Monte Carlo Studies in Item Response Theory

Michael Harwell, Clement A. Stone, Tse-Chi Hsu, and Levent Kirisci
University of Pittsburgh

Monte carlo studies are being used in item response theory (IRT) to provide information about how validly these methods can be applied to realistic datasets (e.g., small numbers of examinees and multidimensional data). This paper describes the conditions under which monte carlo studies are appropriate in IRT-based research, the kinds of problems these techniques have been applied to, available computer programs for generating item responses and estimating item and examinee parameters, and the importance of conceptualizing these studies as statistical sampling experiments that should be subject to the same principles of experimental design and data analysis that pertain to empirical studies. The number of replications that should be used in these studies is also addressed. *Index terms: analysis of variance, experimental design, item response theory, monte carlo techniques, multiple regression.*

Item response theory (IRT) is an important and popular methodology for modeling item response data that have applicability to numerous measurement problems. The enthusiasm for IRT, however, has been tempered by the realization that the validity with which these methods can be applied to realistic datasets (e.g., small numbers of items and examinees, multidimensional data) is often poorly documented. Increasingly, monte carlo (MC) techniques in which data are simulated are being used to study how validly IRT-based methods can be applied. For example, for the period 1994–1995, approximately one-fourth to one-third of the articles in the journals *Applied Psychological Measurement* (*APM*), *Psychometrika*, and the *Journal of Educational Measurement* (*JEM*) used MC techniques.

The popularity of IRT MC studies, coupled with the desire for the results of such studies to be used by measurement professionals, suggests a need for a comprehensive source that describes what MC studies are, a rationale for using these studies in IRT, and guidelines for properly designing and executing a MC study and analyzing the results. Unfortunately, there has been no single source of information about IRT MC studies to consult; although books on these techniques are available, they are devoted to solving statistical rather than measurement problems. This paper was designed to fill this gap.

First, a definition of a MC study is offered and a description of a typical MC study in IRT is provided. Next, references are provided for readers desiring an introduction to these techniques, followed by a brief history of these procedures. The conditions under which a MC approach is appropriate are outlined, along with some advantages and limitations of these techniques. The results of a survey of IRT articles in three prominent measurement journals are reported to illustrate the variety of problems to which these techniques have been applied.

Following the advice of several authors (e.g., Hoaglin & Andrews, 1975; Naylor, Balintfy, Burdick, & Chu, 1968; Spence, 1983), the description of the design and implementation of an IRT MC study treats these studies as statistical sampling experiments that should be subject to the same guidelines and principles that pertain to empirical studies. Spence captured the importance of this perspective:

> The Monte Carlo experiment is a designed experiment and, as such, is capable of displaying the same virtues and vices to be found in designed experiments in more familiar settings. Bad design, sloppy execution, and inept analysis lead to the same kinds of difficulties in the world of computer simulation as they do in the laboratory. (p. 406)

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 2, June 1996, pp. 101–125

0146-6216/96/020101-25$2.50

Finally, some implications for using MC techniques in IRT-based research in the future are presented.

Monte Carlo Studies

The term "monte carlo" is frequently used when random numbers are generated in stochastic simulation (Naylor et al., 1968), although such studies are really statistical sampling experiments with an underlying model whose results are used to address research questions. [Rubinstein (1981, p. 3) describes the role of models in MC studies.] MC studies are in many ways mirror images of empirical studies with one key difference: The data are simulated using a computer program.

A Typical Monte Carlo Study in IRT

A typical IRT MC study might proceed as follows:

1. One or more research questions that define the purpose of the study are specified. For example, "How accurate are parameter estimates for a two-parameter logistic model (2PLM) when the numbers of items and examinees are varied and different prior distributions for the item parameters are specified?"
2. The conditions to be studied are delineated. This includes the independent variables, such as the numbers of examinees and items, and the dependent (outcome) variables, which are used to measure the effects of the independent variables.
3. An appropriate experimental design is specified.
4. Item response data are generated following an underlying model, such as the 2PLM, subject to the conditions studied.
5. Parameters are estimated using the simulated item responses.
6. Outcomes that measure the effect of the conditions being modeled are compared (e.g., median standard error for estimated discrimination parameters across n items).
7. This process is replicated R times for each cell in the design, producing R outcomes per cell.
8. The R outcomes are analyzed using both descriptive and inferential methods. Inferential analyses are guided by the research questions and the experimental design. For example, use of a factorial design might suggest that analysis of variance (ANOVA) should be used. The results from the descriptive and inferential analyses provide evidence concerning the research questions under investigation.

General References for Monte Carlo Techniques

Readers desiring general references for MC techniques will find this literature to be rather extensive, and only references that are relatively well known and have features that are particularly suitable for IRT-based research are mentioned here. Some useful references are Hammersley & Handscomb (1964), Rubinstein (1981), and Ripley (1987). These texts cover the definitions, objectives, and limitations of MC techniques, random number generation, testing of random numbers, variance reduction methods, and selected applications. Readers desiring a brief treatment may find it sufficient to read Chapter 11 of Lehman & Bailey (1968) and Hartley's Chapter 2 in Enslein, Ralston, & Wilf (1977); those interested in the mechanics of generating random numbers should find De Vroye (1986) and Newman & Odell (1971) helpful. For readers interested in the application of these methods to IRT, the article by Wilcox (1988) is a good introduction.

Some texts are written with applications to specific fields. For example, the texts by Smith (1975) and Naylor et al. (1968) are intended for econometricians, and the text by Lewis & Orav (1989) for engineers. In the absence of a text specifically written for psychometricians, the text by Naylor et al. is recommended. Even though the text is relatively old, most of the discussion about the planning and design of statistical sampling experiments is applicable to IRT-based research.

A Brief History of Monte Carlo Techniques

The history of MC techniques can be divided into three periods. The first can be called pre-MC and covers the time before the term was used. During this period, these techniques were not really considered formal methods of research, yet they made significant contributions to the study of statistical problems. For example, statistical sampling experiments figured prominently in the discovery of the distribu-

tions of the correlation coefficient and t by Student in 1908, as cited in Hammersley & Handscomb (1964).

The term "monte carlo" was first used by Metropolis & Ulam (1949), which may be considered the beginning of the second period. These techniques were used as a research tool during World War II in their study of problems related to the atomic bomb, and were popularized by researchers in the immediate post-war years (von Neumann & Ulam, 1949). The third period of MC methods began in the late 1960s and the early 1970s when high-speed computers became accessible to many researchers, and these studies became a popular method of research for statistical problems.

When Is a Monte Carlo Study Appropriate in IRT?

The conditions under which a MC study would be appropriate are summarized in the publication policy of *Psychometrika* (Psychometric Society, 1979):

> Monte Carlo studies should be employed only if the information cannot reasonably be obtained in other ways. The following are probably the most common situations in psychometrics where the method may be appropriate:
> (a) Determination of sampling distribution of test statistics, or comparisons of estimators, in situations where analytic results are difficult to obtain, e.g., when the robustness of a test statistic is investigated.
> (b) Comparison of several algorithms available to perform the same function, or the evaluation of a single algorithm. It is very important that the objectives and limitations of such studies be carefully and explicitly considered. (pp. 133–134)

Thus, a MC study should be considered if a problem cannot be solved analytically, and should be performed only after a compelling rationale for using these techniques has been offered.

Advantages and Limitations of Monte Carlo Studies

The popularity of MC studies in IRT-based research should not be taken as evidence that these techniques are methodological panaceas; to the contrary, their success depends in large part on the skills of the researcher performing the study. Still, it is possible to list some general advantages and limitations of these studies (Lehmann & Bailey, 1968; Naylor et al., 1968).

Perhaps the most compelling advantage is that MC studies can often be conducted when an analytic solution for a problem does not exist or is impractical because of its complexity. As suggested by *Psychometrika's* publication policy, research questions should ideally be solved in a precise analytic or mathematical way, similar to the way that analytic methods are sometimes used to study the properties of statistical tests (e.g., Box, 1954; Rogosa, 1980). These methods deduce from postulates and are exact in the sense that if a set of underlying assumptions is true, the results are highly generalizable because they involve no uncertainty. However, the results may be of little value if the assumptions underlying an analytic solution are unrealistic.

For example, suppose that it was necessary that estimated standard errors for discrimination parameters in a dichotomous 2PLM be less than .15. How many examinees would be needed to achieve this for a given number of test items if the prior variance of the distribution of discrimination parameters doubled? Would the needed number of examinees be the same for the 2PLM versus the three-parameter logistic model (3PLM) and for a 10-item versus a 15-item test, and what would be the effect (if any) of a skewed trait (θ) distribution on the standard errors? Solutions to these kinds of questions can be quite useful and ideally would involve no uncertainty; unfortunately, analytic solutions to questions such as these in IRT are typically very difficult or simply impossible.

Alternatively, an experiment could be conducted to provide information about the conditions needed to produce standard errors less than .15. Experiments use a process of induction, study one behavior or phenomenon in realistic settings, and produce results of quality that depend heavily on the size of the sampling error and whose generalizability depend on the representativeness of the

sample. A statistical sampling experiment (i.e., a MC study) satisfies these conditions.

Other advantages of MC studies include the capability of specifying and manipulating values of parameters and studying the effects of several factors at one time. Also, MC studies are sometimes the fairest way of comparing alternative actions and may be less expensive than studies involving humans.

But researchers should not uncritically turn to MC studies, because indiscriminate use of these techniques may produce misleading results or results with little utility. One limitation of these studies is that the usefulness of the results is highly dependent on how realistic the conditions modeled are (e.g., assumed distribution of the parameters or data). Another drawback is that the quality of the random number generator is difficult to assess, especially for a long series of numbers. MC results may also vary depending on the number of replications used and the numerical precision of the computer (Stone, 1993).

The Use of Monte Carlo Studies in IRT

Applications of MC techniques in IRT have typically involved one or more of the following: (1) evaluating estimation procedures or parameter recovery, (2) evaluating the statistical properties of an IRT-based statistic (e.g., a goodness-of-fit measure), or (3) comparing methodologies used in conjunction with IRT (e.g., differential item functioning or multidimensionality assessment). All involve generating random samples using an assumed model for the purpose of comparing results when the "truth" is known. However, Types 1 and 2 involve generating and analyzing empirical sampling distributions, whereas Type 3 involves generating data and comparing the extent to which methodologies detect manipulated characteristics in the data.

To document the problems to which MC studies have been applied, 26 published studies using these techniques, either exclusively or partially, and appearing in *APM*, *Psychometrika*, or *JEM* between 1981 and 1991 inclusive, were identified, classified, and tabulated according to the nature of the problem investigated. Studies that used MC techniques partially were those that relied on either simulated data or empirical data, whereas studies classified as using MC techniques exclusively relied entirely on simulated data. Studies in the two problem areas with the highest frequencies were further examined, classified, and tabulated according to problem characteristics and how well they satisfied standards for evaluating MC studies adapted from *Psychometrika* (Psychometric Society, 1979) and Hoaglin & Andrews (1975). The purpose was to illustrate the problems to which MC studies have been applied and to evaluate how well these studies were conducted. The standards used in evaluating the MC studies assessed whether (1) the problem could be solved analytically, (2) the study was a minor extension of existing results, (3) an appropriate experimental design and analysis of MC results was used, (4) locally-written software or modifications of public software were properly documented, (5) the results depended on the starting values for iterative parameter estimation methods, and (6) the choice of distributional assumptions and independent variables and their values were realistic.

Problems Investigated With Monte Carlo Techniques

Table 1 reports frequencies of a sample of IRT MC studies classified according to the journals that published the study and the problems studied. *APM* published most of the studies; *Psychometrika* and *JEM* published about equal numbers of MC studies. The two problem areas with the highest frequencies were parameter estimation and dimensionality.

Parameter Estimation

Research characteristics of the 26 parameter estimation studies reported in Table 2 indicate that the focus has been on relatively short tests (e.g., ≤ 25 items) and a range of examinee sample sizes (N) and IRT models. The various criterion variables used, [e.g., the root mean square deviation (RMSD), defined as the square root of the average of the squared deviations between estimated and true parameter values], were used with approximately the same frequency. Maximum likelihood estimation and its variations were used most frequently to estimate

Table 1
Frequencies of Studies in *APM*, *Psychometrika* (*PSY*), and *JEM* Classified According to Major Research Problem Areas

Problem Area	*APM*	*PSY*	*JEM*	Total
Parameter Estimation (θ and/or Item)	15	10	1	26
Dimensionality	10	3	4	17
Goodness of Fit (Item, Person, Model)	9	0	2	11
Equating	7	1	2	10
Computerized/Adaptive Testing	9	0	1	10
Differential Item Functioning	6	0	3	9
Item Banking/Test Construction	5	1	2	8
Polytomous Models	2	2	1	5
Computer Packages	4	1	0	5
Criterion-Referenced Assessment	1	0	2	3
Miscellaneous	8	6	3	17
Total	76	24	21	121

parameters, and Bayesian estimation was rarely used. 11 studies used MC techniques exclusively and 15 partially. Virtually all of the studies that sampled item discrimination (*a*) or difficulty (*b*) parameters assumed a normal or uniform distribution for these parameters.

Most of the parameter estimation studies failed to satisfy two or more of the standards (Table 3). For example, 8 studies that used locally developed computer programs to generate item responses and estimate parameters failed to provide any documentation supporting the adequacy of those programs. 16 of these studies also used unreplicated designs and simplistic data analyses. Unreplicated designs are particularly vulnerable to the effects of sampling error associated with simulating data, which may affect the validity of the results (Hauck & Anderson, 1984; Stone, 1992). Finally, 20 of the 26 studies failed to provide any evidence that the selection of particular θ distributions and item parameters was realistic.

Dimensionality

17 studies dealing with dimensionality were examined, most of which investigated methods for detecting multidimensionality (9 studies) or studied the effect of multidimensional data when unidimensional models were used (6 studies; see Table 4). Once again, several studies failed to satisfy two or more of the standards for evaluating MC studies (Table 5). For example, only one of the studies (5%) provided an adequate description of the experimental design (compared to 12% of the parameter recovery studies in Table 3), and only 4 (24%) documented the adequacy of the random number generators or other computer resources (compared to 38% of the parameter recovery studies). In addition, 12 (71%) of these studies used an unreplicated design (compared to 61% of the parameter recovery studies). On the whole, the studies involving multidimensionality did a slightly poorer job of satisfying the *Psychometrika* and Hoaglin and Andrews standards than the parameter recovery studies.

Major Steps in Implementing an IRT Monte Carlo Study

Naylor et al. (1968) described several steps for implementing MC studies. Adapted to IRT, these steps include (1) formulating the problem; (2) designing the study, which includes specification of the independent and dependent variables, the experimental design, the number of replications, and the IRT model; (3) writing or identifying and validating computer programs to generate item responses and to estimate parameters; and (4) analyzing the MC results.

Formulating the Problem

Formulating a proper research problem is critical in any research endeavor and MC studies are no exception. The researcher must determine the problem

Table 2
Frequencies of 26 Parameter Estimation Studies Classified According to Research Characteristics (1PLM = One-Parameter Logistic Model)

Estimation Method	IRT Model Investigated
Maximum Likelihood: 14 Studies	1PLM: 8 Studies
Bayesian: 2 Studies	2PLM: 7 Studies
Other Estimation Methods: 7 Studies	3PLM: 10 Studies
Comparison of Methods: 3 Studies	Others: 5 Studies
Variables Most Frequently Manipulated	Distribution (Including Prior) Simulated
Test Length	θ Parameter
5 Items: 2 Studies	Normal: 15 Studies
10 Items: 3 Studies	Uniform: 6 Studies
15 Items: 4 Studies	Fixed: 3 Studies
20 Items: 5 Studies	Others: 5 Studies
25 Items: 4 Studies	*a* Parameter
30 Items: 2 Studies	Uniform: 9 Studies
35 Items: 3 Studies	Fixed: 6 Studies
40 Items: 2 Studies	Others: 3 Studies
50 Items: 2 Studies	*b* Parameter
60 Items: 2 Studies	Normal: 10 Studies
Calibration Size	Uniform: 9 Studies
N = 100 or Less: 9 Studies	Fixed: 2 Studies
N = 150–200: 5 Studies	Others: 1 Study
N = 300–500: 13 Studies	*c* Parameter
N = 900–1,000: 11 Studies	Normal: 6 Studies
N = More Than 1,000: 7 Studies	Fixed: 3 Studies
θ Level: 6 Studies	Criterion Variable
IRT Model: 6 Studies	RMSE: 6 Studies
b Parameter: 2 Studies	RMSD: 6 Studies
Use of MC Techniques	Standard Error/Error Variance: 6 Studies
Exclusively: 11 Studies	θ Estimates: 4 Studies
Partially: 15 Studies	Bias (True – Estimate): 4 Studies
	Item Parameter Estimates: 2 Studies

to be investigated, the questions to be asked, the hypotheses to be tested, and the effects to be measured. Accepted principles governing these activities in empirical studies should be used (e.g., Babbie, 1989). In general, the formulation of research questions relies heavily on knowledge of a literature; the hypotheses to be tested represent an operationalization of the research questions; and the effects to be measured must be sensitive to the variables being manipulated.

For example, Harwell & Janosky (1991) studied the effect of varying prior variances of the distribution of *a* on item parameter estimation for different test lengths and *N*s using the BILOG computer program (Mislevy & Bock, 1986). They reported that the IRT literature offered little guidance in this area, and hypothesized that smaller prior variances would require fewer examinees to obtain stable parameter estimates and that this phenomenon would be more pronounced for shorter tests. Because this question could not be answered analytically, a MC study was used to provide specific values of the number of examinees and the prior variance needed to produce stable parameter estimates. They used the RMSD as the dependent variable because it was expected to be sensitive to the effects of the variables being manipulated. The survey of published IRT studies summarized in Tables 2 and 4 provides other examples of problems that have been investigated, hypotheses that have been tested, and effects that have been measured.

Designing a Monte Carlo Study

A recurring theme in empirical research is the

Table 3
26 Parameter Estimation Studies Classified According to Frequency of Satisfying *Psychometrika* and Hoaglin & Andrews (1975) Standards for Evaluating Monte Carlo Studies

- Examples of Appropriate Uses of MC Techniques
  1. Determination of sampling distributions of test statistics, or comparisons of estimators, in situations in which analytic results were difficult to obtain: 5 Studies
  2. Comparison of several algorithms available to perform the same function, or the evaluation of a single algorithm: 18 Studies
- Expert Treatment of Design and Analysis
  - Adequate Description of Experimental Design: 3 Studies
- Types of Statistics Reported
  - Descriptive Statistics: 26 Studies
  - Graph/Plot: 14 Studies
  - Correlation/Regression: 10 Studies
  - Inferential Statistics: 2 Studies
- Trustworthiness of the Results
  - If any of the following was reported: 19 Studies
    - The accuracy of the results assessed.
    - The extent of agreement with theoretical results.
    - The details of the simulation.
  - Number of Replications
    - 1 Replication: 16 Studies
    - 2–10 Replications: 3 Studies
    - 11–50 Replications: 3 Studies
    - 51–100 Replications: 1 Study
    - More Than 100 Replications: 2 Studies
- Adequacy of Random Number Generators and/or Other Computer Resources
  - If any of the following was reported: 10 Studies
    - The numerical algorithms used.
    - The computer, programming language, major software components.
  - Software used
    - LOGIST: 11 Studies
    - DATAGEN: 2 Studies
    - IMSL: 7 Studies
    - GENIRV: 2 Studies
    - BILOG/MULTILOG: 3 Studies
    - Self-Developed Program or Not Mentioned: 17 Studies
- Distributional Assumptions of Error Models Realistic
  - If any of the following was mentioned: 6 Studies
    - Simulation of real life data.
    - Justifications for selecting the error model.

importance of carefully designing all aspects of the study to allow hypotheses to be properly tested and research questions to be properly answered. Such studies are marked by efficient modeling of variation in the dependent variable, straightforward estimation and testing of effects with widely available computer programs, and an evaluation of the threats to internal and external validity (Cook & Campbell, 1979). Carefully designed MC studies will have these same desirable characteristics.

Issues related to the design of the MC study include selecting the independent and dependent variables, the experimental design, the number of replications, and the IRT model. The overarching goal of these choices is to maximize the generalizability and replicability of the results.

Selecting the Independent Variables and Their Values

The research questions should dictate the independent variables to be included in the simulation as well as suggest values of these variables, which are discrete and typically represent fixed effects. In the Harwell & Janosky (1991) study, N, test length, and prior variance served as independent variables. Values for these variables were suggested by the research questions, which focused on small Ns and short tests, and previous work in this area.

Model parameters also represent an independent variable in a MC study, although they are rarely treated as such. These parameters are often represented as equally spaced values across a fixed range or as estimates from a previously calibrated test, implying a fixed effect. Random sampling of a and b values from specified distributions, however, identifies item parameters as a random effect. An advantage of randomly selecting model parameters is that some generalizability is obtained, although it is possible to obtain unusual combinations of parameters. In any event, a justification must be provided for the values

Table 4
Frequencies of 17 Dimensionality Studies Classified According to Research Characteristics (MD = Multidimensional)

Research Problem Investigated	MD Models Used to Generate Data
Dimensionality Detection Method: 9 Studies	Sympson (1978): 4 Studies
Comparison of MD Models: 2 Studies	Reckase & McKinley (1985): 3 Studies
When Unidimensional Models Were Used: 6 Studies	Drasgow & Parsons (1983): 2 Studies
Variables Manipulated	Doody-Bogan & Yen (1983): 2 Studies
Test Length: (30, 60): 8 Studies; (60, 45): 5 Studies; (26, 40, 50): 2 Studies; (20, 30, 40, 50, 60): 2 Studies	Distribution (Including Prior) Simulated
Calibration Size: (1,000, 2,000): 13 Studies; (125, 500, 2,000): 2 Studies; (750, 2,000, 20,000): 1 Study; (750, 2,000): 1 Study	θ Parameter
Correlation Between Traits/Factors: (0.0, .3, .6, .9): 4 Studies; (0.0, .3, .6, .9, .95): 1 Study; (.1, .6): 8 Studies; (0.0, .3, .5, .8, 1.0): 2 Studies; (.10 – .90): 3 Studies; (Low, Intermediate, High): 1 Study	Bivariate Normal: 5 Studies
	Normal: 3 Studies
	a Parameter
	Uniform: 2 Studies
	b Parameter
	Uniform: 2 Studies
	Normal: 1 Study
	c Parameter
	Fixed: 2 Studies
Use of MC Techniques	Criterion Variables
Exclusively: 7 Studies	*a* Estimates: 7 Studies
Partially: 10 Studies	*b* Estimates: 7 Studies
	Eigenvalues: 5 Studies
	θ Estimates: 4 Studies
	Reliability Indexes: 4 Studies

of the independent variables that are selected.

Researchers also must consider the relationship between the number of independent variables, the efficiency of the study, and the interpretability of the results. As the number of variables increases, the breadth of the study increases but more time is needed to perform the simulation. For example, the Harwell & Janosky (1991) study involved a $6 \times 2 \times 5$ fully-crossed design, which would require 60, 600, and 6,000 BILOG analyses with 1, 10, and 100 replications, respectively. If a fourth independent variable was added with W levels, the number of analyses would increase by a factor of W. Advances in computing speed and power imply that the issue of efficiency is perhaps not as critical as it once was, but the computer time needed to estimate model parameters for replicated studies can still be substantial. Moreover, as noted by Naylor et al. (1968), if too many variables are included the interpretation of their joint effect may be difficult.

Selecting an Experimental Design

The nature of the independent variables frequently suggests an appropriate experimental design. For example, for a small number of independent variables with relatively few values, a factorial design may be appropriate. In general, selection of an appropriate experimental design for a MC study should take the goals and computing resources of the study into account. Careful selection of a design also helps to delineate the analyses of the MC results that are permissible (Lewis & Orav, 1989).

In the Harwell & Janosky (1991) study, the manipulated variables were N, test length, and variance of the prior distribution of a, which served as independent variables in a completely between-subjects factorial design. A MC study by Yen (1987) typifies another useful experimental design. Yen compared the BILOG and LOGIST (Wingersky, Barton, & Lord, 1982) item analysis programs on, among other things, CPU time, for combinations of various test lengths and θ distributions. This, too, could be represented as a factorial design, but one in which test length and θ distribution served as between-subjects factors and item analysis program as a within-subjects factor.

Table 5
17 Dimensionality Studies Classified According to Frequency of Satisfying *Psychometrika* and Hoaglin & Andrews (1975) Standards for Evaluating Monte Carlo Studies

Examples of Appropriate Use of MC Techniques
1. Determination of sampling distributions of test statistics, or comparisons of estimators, in situations where analytic results are difficult to obtain: 1 Study
2. Comparison of several algorithms available to perform the same function, or the evaluation of a single algorithm: 13 Studies

Expert Treatment of Design and Analysis
- Adequate Description of Experimental Design: 1 Study
- Type of Statistic Reported
  - Descriptive Statistics: 16 Studies
  - Graph/Plot: 5 Studies
  - Correlation/Regression: 6 Studies
  - Inferential Statistics: 1 Study

Trustworthiness of the Results
- If any of the following was reported: 10 Studies
  - The accuracy of the results assessed.
  - The extent of agreement with theoretical results.
  - The details of the simulation.
- Number of Replications
  - 1 Replication: 12 Studies
  - 2–10 Replications: 1 Study
  - 11–50 Replications: 2 Studies
  - 51–100 Replications: 2 Studies
  - More Than 100 Replications: 1 Study

Adequacy of Random Number Generators and/or Other Computer Resources
- If any of the following was reported: 4 Studies
  - The numerical algorithms used.
  - The computer, programming language, major software components.
- Software Used
  - LOGIST: 9 Studies
  - BILOG/MULTILOG: 3 Studies
  - IMSL: 3 Studies
  - MIRTE: 1 Study
  - Self-Developed Program or Not Mentioned: 8 Studies

Distributional Assumptions of Error Models Realistic
- If any of the following was mentioned: 8 Studies
  - Simulation of real life data.
  - Justifications for selecting the error model.

Selecting Dependent Variables

The problem specification should also delineate a class of appropriate dependent variables with which to measure the effects of the manipulated variables. These variables must be sensitive to the independent variables being manipulated, but it is also useful if they are in a form (or can be transformed to a form) that simplifies inferential analyses of the results.

For example, studies evaluating parameter estimation procedures have typically used dependent variables reflecting the success of parameter recovery, such as the RMSD (e.g., Harwell & Janosky, 1991; Kim, Cohen, Baker, Subkoviak, & Leonard, 1994; Stone, 1992). An advantage of the RMSD is that it can be transformed (using a log transformation) so that, if the variables used to compute the RMSD are normally distributed, it has an approximate normal distribution, which is useful for inferential analyses. In studies comparing IRT-based methodologies designed to detect characteristics of a test (e.g., dimensionality), or differentially functioning items (DIF) or persons [e.g., appropriateness measures (Levine & Rubin, 1979)], percentages reflecting detection rates can be used as outcomes. Of course, multiple criterion variables in IRT MC studies are desirable because they can provide complementary evidence of the effect of an independent variable (e.g., RMSD and the correlation between true and estimated parameters) although, as noted by Naylor et al. (1968), too many outcome measures may decrease the efficiency of the study and increase the occurrence of chance differences.

The correlation between estimated and true parameters is also used as a criterion variable in MC studies. An advantage of using correlations is that variables with different metrics can be correlated to provide evidence about the factors being manipulated; for example, the correlation between estimated parameter values and their standard errors. A disadvantage is that these correlations only reflect the rank ordering of the variables being correlated and, as

such, produce only relative evidence of the effects of the independent variables. For example, a correlation of .9 between true and estimated a parameters implies that, on average, true a values that exceed the mean of the true as are associated with estimated a values that are above their mean. But that does not guarantee that the true and estimated a values will be close in value, nor is it clear how much better the estimation is for a correlation between true and estimated values of, for example, .8 versus .9. Finally, the assumptions underlying valid interpretations of these correlations (e.g., linearity, homoscedasticity, no truncation or outliers) may not be routinely satisfied when these indexes are used.

Selecting the Number of Replications

The number of replications in a MC study is the analogue of sample size, and criteria used to guide sample size selection in empirical studies apply to MC studies. In IRT research, the number of replications is influenced by the purpose of the MC study, by the desire to minimize the sampling variance of the estimated parameters, and by the need for statistical tests of MC results to have adequate power to detect effects of interest.

The purpose of the study has an important effect on the number of replications selected. For example, a parameter recovery study in which the significance of a statistic is tested must generate empirical sampling distributions for the statistics. Thus, a fairly large number of replications may be needed (e.g., 500). However, when comparing IRT-based methodologies (e.g., comparing the number of DIF items correctly detected by competing methods), empirical sampling distributions are not necessarily obtained and a small number of replications may be sufficient (e.g., 10).

Replications and precision. The number of replications also has a direct influence on the precision of estimated parameters—larger samples (i.e., more replications) produce parameter estimates with less sampling variance. The importance statisticians attach to minimizing sampling variance is reflected in the fact that MC studies in statistics typically use several thousand replications (Harwell, Rubinstein, Hayes, & Olds, 1992). Unfortunately, IRT-based MC research has lagged behind, typically using no replications (e.g., Hambleton, Jones, & Rogers, 1993; Harwell & Janosky, 1991; Hulin, Lissak, & Drasgow, 1982; Qualls & Ansley, 1985; Yen, 1987). The danger in using no replications is that the sampling variance will be large enough to seriously bias the parameter estimates, reducing the reliability and credibility of the results.

Among the techniques available to reduce the variance of estimated parameters (see Hammersly & Handscombe, 1964; Lewis & Orav, 1989), increasing the number of replications is particularly attractive. The advantages of replicated over unreplicated IRT MC studies are the same as those that accrue in empirical studies; that is, aggregating results over replications produces more stable and reliable results. For example, consider a parameter recovery study in which estimated and true parameters are to be compared. When no replications are used, the only information available is from a single parameter estimate, and summary statistics such as the RMSD can only be calculated across the estimated parameters (e.g., across the set of test items). The equation for computing the RMSD for estimated a parameters aggregated across n items is

$$\text{RMSD} = \left[\frac{\sum_{i=1}^{n} (\hat{a}_i - a_i)^2}{n} \right]^{1/2}, \qquad (1)$$

where $\hat{a}_i$ is the estimated parameter and a_i is the true parameter. The RMSD values can be compared across the experimental conditions to assess the degree to which the factors affect parameter recovery.

In replicated studies, parameter recovery is generally assessed by comparing the difference between an item parameter estimate and the corresponding parameter value across replications. Gifford & Swaminathan (1990) demonstrated that the mean squared difference for any particular item parameter across $r = 1, 2, \ldots, R$ replications can be separated into two components—one reflecting bias in estimation and the other the variance of the estimates across replications. With respect to a, this may be expressed as

$$\frac{\sum_{r=1}^{R}(\hat{a}_{ir}-a_i)^2}{R}=(\bar{\hat{a}}_i-a_i)^2+\frac{\sum_{r=1}^{R}(\hat{a}_{ir}-\bar{\hat{a}}_i)^2}{R}, \quad (2)$$

where $\bar{\hat{a}}_i$ is the mean of the estimated a parameters for the ith item across R replications. The expression to the immediate right of the equal sign reflects estimation bias. This index provides evidence of the effect on parameter estimation of the conditions being modeled, with smaller values of this index suggesting little effect. The right-most term in Equation 2 reflects the variance of the estimates across replications and functions as an empirical error variance. Smaller values of this index suggest that the estimates are fairly stable and hence reliable, whereas larger values serve as a warning that the estimates may be unreliable. The components of Equation 2 are correlated in the sense that parameter estimates showing little bias would also be likely to show little variance over replications (i.e., estimates are close to parameters), whereas a large bias would be expected to be accompanied by a larger variance (estimates are not close to parameters). Both components of Equation 2 are important in assessing parameter recovery, and can only be distinguished in replicated studies.

Replications and power. Finally, the number of replications should be selected so that the power of statistical tests used to analyze MC results is large enough to detect effects of interest. Stone (1993) used a two-step procedure to study the relationship between number of replications and power. First, a MC study was used to investigate the effects of multiple replications with N (N = 250, 500, 1,000), test length (L = 10, 20, 30), and assumed distribution of θ (D = normal, skewed, platykurtic) as factors. He also investigated the effects of analyzing the results for each item (item level) or aggregated across items (test level). A fixed effects, completely between-subjects factorial design was used, and the RMSD was the dependent variable. Stone used a 2PLM and specified the a and b values for each item. Then, R = 10 item response datasets were simulated for each condition. Graphical displays of the resulting RMSD values were followed by an ANOVA of these values to determine which effects were significant. The magnitude of significant effects was estimated using the squared correlation ratio η^2, computed as the ratio of the sum of squares (SS) of an effect to the total SS. (The log of the RMSD values was used in the ANOVA to better satisfy the assumption of normality).

In the second step, the η^2s obtained from Stone's MC study were used to estimate the power of the ANOVA F test to detect particular effects across different numbers of replications (R = 10, 25, 50, 100) using the empirical η^2s as effect size estimates. The computer program STAT-POWER (Bavry, 1991) was used to estimate power for α = .05, the degrees of freedom (df) for the effect, the total number of cells in the design (27), and the η^2 statistic from Stone's analysis. This procedure allowed changes in the power of the ANOVA F test across R to be examined. Table 6 reports these analyses separately for a and b for both item and test levels. Power was calculated only for effects that were significant at $p < .01$ in Stone's original analysis of the RMSD values. The η^2s reported in Table 6 are from Stone's study.

Consider the power of the F test to detect the main effect of N at the test level for a. For all values of R, the power was 1.0, which is not too surprising given the fairly large η^2 of .42. The power to detect the test length (L) × distribution (D) interaction with η^2 = .02, however, was much lower for R = 10 (.77) than for $R \geq 25$ (1.0). Thus, it would be necessary to use at least 25 replications to have good power to detect this effect. More interaction effects were also detected for the bs than for the as, probably because the RMSD values showed less variability for the bs than for the as. That is, estimated bs were typically more stable than estimated as, and the bs exhibited stronger and fairly unambiguous relationships with many effects (e.g., N). For the test level power analysis, the power of the F test to detect relationships between the RMSD for the as and various effects was sometimes less than that for the bs, as suggested by the nonstatistically significant results for the $N \times D$ and $N \times L \times D$ interactions. Although the a and b parameters had similar and quite small η^2 values for these interactions, only the effects for the b parameters were statistically significant.

The relationship between power and the number of replications was also explored for increas-

Table 6
Statistical Power of the Analyses of Monte Carlo Results as a Function of the Number of Replications for *N*, *L*, and *D*

Level of Analysis and Effect	*a* Parameter					*b* Parameter				
	Number of Replications					Number of Replications				
	10	25	50	100	η^2	10	25	50	100	η^2
Test Level Power Analysis										
N	1.0	1.0	1.0	1.0	.42	1.0	1.0	1.0	1.0	.41
L	1.0	1.0	1.0	1.0	.12	.13	.29	.53	.85	.002
D	.56	.95	1.0	1.0	.01	1.0	1.0	1.0	1.0	.05
L × *D*	.77	1.0	1.0	1.0	.02	.84	1.0	1.0	1.0	.03
N × *D*						.49	.93	1.0	1.0	.01
N × *L* × *D*						.13	.31	.62	.94	.004
Item Level Power Analysis										
(*a* = .83; *b* = .01)										
N	.87	1.0	1.0	1.0	.046	.95	1.0	1.0	1.0	.059
(*a* = 1.9; *b* = 1.7)										
N	.96	1.0	1.0	1.0	.063	.85	1.0	1.0	1.0	.042
L	.23	.23	.87	1.0	.008	.24	.56	.87	1.0	.008
D						.39	.81	1.0	1.0	.014
L × *D*	.18	.44	.78	.98	.008	.23	.59	.91	1.0	.011
N × *D*						.18	.44	.79	.98	.008
(*a* = 3.0; *b* = 2.18)										
N	.84	1.0	1.0	1.0	.043	.90	1.0	1.0	1.0	.05
L	.59	.96	1.0	1.0	.023					
D	.27	.67	.95	1.0	.013	.16	.36	.66	.66	.005

ingly extreme *a*s and *b*s and larger values of *R* at the item level. Table 6 shows that for $a = .83$, for example, the power of the *F* test to detect this effect was the same (1.0) for $R \geq 25$. For $a = 1.9$, however, power values varied across number of replications. The main effect for *L* and the $L \times D$ interaction showed poor power (< .45) for 10 or 25 replications, and good power (.78) for 50 replications. For $a = 3.0$ and $N = 250$, the power to detect a distribution effect (not reported in Table 6) was .29, .53, .72, .84, and .92 for $R = 100$, 200, 300, 400, and 500, respectively. Likewise, the power to detect a distribution effect for $b = -2.18$ was .66, .93, .99, 1.0, and 1.0 across the same values of *R* (also not reported in Table 6). These results suggest that at least 500 replications may be needed in order to detect the effects of manipulated factors when extreme parameter values (e.g., $a = 3.0$) are combined with less than ideal data conditions.

Increasing the number of replications to 500 may not be necessary when *N* and *L* are sufficiently large; less than 100 replications may be adequate under such conditions. One strategy is to use varying numbers of replications in the study, using fewer replications for conditions in which there are no estimation problems and more for conditions in which estimation problems occur. A related approach to minimizing *R* is to fix a value of $R \times N$ and then use values for *N* and *R* that produce this value. For example, if $R \times N$ were set to 25,000, combining $R = 25$, 50, 100, and 200 with $N = 1{,}000$, 500, 250, and 125, respectively, would produce the desired 25,000.

Thus, the number of replications needed to reliably detect effects in MC results is higher when (1) an empirical sampling distribution is needed (e.g., to investigate the properties of a statistic or significance test), (2) interest centers on item level analyses in which sampling variances may be large, (3) the effects increase in complexity (e.g., interactions versus main effects), and (4) model parameters become more extreme (e.g., $a = 3.0$ vs. $a = .83$). Clearly, more research in this area is needed before there is a definitive answer concerning the number of replications that should be used, given the purpose and conditions of a particular MC study. Based on the

available evidence, however, a minimum of 25 replications for MC studies in IRT-based research is recommended.

Formulating the Mathematical Model

Another design-related issue is selecting the model that will govern data generation. The model refers to the abstract representation or description of the process being simulated. Thus, the mathematical model is that underlying data generation.

Selection of an IRT model is, of course, dictated by the specification of the problem. Although IRT models have a common form (Baker, 1992, chap. 2), they vary in how they are implemented according to the nature of the problem. For example, parameter recovery studies for the 2PLM would naturally use this as the underlying mathematical model in the study (e.g., Drasgow, 1989; Harwell & Janosky, 1991; Reise & Yu, 1990). Similar comments hold for other IRT-based research settings that use IRT models whose implementation depends on the nature of the problem; for example, studies of equating methods (e.g., Baker & Al-Karni, 1991; Skaggs & Lissitz, 1988), dimensionality (e.g., Ackerman, 1989; Nandakumar, 1991), person/item goodness of fit (e.g., Holt & Macready, 1989; McKinley & Mills, 1985), adaptive testing/computerized testing (e.g., De Ayala, Dodd, & Koch, 1990), test speededness (e.g., Oshima, 1994), and criterion-referenced assessment (e.g., Plake & Kane, 1991). A particularly large class of modeling problems involves DIF in which the model underlying data generation must be justified with regard to the number of DIF items, the parameters that will reflect DIF and the degree of DIF, whether group differences are modeled, and how the matching criteria should be calculated (e.g., Candell & Drasgow, 1988; Gifford & Swaminathan, 1990; McCauley & Mendoza, 1985; Zwick, Donoghue, & Grima, 1993).

Writing and Validating or Selecting Computer Programs

IRT MC studies often use different computer programs to generate data, estimate model parameters, and calculate outcomes such as the RMSD. For example, Harwell & Janosky (1991) used the GENIRV computer program (Baker, 1989) to generate dichotomous item response data following a 2PLM, BILOG to estimate IRT model parameters, and a locally developed program to compute RMSDs. Efficiency concerns often require that, to the extent possible, these programs be integrated, in which case knowledge of a common purpose programming language such as FORTRAN or PASCAL is helpful.

Each component of the computer program must be evaluated for accuracy. Naylor et al. (1968) posed two questions in this regard:

> First, how well do the simulated values of the ... output variables compare with known historical data.... Second, how accurate are the simulation models predictions of the behavior of the actual system in future time periods? (p. 21)

In the context of IRT MC studies, the first question concerns both the validity of the model or process that is being simulated and the validity of the generated values. In most cases, the process of responding to items is inherent in the IRT model being studied and therefore need not be validated. Still, if the process is intended to reflect data conditions observed in practice, such as DIF or test speededness, it is important that evidence of this be provided. Documenting the validity of the generated values requires verifying that the program produces the correct numbers.

Naylor et al.'s (1968) question about predicting future behavior has received surprisingly little attention in published IRT MC studies. Essentially, this is concerned with how well the simulation results stand up over time in empirical studies. For example, the results of Harwell & Janosky (1991) suggested that a small prior variance for as leads to stable parameter estimates with only $N = 150$. Validation evidence here might take the form of results from empirical studies that stable estimates were obtained under these conditions. This kind of validation evidence is more difficult to obtain than, for example, evidence of model-data fit, but it is more critical.

Generating Item Responses

Generating item responses begins with a numerical seed to initiate the generation of random num-

bers, which are translated to response probabilities and then to discrete responses. These steps typically occur within the same computer program.

Choice of seed values. Data generation is initiated by a number or seed value provided by the user when prompted, or supplied by the data generation program (usually using the computer's internal clock). The advantage of a user-supplied number is that it is easy to reproduce the same item responses later, perhaps as a check on the software. This is also true for clock time if this number is reported; however, to do this requires saving the binary version of clock time.

An important issue in the choice of seed values relates to the sampling error that inevitably accompanies the generation of random numbers. As noted earlier, in IRT parameter estimation this increases the variance in parameter estimates. One technique to reduce chance variation is to use common item parameters and common seed values whenever possible in generating datasets. For example, if item responses for 20- and 30-item tests were to be generated, the model parameters used for the 20-item case could serve as the model parameters for the first 20 items of the 30-item case, and the seed used to initiate the simulation of item responses would be the same for both test lengths. This lowers the "noise" in the simulated data and helps to minimize the effects of random error on the parameter estimates; but these data then are dependent, making inferential analyses of the MC results more complex.

Another technique is to simulate a large number of responses and then randomly sample from this population, with each sample serving as a replication. For example, suppose 10 replications were to be generated for a 30-item test and $N = 200$. Rather than using multiple seeds to generate 10 datasets reflecting the responses of 200 examinees to 30 items, a single seed could be used to generate a dataset containing the responses of $N = 1{,}000$ to 30 items; 10 samples of size 200 then would be randomly sampled. Using one seed as opposed to 10 should help minimize random error. Using different model parameters for the 20- and 30-item cases and different seed values eliminates the dependency in the responses but would be expected to increase the random error compared to using the same model parameters and seed value. The preferred method is to simulate all of the needed datasets using different model parameters but a common seed value.

Generating random numbers. Uniformly distributed random numbers are the most frequently generated, and congruential generators are the most widely used for producing uniform random numbers (Lewis & Orav, 1989). A congruential generator makes use of modular arithmetic in producing random integers that depend only on the previous integer. The random integers extend from 0 to a number m, which is typically defined in the random number generation software. The integers running from 0 through m constitute a cycle and the length of a cycle is known as its period. Dividing the randomly generated integers by the modulus m produces (0,1) uniform variates (Lewis & Orav, 1989, p. 75).

The literature describing the statistical properties and efficiency of congruential generators indicates that they demonstrate sufficiently long period lengths (i.e., the length between repeated sequences of random numbers) and produce results with desirable statistical characteristics, such as randomness (Newman & Odell, 1971; Ripley, 1987). Standard-normal variates are widely used in MC studies and are usually generated by transforming uniform random numbers to a N(0,1) form. Among the algorithms available to perform this transformation, the Box-Muller (1958) algorithm, Marsaglia's (1964) polar method, and the ratio-of-uniforms method seem to do equally well (Ripley, 1987).

Table 7 summarizes several available computer packages for generating random numbers from a specific distribution. These computer packages are not specific to IRT, and additional software would have to be used to generate the desired item responses. To help organize the information in this table, readers might wish to follow the suggestions of Ripley (1987) for selecting a data generation algorithm:

> (a) The method should be easy to understand and to program... (b) The programs produced should be compact... (c) The final code should

> execute reasonably rapidly... (d) The algorithms will be used with pseudo-random numbers and should not accentuate their deficiencies. (p. 53)

For example, the computer program MINITAB (1989) might be used to generate uniformly distributed random numbers that are subsequently translated into standard-normal deviates that serve as θ parameters. MINITAB is widely known and easy to understand and use, whereas other programs are perhaps less well known and more difficult to use [e.g., LISREL (Jöreskog & Sörbom, 1989)].

The IMSL (IMSL Inc., 1989) collection of FORTRAN subroutines appears to be the most complete package in terms of variety of random number generators, but the user must develop a program that incorporates the necessary routines. MINITAB offers a wide selection of univariate distributions; however, SAS (SAS Institute Inc., 1988), SPSS Windows (SPSS Inc., 1993), and SYSTAT (SYSTAT Inc., 1991) offer, by comparison, fewer subroutines for generating data. An often overlooked feature of the microcomputer version of LISREL (Jöreskog & Sörbom, 1989) is the availability of routines for generating multivariate non-normal distributions. With the exception of SYSTAT, which is available only on microcomputers, the other programs are available for both mainframe computers and microcomputers.

In addition to the software packages and subroutines mentioned above, there are collections of subroutines for random number generation provided in textbooks. Press, Flannery, Teukolsky, & Vetterling (1986), for example, provided routines to generate random numbers in PASCAL and FORTRAN.

Table 7
Software Packages for Generating Various Distributions

	Software Package and Subroutines					
Distribution	IMSL	MINITAB	SAS	SYSTAT	LISREL	SPSS Windows
Uniform (Continuous)	RNUN RNUNF	UNIFORM	UNIFORM RANUNI	URN	GENRAW2	RV.UNIFORM
Normal	RNNOA RNNOF RNNOR	NORMAL	NORMAL RANNOR	ZRN	GENRAW	RV.NORMAL
Gamma	RNGAM	GAMMA	GAMMA RANGAM	GRN		RV.GAMMA
Exponential	RNEXP	EXPONEN	RANEXP	ERN		RV.EXP
χ^2	RNCHI	CHISQUARE		XRN	GENRAW2	RV.CHISQ
Logistic		LOGISTIC				RV.LOGISTIC
Lognormal	RNLNL	LOGNORMAL				RV.LNORMAL
Laplace		LAPLACE				RV.LAPLACE
Cauchy	RNCHY	CAUCHY	RANCAU			RV.CAUCHY
t		T		TRN		RV.T
F		F		FRN		RV.F
Weibull	RNWIB	WEIBULL				RV.WEIBULL
Beta	RNBET	BETA		BRN		RV.BETA
Triangular	RNTRI		RANTRI			
Uniform (Discrete)	RNUND	INTEGER			GENRAW2	
Binomial	RNBIN	BINOMIAL	RANBIN		GENRAW2	RV.BINOM
Poisson	RNPOI	POISSON	POISSON RANPOI			
Multinomial	RNMTN					
Multivariate Normal	RNMVN			GENRAW		
Multivariate Non-normal				GENRAW2		

The accuracy of randomly generated numbers depends on the random number generator and the computer that is used: It is strongly recommended that researchers test the generator and document its capabilities (Hauck & Anderson, 1984; Hoaglin & Andrews, 1975). Even well-known generators, such as those associated with IMSL, should be tested to ensure their adequacy for a particular application. Locally developed generators, however, require extensive testing and documentation, and the absence of this information means that the credibility of the study may be suspect. Tests of the adequacy of random number generators include the Kolmogorov-Smirnov and Cramer-von-Mises goodness-of-fit procedures, the serial and up-and-down tests, the gap and maximum tests, and a test for randomness (Rubinstein, 1981). However, Wilcox (1988) cautioned researchers to not expect a sequence of random numbers to pass all tests and to not use the same sequence more than once.

Transforming random numbers into item responses. Initially, a vector of N θ values is generated by sampling from a specified distribution for a given number of examinees. These θs are often sampled from a N(0, 1) distribution for studies involving unidimensional IRT models, or from a multivariate normal distribution with various values for the intercorrelations for a study involving multidimensional IRT models (non-normal distributions can also be used).

Table 8 compares several computer programs for generating item response data on various characteristics. [Kirisci, Hsu, & Furqon (1993) provided a detailed comparison of the performance of the DATAGEN, GENIRV, and ITMNGR programs listed in Table 8.] The advantage of these programs over those in Table 7 is that conditions particular to IRT have been incorporated into the algorithms. For example, the 3PLM has been incorporated into the programs in Table 8 and can be used to produce responses following this model, with a few commands. However, the programs in Table 8 do not generate skewed or platykurtic distributions in either the univariate or multivariate cases, whereas the programs in Table 7 can be used to generate N(0, 1) variates that can be transformed to univariate and multivariate non-normal distributions using procedures described by Fleishman (1978) and Vale & Maurelli (1983).

The N θs are used to produce an $N \times n \times K - 1$ matrix of response probabilities for N examinees, n items, and $K - 1$ response categories for the test. For example, in the case of a 3PLM with dichotomous response data, the response probabilities (P_i, Q_i) are obtained from

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{\exp\left[a_i D(\theta_j - b_i)\right]}{1 + \exp\left[a_i D(\theta_j - b_i)\right]}, \qquad (3)$$

where

$P_i(\theta_j)$ is the probability of a correct response for the ith item, conditional on the θ of the jth examinee,

c_i represents a guessing parameter,

D represents a scaling constant that brings the logistic and normal ogives into agreement (Lord & Novick, 1968, p. 400), and

$Q_i = 1 - P_i$.

The response probabilities are then translated into discrete item responses by comparing the probabilities with random numbers drawn from a uniform distribution. In the dichotomous response case, if the probability of a correct response for an examinee to an item is greater than or equal to the random number, then a 1 is typically assigned to that item; otherwise a 0 is assigned. In the polytomous response case, if the random number falls between response category k and $k + 1$, the item response $k + 1$ is assigned to the item. Rather than comparing two boundaries to translate category response probabilities into a polytomous response, cumulative probabilities can also be used (Ankemann & Stone, 1991). In all cases, the process is repeated with different random numbers for each item and for all examinees.

Examples of generating item responses. In the Harwell & Janosky (1991) parameter recovery study, unidimensional dichotomous response data following a 2PLM were simulated for, among other conditions, $N = 500$ examinees and $n = 25$ items. The as and bs for each item were sampled at random from a uniform and normal distribution, respectively. Using GENIRV, the data generation con-

Table 8
Computer Programs for Generating Item Responses

Features	DATAGEN	GENER	GENIRV	ITMGNR	RESGEN	YeomanDG
Reference	Hambleton & Rovinelli (1973)	Carlson (1993)	Baker (1989)	Kirisci (1992)	Muraki (1992)	DeAyala (1993)
Maximum Number of Items	400	N/A	100	500	100 (Default)	150
Maximum Number of Examinees	4,000	N/A	4,000	10,000	1,000 (Default)	32,000
PC or Mainframe	Mainframe	PC	PC	Both	PC	MAC II
Type of Response						
Dichotomous	YES	YES	YES	YES	YES	YES
Graded	NO	NO	YES	NO	YES	YES
Nominal	NO	NO	YES	NO	YES	YES
Item Response Model						
Normal Ogive	YES	YES	YES	NO	NO	N/A
Logistic	YES	YES	YES	YES	YES	YES
Distribution	1. Normal 2. Uniform 3. User Specified Parameter Values	1. Normal	1. Normal 2. Uniform	1. Normal 2. Uniform 3. Gamma 4. χ^2 5. Exponential 6. Beta 7. Non-normal 8. User Specified Parameter Values 9. Uniform Over Subintervals	1. Normal 2. Uniform 3. Lognormal 4. Gamma 5. Multivariate Normal	1. Normal 2. Uniform 3. Beta
Options	Parallel Tests	Multidimensional Model			1. Multiple Sampling 2. Multiple Blocks 3. Multiple Subtests 4. Multiple Groups 5. Multidimensional Model	Multidimensional Model

sisted of creating a vector of 500 θ levels sampled at random from a N(0, 1) distribution. Next, a 500 × 25 matrix of response probabilities was created by GENIRV by computing, for each simulated examinee, $P_i(\theta_j)$ using Equation 3 (with $c = 0$). Finally, each of the $P_i(\theta_j)$ values in the 500 × 25 matrix was compared to a randomly selected value from a uniform distribution to generate a 1 (correct response) or 0 (incorrect response).

Ansley & Forsyth (1985) studied the effect on parameter estimation assuming unidimensionality when the data were multidimensional. Two-dimensional ($M = 2$) dichotomous response data following Sympson's (1978) two-parameter multidimensional IRT model (with $c = .2$) were generated for, among other conditions, $N = 1{,}000$ examinees and $n = 30$ items. Initially, $M \times N$ θ values were sampled at random from a bivariate normal distribution with specified correlation between the two θ variables. Next, a 1,000 × 30 matrix of response probabilities was created using Sympson's model by computing, for each examinee, $P_i(\theta_j)$ using selected as and bs (each item had two a parameters and two b parameters). Finally, each of the $P_i(\theta_j)$ values in the 1,000 × 30 matrix was compared to a randomly selected value from a uniform distribution to generate 1s or 0s.

The last example of generating item responses involves DIF. Most MC studies define DIF as two groups of examinees showing differences on item parameters (or a function of item parameters) prior to generating item responses. For example, Rudner, Getson, & Knight (1980) altered the standard deviation of group differences on item parameters; Swaminathan & Rogers (1990) simulated DIF by varying the group item response functions; and Candell & Drasgow (1988) and Furqon & Hsu (1993) simulated DIF by increasing or decreasing values of item parameters before generating item responses for one group.

Each of these approaches has merit, but for illustrative purposes, the procedure used by Furqon & Hsu (1993) is presented in which data for two groups of 500 examinees each were simulated for a 60-item test (12 items showed DIF) for a 3PLM with $c = .2$. Data were simulated for each group separately, and the a and b parameters were selected at random from specified distributions. Initially, a vector of 500 θ values for the first group were sampled from a N(0, 1) distribution, then another 500 θ values were sampled from a N(.5,1) distribution for the second group. A constant was then added or subtracted to the a and b parameters for 12 items for the first group to produce the desired DIF. The a and b parameters for the second group were not modified. Next, a 500 × 60 matrix of response probabilities was created and translated into item responses for each group using the procedure described above.

The above examples illustrate that the procedure for generating item responses is common to a variety of research problems in IRT. This procedure can be replicated to produce multiple datasets based on the same conditions; that is, a new set of θ values can be generated and response probabilities computed and translated, and so on.

Estimating Model Parameters

Once item response datasets have been generated, the next step may be to estimate the model parameters. Comprehensive descriptions of various estimation methods can be found in Baker (1987, 1992). Unlike other steps in implementing an IRT MC study, the research questions may not dictate the estimation method. Researchers can rely on commercial item analysis programs, such as BILOG or MULTILOG (Thissen, 1986) or XCALIBRE (Assessment Systems Corporation, 1995), to estimate model parameters or can write their own software. If the latter method is used, it is critical that the adequacy of the estimation algorithm be extensively documented. Validation evidence could take the form of using the program to analyze a well-known dataset such as the LSAT-6 data (Bock & Lieberman, 1970) and comparing the results to published parameter estimates, or analyzing item responses that show a perfect fit with an item response function (i.e., the empirical proportions of correct responses fall exactly on the function), in which case the estimation program should return the exact item parameters.

Two issues are especially pertinent in selecting or writing an item analysis program to estimate parameters: The handling of starting values and non-con-

vergent solutions. All of the estimation procedures involve iterative algorithms and hence require starting values for the parameters in the algorithm. Item analysis programs have default starting values but most offer users the option of specifying starting values. In many cases these default values are sufficient, particularly for well-structured datasets with a large number of examinees (e.g., $N = 1{,}000$). If the computer time needed to estimate model parameters is excessive, one strategy is to use the parameter values as the starting values (assuming that comparing estimation methods is not the purpose of the MC study). Stone (1992) used parameters as starting values in an evaluation of the marginal maximum likelihood estimation procedures in MULTILOG, citing Bock's (1991) discussion that the choice of starting values is not critical in the EM estimation algorithm. However, it is important to establish that the solution is not dependent on particular starting values (Mislevy, 1986).

A second issue is how nonconvergence is handled. If the estimation algorithm fails to converge to a solution in the MC study, researchers can (1) ignore the nonconvergence but use the estimates only after a large number of iterations were performed, (2) exclude the estimates in the calculation of summary statistics such as RMSDs, or (3) use a different estimation algorithm [e.g., a fully Bayesian approach (Mislevy, 1986)] to constrain values of the estimated parameters. It is also important to equate the parameter estimates to the metric of the parameter values; otherwise, the estimates may be biased (Baker, 1990).

Problems of poorly specified starting values and nonconvergent solutions are likely to be more pronounced with smaller datasets (e.g., $N = 200$ and a 40-item test) and more complex IRT models (e.g., the 3PLM). These problems can arise in both commercial and locally-developed estimation programs. The fully Bayesian approach can help to mediate these difficulties by careful specification of prior distributions and the parameters of these distributions (i.e., hyperpriors and hyperparameters). The ASCAL program (Assessment Systems Corporation, 1988) implements a Bayesian modal approach to IRT item parameter estimation. There is, however, little evidence that a fully Bayesian approach possesses strong practical advantages over other estimation methods (Kim et al., 1994).

Analyzing the Results of a Monte Carlo Study

Based on the research questions and the experimental design, statistical hypotheses and data analysis procedures can be determined. As documented by Hsu (1993), however, many MC studies do not use any discernible experimental design—which can have several negative consequences.

One is that analyses of MC results often consist entirely of tabular summaries, simple descriptive statistics, or graphical presentations. This seems to be an unreliable way of detecting important effects and of estimating the magnitude of effects—a problem that is exacerbated when the amount of reported data is substantial. For example, Ansley & Forsyth (1985) reported 144 values in four tables, Kim et al. (1994) reported 240 values in three tables, Harwell & Janosky (1991) reported 360 values in two tables, and Yen (1987) reported 1,639 values in 10 tables. Readers are left attempting to verify an author's conclusions by looking at hundreds (and possibly thousands) of values, which is a daunting task. Harwell (1991) argued that this increases the chance that important effects will go undetected and that the magnitude of effects will be misestimated; thus, the solution is to use both descriptive and inferential analyses.

Inferential Analysis of Monte Carlo Results

A variety of inferential analyses can be performed, but regression and ANOVA methods are often suitable (Harwell, 1991). Naturally, if the independent variables are nominal (e.g., item analysis program) then ANOVA may be more appropriate, whereas for metric variables (e.g., number of examinees) a regression approach may be preferred. Regression methods are somewhat more attractive because most independent variables in IRT MC studies are metric, and because nonlinear and heteroscedastic prediction models are readily available. Still, there seems to be a preference for ANOVA among IRT researchers (e.g., De Ayala, 1994; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993), perhaps because of a prefer-

ence for means over regression coefficients (of course, under certain conditions the two procedures yield the same results).

The population regression model can be written as

$$\tau_s = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_T X_T \quad (4)$$

and

$$\hat{\tau}_s = \tau_s + \varepsilon_s \quad (5)$$

(Timm, 1975, pp. 267–268). In Equation 4, τ_s represents the sth outcome (e.g., RMSD) that depends on a set of T fixed predictor variables X_s ($s = 0, \ldots, T$), β_0 is an intercept, β_T represents a population regression coefficient that captures the relationship between a predictor variable and the outcome, ε_s is an error term, and $\hat{\tau}_s$ is an observed outcome. The estimated model is

$$\tau_s = b_0 + X_1 b_1 + X_2 b_2 + \ldots + X_T b_T . \quad (6)$$

Standard normal theory regression assumes that the errors are independent in the sense that the generated data would be declared independent by a statistical test of independence; independence of the τ_ss may be assumed by virtue of the random number generator. Inferential analyses also require that the τ_ss be normally distributed (Timm, 1975, p. 267). In addition to relying on the well-documented robustness of normal theory tests to non-normal distributions, it may be useful in some cases (and necessary in others) to perform a nonlinear transformation on the $\hat{\tau}_s$s to increase the likelihood that the normality assumption is approximately satisfied.

For example, rather than working with RMSD, the log(RMSD) could be analyzed. A log transformation of a standard deviation results in values that, if the values used to compute the RMSD (e.g., estimated a parameters) are themselves normally distributed, are asymptotically normally distributed with a known mean and variance that depends on the number of replications (Bartlett & Kendall, 1946). Heteroscedasticity can be handled with weighted least squares, provided reasonable estimates of the unknown variances are available. If the assumptions of independence and normality are tenable, the F test can be used to test whether there is a relationship between the dependent variable and the set of predictors. If transformations are unsuccessful in reducing skewness in the dependent variable, or if the distribution of this variable can only be arbitrarily specified, nonparametric procedures that do not require normality can be used (Harwell & Serlin, 1989).

An example. Data from the Harwell & Janosky (1991) study are used to illustrate regression and ANOVA approaches to analyzing IRT MC results. Recall that Harwell and Janosky simulated dichotomous response data using a 2PLM, estimated model parameters with BILOG, and compared the estimated as and bs for each item to the true values using RMSD.

Relying on traditional descriptive methods, Harwell and Janosky concluded that, for the as and bs, (1) when $N > 250$, the prior variance had little effect on the accuracy of estimation for both 15- and 25-item tests; (2) when $N < 250$ and a 15-item test was used, the prior variance played a large role in the quality of the estimation, with smaller prior variances offering better accuracy; and (3) for a 25-item test, the effect of the prior variance on the accuracy of estimation was neutralized when $N > 100$. These conclusions suggest a prior variance $\times$ N interaction.

The design of this study, as noted earlier, was a $6 \times 2 \times 5$ factorial with number of examinees (N), test length (L), and prior variance (PV) serving as independent variables and RMSD as the dependent variable. The analysis of the Harwell and Janosky results is complicated by the fact that the same random seed was used for the 15- and 25-item cases; however, for illustrative purposes the results for the 15- and 25-item tests were analyzed together. Only results for the as are presented. A log-transformation of the RMSD, log(RMSD), was used to increase the likelihood of satisfying the assumption of normality needed in hypothesis testing. Because all three independent variables were metric, a regression analysis was conducted. The results are reported in Table 9.

Initially, a (main effects) model with the predictors N, L, and PV was fit to the log(RMSD) values for the as (Model 1a), followed by the three possible, two-variable-at-a-time interaction predictors (Model

Table 9
Regression *df* (df_R) and Sum of Squares (SS_R), Residual *df* (df_E) and Sum of Squares (SS_E), and R^2 for the Estimated *a*s of Harwell & Janosky (1991), Under Several Main Effects (a) and Main Effects Plus Interaction Models (b)

Model	df_R	df_E	SS_R	SS_E	R^2
1a	3	56	9.32	5.33	.62
1b*	6	53	11.00	3.63	.72
2a	3	16	.53	.66	.34
2b*	6	13	.98	.21	.75
3a	2	12	2.61	.86	.71
3b	3	11	2.82	.65	.76
4a	2	17	2.94	.30	.90
4b	3	16	2.97	.25	.91

Note. All models were significant at $p < .05$.
*Indicates a significant difference between the main effects and main effects + interactions models.

lb). Each predictor was centered to minimize colinearity problems. Because this study was not replicated within cells, estimates of the highest-order interaction could not be obtained. The results indicated a strong relationship between log(RMSD) and the prediction models, with the R^2s suggesting that this set of predictors was sensitive to variations in log(RMSD) (although the contribution of the interactions appeared to be modest). Note that the R^2s were adjusted for the number of predictors (see Draper & Smith, 1981, pp. 91–92). Thus, the accuracy of estimating *a*s in BILOG appeared to depend heavily on these predictors. The estimated (standardized) slopes $b_{PV} = -1.08$ and $b_{N\times L} = -.69$ suggests that these variables played especially prominent roles. (Each estimated regression coefficient was tested using $\alpha = .01$.) The significant $N \times PV$ interaction term helped to clarify the general conclusion of Harwell and Janosky; that is, the effect of the prior variance on estimation depended on *N*.

The regression analysis also indicated that this interaction accounted for 4% of the variance of log(RMSD), suggesting that this is probably not as critical a factor in the accuracy of estimating *a*s as suggested by Harwell and Janosky. To extract additional information about the role of *PV*, which figured prominently in the conclusions of Harwell and Janosky, Models 1a, 1b, and 2 were repeated for $N > 250$; Model 3 for $N < 250$ and a 15-item test; and Model 4 for $N > 100$ and a 25-item test (see Table 9). For $N > 250$, *PV* accounted for 1% of the variance with $b_{N\times L} = -2.97$, $b_N = 1.35$; for a 15-item test and $N < 250$, *PV* accounted for 21% of the variance with $b_{PV} = -1.62$; for a 25-item test and $N > 100$, *PV* accounted for 0% of the variance with $b_N = -1.19$.

The results of Harwell and Janosky were also analyzed using completely between-subjects factorial ANOVA. These results are reported in Table 10 and are similar but not identical to the regression results. Using $\alpha = .05$ for each hypothesis tested, all seven effects were significant. Among the interactions, the two-way $N \times PV$ interaction effect accounted for the most variance [$\eta^2 = SS_{(N\times PV)}/SS_{Total} = 13\%$]. Following the advice of Rosnow & Rosenthal (1989), a graph of the residual cell means (i.e., after removing variation due to the other interactions and the main effects) suggested that smaller *N*s need smaller prior variances to maintain the accuracy of the estimation, but for larger *N*s the prior variance seems less important. Tetrad contrasts could be tested in a post hoc analysis to further clarify the nature of this interaction (see Toothaker, 1991). The results of the regression and ANOVA analyses of MC results generally supported all three conclusions of Harwell and Janosky and extended the descriptive results they reported.

Table 10
Results of ANOVA for the Estimated *a*s of Harwell & Janosky (1991)

Source	*df*	*F*	*p*	η^2
L	1	27.5	0.000	.06
N	5	61.8	0.000	.65
PV	4	6.8	.001	.06
$L \times N$	5	3.8	.015	.04
$L \times PV$	4	3.1	.037	.03
$N \times PV$	20	2.9	.009	.13
Error	20			

Conclusions

The importance of MC techniques in IRT research is likely to increase because of their ability to model realistic data conditions and to compare competing statistics or methodologies in ways not possible with empirical data. Along with increases in desktop

computing power, the flexibility of these techniques makes them an increasingly attractive research tool.

MC techniques will continue to make a contribution to important problems in IRT. Any problem in which analytic solutions are unwieldy or impossible is a good candidate for a MC study. The list is extensive, but a sampling of problems includes information about the number of examinees necessary to produce stable parameter estimates, the comparison of methodologies used in conjunction with IRT, such as procedures designed to detect multidimensionality and differential item functioning, and the properties of goodness-of-fit tests. The litmus test of the usefulness of the results of such studies will be how realistically they model empirical problems.

In the future, replicated IRT MC studies should become the rule rather than the exception. These studies also should be expected to use increasingly sophisticated experimental designs and data analyses and to expand the number of conditions modeled. These advances should increase the value of MC studies in IRT research.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113–127.

Ankemann, R. D., & Stone, C. A. (1991, April). *More results on parameter recovery in the graded model using MULTILOG.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37–48.

Assessment Systems Corporation. (1988). *ASCAL: 2- and 3-parameter IRT calibration program.* St. Paul MN: Author.

Assessment Systems Corporation. (1995). *XCALIBRE: Marginal maximum-likelihood estimation program.* St. Paul MN: Author.

Babbie, E. (1989). *The practice of social research* (5th ed.). Belmont CA: Wadsworth.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11,* 111–141.

Baker, F. B. (1989). *GENIRV: A program to generate item response vectors.* Unpublished manuscript, University of Wisconsin, Laboratory of Experimental Design, Madison.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement, 14,* 139–150.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147–162.

Bartlett, M. S., & Kendall, D. G. (1946). The statistical analysis of variance heterogeneity and the logarithmic transformation. *Journal of the Royal Society,* (Supplement 8), 128–138.

Bavry, J. L. (1991). *STAT-POWER: User's guide.* Chicago: Scientific Software.

Bock, R. D. (1991, April). *Item parameter estimation.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Bogan, D. E., & Yen, W. M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. I: Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290–302.

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Journal of the Royal Statistical Society, 29,* 610–613.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253–260.

Carlson, J. (1993). *GENER: Program to generate compensatory MIRT data.* Unpublished program, Educational Testing Service, Princeton NJ.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

De Ayala, R. J. (1993). YeomanDG: A data generation program (Version 1.0). *Applied Psychological Measurement, 17,* 393.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18,* 155–170.

De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1990). A simulation and comparison of flexilevel and Baye-

sian computerized adaptive testing. *Journal of Educational Measurement, 27,* 227–239.

De Vroye, L. (1986). *Non-uniform random variate generation.* New York: Springer-Verlag.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77–90.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199.

Fleishman, A. (1978). A method for simulating nonnormal distributions. *Psychometrika, 43,* 521–532.

Furqon, S., & Hsu, T. C. (1993). *The use of flexible logistic regression when assessing differential item functioning.* Unpublished paper, University of Pittsburgh, Department of Psychology in Education.

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14,* 33–43.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30,* 143–155.

Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic models. *Behavioral Science, 17,* 73–74.

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

Hartley, H. O. (1977). Solution of statistical distribution problems by Monte Carlo methods. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (Vol. III; pp. 16–34). New York: Wiley.

Harwell, M. R. (1991, April). *Analyzing and reporting the results of Monte Carlo studies in educational and psychological research.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15,* 279–291.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17,* 315–339.

Harwell, M. R., & Serlin, R. C. (1989). A nonparametric test statistic for the general linear model. *Journal of Educational Statistics, 14,* 351–371.

Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician, 38,* 214–216.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *American Statistician, 29,* 122–126.

Holt, J. A., & Macready, G. B. (1989). A simulation study of the difference chi-square statistic for comparing latent class models under violation of regularity conditions. *Applied Psychological Measurement, 13,* 221–231.

Hsu, T. (1993, July). *Contributions of Monte Carlo techniques to IRT research: A methodological review.* Paper presented at the European meeting of the Psychometric Society, Barcelona.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6,* 249–260.

IMSL, Inc. (1989). *IMSL library reference manuals.* Houston TX: Author.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7 user's guide.* Chicago: Scientific Software.

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59,* 405–421.

Kirisci, L. (1992). *ITMGNR: A Fortran IV computer program to generate item response data.* Unpublished program, University of Pittsburgh.

Kirisci, L., Hsu, T., & Furqon, S. (1993, July). *Computer resources for Monte Carlo-based IRT research.* Paper presented at the European meeting of the Psychometric Society, Barcelona.

Lehman, R. S., & Bailey, D. E. (1968). *Digital computing: Fortran IV and its applications in behavioral science.* New York: Wiley.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–290.

Lewis, P. A. W., & Orav, E. J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers* (Vol. 1). Pacific Grove CA: Wadsworth and Brooks/Cole.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Marsaglia, G. (1964). Generating a variable from the tail of the normal distribution. *Technometrics, 6,* 101–102.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9,* 389–400.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psycho-*

logical Measurement, 9, 49–57.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 44,* 335–341.

Minitab, Inc. (1989). *MINITAB handbook.* Boston: Duxbury Press.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–194.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Muraki, E. (1992). *RESGEN: Item response generator.* Unpublished program, Educational Testing Service, Princeton NJ.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99–117.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18,* 315–328.

Naylor, T. H., Balintfy, J. L., Burdick, D. S., & Chu, K. (1968). *Computer simulation techniques.* New York: Wiley.

Newman, T. G., & Odell, P. L. (1971). *The generation of random variates.* New York: Hafner Press.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31,* 200–219.

Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement, 28,* 249–256.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes.* Boston: Cambridge University Press.

Psychometric Society. (1979). Publication policy regarding Monte Carlo studies. *Psychometrika, 44,* 133–134.

Qualls, A. L., & Ansley, T. N. (1985, April). *A comparison of item and ability parameter estimates derived from LOGIST and BILOG.* Paper presented at the Annual Meeting of the National Conference on Measurement in Education, Chicago.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Ripley, B. D. (1987). *Stochastic simulation.* New York: Wiley.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116.

Rogosa, D. R. (1980). Comparing non-parallel regression lines. *Psychological Bulletin, 88,* 307–321.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105,* 143–146.

Rubinstein, R. V. (1981). *Simulation and the Monte Carlo method.* New York: Wiley.

Rudner, L. M., Getson, P. M., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213–233.

SAS Institute, Inc. (1988). *SAS user's guide: Statistics.* Cary NC: Author.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12,* 69–82.

Smith, V. K. (1975). *Monte Carlo methods: Their role for econometrics.* Lexington MA: Lexington Books.

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement, 7,* 405–425.

SPSS, Inc. (1993). *SPSS for windows: Base system user's guide.* Chicago: Author.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16,* 1–16.

Stone, C. A. (1993, July). *The use of multiple replications in IRT based Monte Carlo research.* Paper presented at the European Meeting of the Psychometric Society, Barcelona.

Swaminathan, H., & Rogers, J. A. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

SYSTAT, Inc. (1991). *SYSTAT: The system for statistics.* Evanston IN: Author.

Thissen, D. (1986). *MULTILOG user's guide.* Mooresville IN: Scientific Software.

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology.* Monterey CA: Brooks/Cole.

Toothaker, L. E. (1991). *Multiple comparisons for researchers.* Newbury Park CA: Sage.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48,* 465–471.

von Neumann, J., & Ulam, S. J. (1949). Various techniques used in connection with random digits. *Jour-*

nal of Research of the National Bureau of Standards, 12, 16–38.

Wilcox, R. R. (1988). Simulation as a research technique. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 134–137). New York: Pergamon.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST 5.0 version 1.0 user's guide.* Princeton NJ: Educational Testing Service.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52,* 275–291.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Acknowledgments

This paper is a consolidation of four papers presented in a symposium on the use of monte carlo studies in item response theory research at the European meeting of the Psychometric Society, Barcelona, 1993. The authors are grateful to Michelle Liou and Brian Junker for their comments on the original papers. The comments of the editor and the reviewers were also of great value in structuring the paper.

Author's Address

Send requests for reprints or further information to Michael Harwell, 5C01 Forbes Quad, University of Pittsburgh, Pittsburgh PA 15260, U.S.A. Email: harwell@cis.vms.pitt.edu.