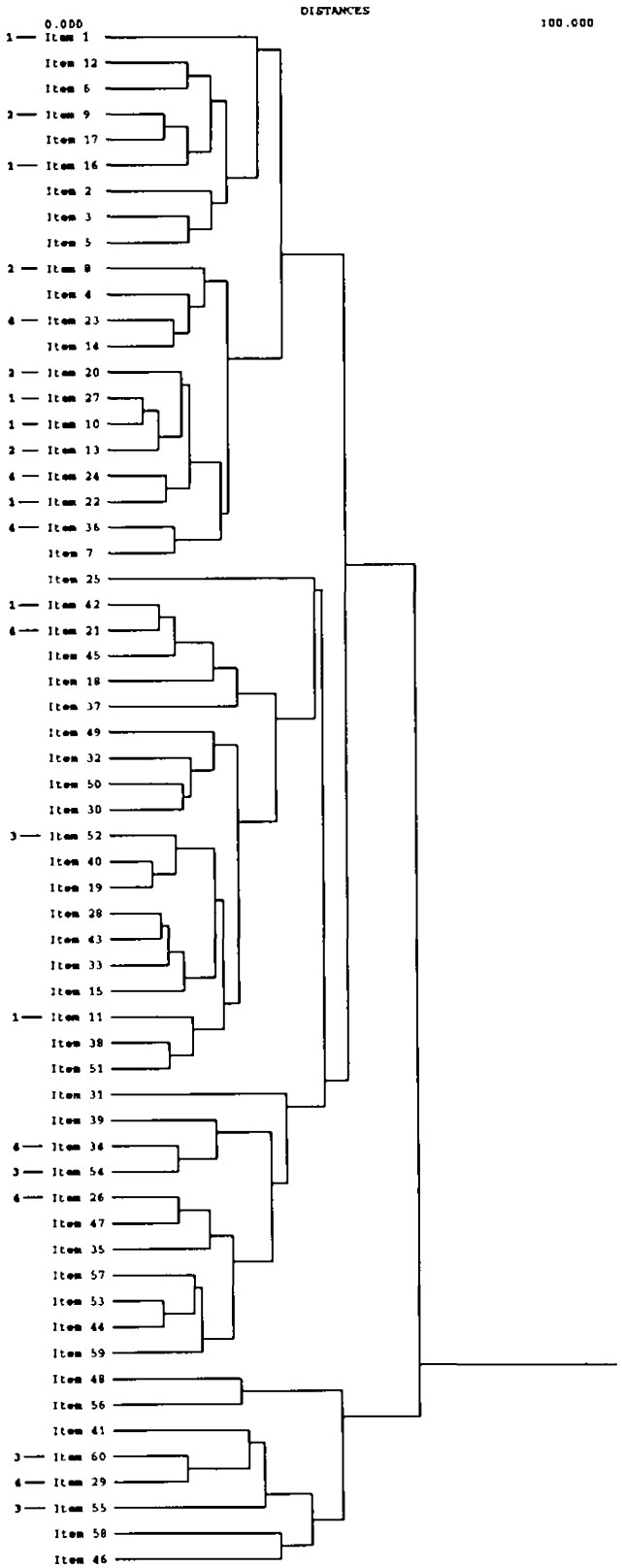

Figure 10
Cluster Diagram from Multidimensionally Simulated Math Dataset

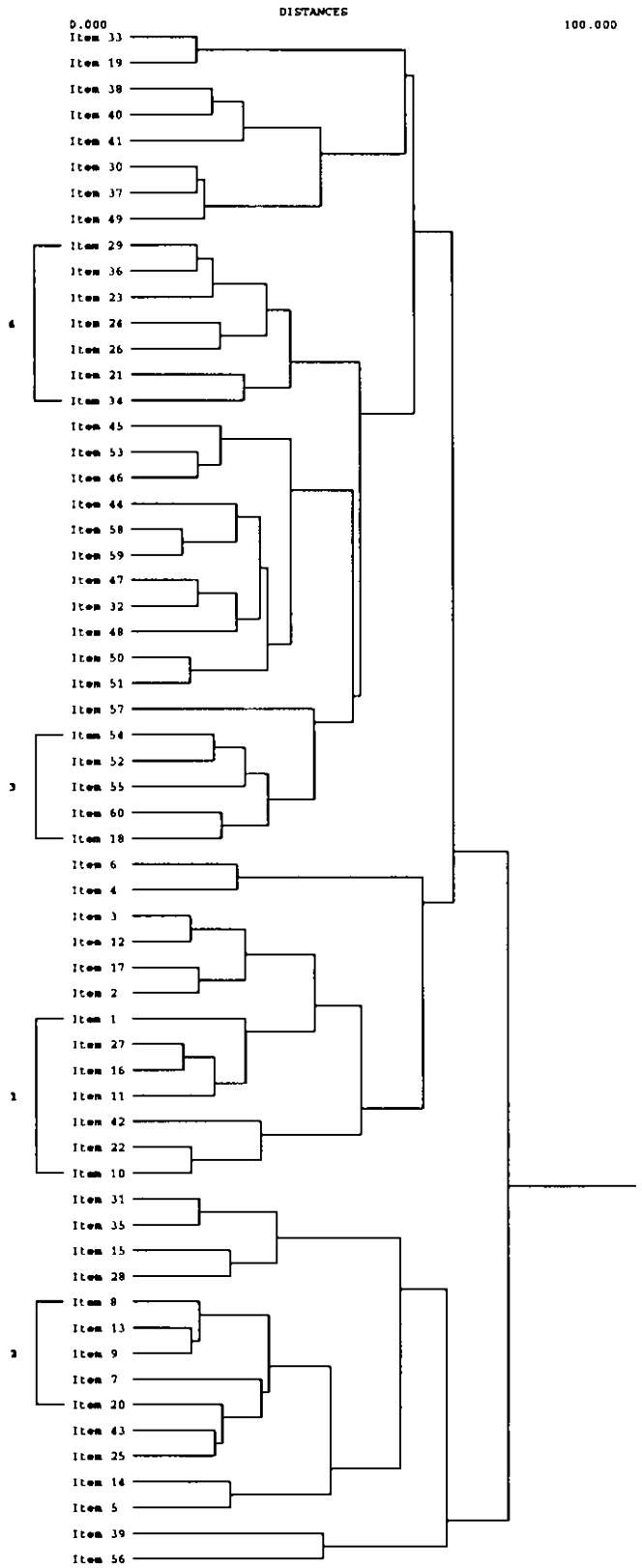

Figure 11
Conditional Variances for English Datasets

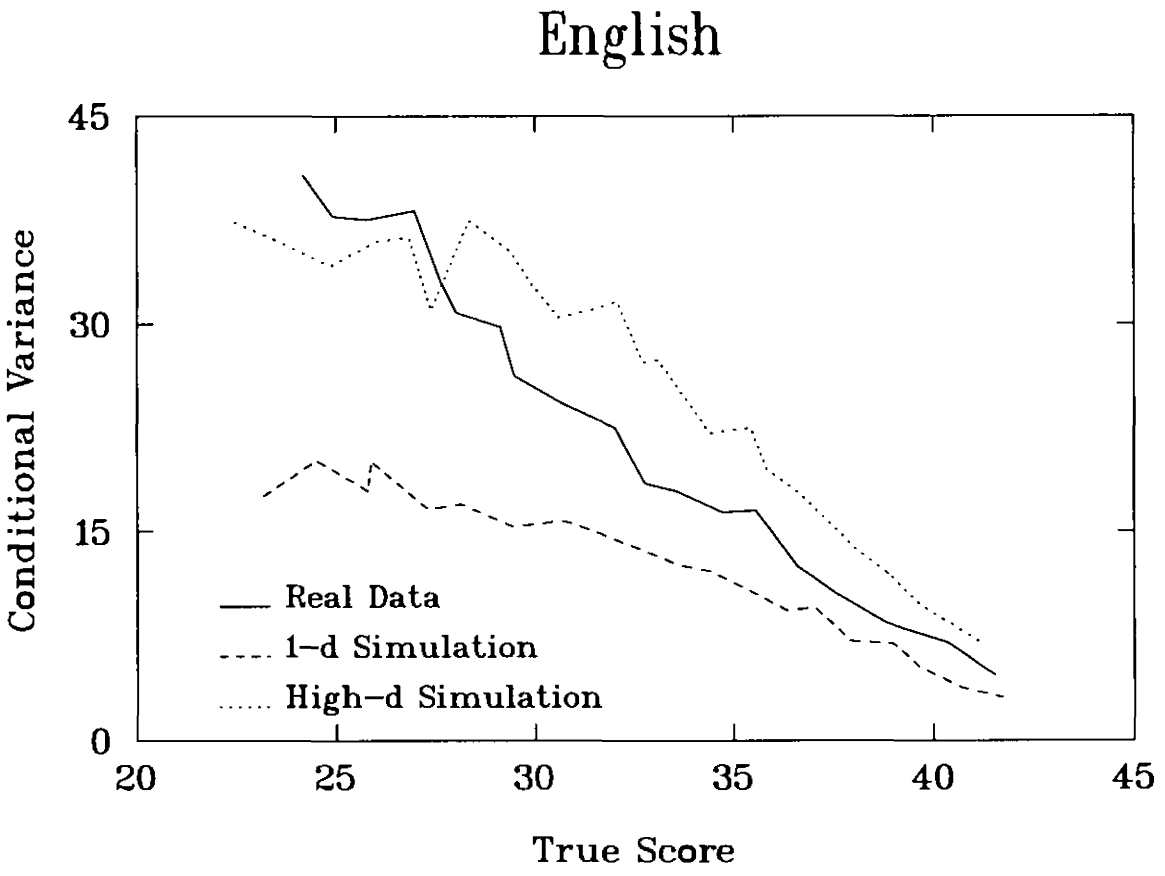

Figure 12
Conditional Variances for Math Datasets