Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2016, Vol. 35, No. 2, pp. 36–49

Conducting Simulation Studies in Psychometrics

Richard A. Feinberg and Jonathan D. Rubright*, *National Board of Medical Examiners*

Simulation studies are fundamental to psychometric discourse and play a crucial role in operational and academic research. Yet, resources for psychometricians interested in conducting simulations are scarce. This Instructional Topics in Educational Measurement Series (ITEMS) module is meant to address this deficiency by providing a comprehensive introduction to the topic of simulation that can be easily understood by measurement specialists at all levels of training and experience. Specifically, this module describes the vocabulary used in simulations, reviews their applications in recent literature, and recommends specific guidelines for designing simulation studies and presenting results. Additionally, an example (including computer code in R) is given to demonstrate how common aspects of simulation studies can be implemented in practice and to provide a template to help users build their own simulation.

Keywords: psychometrics, research design, simulation study

Analyzing examinee responses to test questions is often indispensable when making educational policies. Many government-sponsored initiatives collect data for this very purpose—for example, the National Assessment of Educational Progress (NAEP); however, real data can be time consuming and costly to collect and it is often incomplete. Moreover, even when data are complete and readily available, they may still be inadequate to support the analyses, inferences, and conclusions intended by the researcher—indeed, some research questions simply cannot be answered using empirical data. It is on these occasions, when it is impractical to collect the necessary empirical data, or when a given research question requires knowledge of the "true" parameters, or when the ability to manipulate study conditions are unsupported by real data, that simulating data may be necessary.

Simulation studies gained popularity in the late 19th and early 20th century, and have led to a number of noteworthy advances across a broad range of practical fields—from the quality of beer (perhaps the first simulation; Student, 1908) to nuclear warfare (von Neumann & Ulam, 1949). In psychometrics, particularly in the context of item response theory (IRT), simulations are widely used to investigate a broad range of research questions from evaluating the success of various models under different conditions (e.g., small sample conditions) to the robustness of fitting unidimensional models to multidimensional data (Harwell, Stone, Hsu, & Kirisci, 1996). In some cases, simulations have also been profitably used for routine activities such as evaluating model–data fit (Hambleton, Swaminathan, & Rogers, 1991). Recently, simulation methods were highlighted at the 2012 National Council on Measurement in Education (NCME) annual meeting in a coordinated session entitled "Design Issues in Equating Studies using Simulation and Resampling" (Carlson, 2012; Dorans, 2012; Livingston, 2012; Walker, 2012).

The popularity and utility of simulations suggests that researchers and practitioners in measurement should be familiar with their appropriate use, the kinds of questions that simulations can answer best, and the most effective way to present or communicate simulation results. Nevertheless, students in graduate level measurement programs may have limited exposure to simulation techniques or applications during their training, and published resources and guidelines tailored to the needs of measurement specialists are scarce. Moreover, those few resources that are available tend to be written for advanced readers or to be focused on software and techniques that are now considered obsolete (e.g., Harwell et al., 1996; Spence, 1983; Wilcox, 1988).

The purposes of this ITEMS module are fourfold: we (i) introduce readers to the vocabulary and applications specific to simulations; (ii) summarize current simulation research by reviewing recently published articles in educational and psychological measurement; (iii) recommend specific guidelines for designing simulation studies and summarizing the results; and (iv) illustrate how to conduct a simulation study with an example—including sample computer code in R (R Core Team, 2015). In summary, the intent of this article is to provide readers with a comprehensive introduction to the topic of simulation that can be easily understood by measurement specialists at all levels of training and experience.

Uses of, and Common Terms in, Simulation

In a simulation, data are created by researchers based on a model. This model can take any form—as simple as sampling from a normal distribution with a given mean and standard deviation, or, as complex as generating responses based on a multidimensional IRT model. One advantage of simulation studies is that they allow researchers to compare *estimated*

Richard A. Feinberg and Jonathan D. Rubright, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; rfeinberg@nbme.org; jrubright@nbme.org.
**Work completed while at the American Institute of Certified Public Accountants.*

parameters against their respective *true* parameters, which are unknown for real data applications.

Because true parameters are not known in practice, they must be decided upon by the researcher prior to simulating data. This creates the opportunity to specify a model that may or may not closely align with, or generalize to, reality. Thus, despite the benefits of simulation, they should be approached with caution and skepticism, ever mindful that unlike more data collection based scientific experiments, simulations, even when constructed according to an experimental design, are deductive demonstrations that are consistent with the assumptions used to generate the data (Davey, Nering, & Thompson, 1997). This stands in contrast to the inductive reasoning that arises when real data are used to approximate the truth (Dorans, 2012). In some cases, the extent to which findings based on simulated data generalize to practice will depend on how well the simulated data conform to the empirical data that arise in practice. Of course, findings from real data are also limited in generalizability based on how those data were collected.

A simulation may also be referred to as a *computer simulation* or a *Monte Carlo (MC) simulation*. Nicholas Metropolis coined this term MC after the Monte Carlo casino, which his colleagues' uncle was known to frequent (Metropolis, 1987). Simulations share some characteristics with resampling methodologies such as the *bootstrap* and the *jackknife* (Wu, 1986). For bootstrapping, random samples are drawn from an empirical data set to create a sampling distribution for the parameter of interest. Under jackknifing, the parameter is estimated using the actual data with one or more observations removed. Using resampling techniques, the true parameters remain unknown and the researcher must identify another criterion by which to evaluate estimated quantities such as large sample estimates. A related technique, Markov Chain Monte Carlo (MCMC), uses random number selection to solve difficult modeling problems (Gelfand & Smith, 1990) and was introduced to the field of psychometrics by Patz and Junker (1999a,b). Although parts of the discussion below may apply to any of these approaches, the remainder of this article focuses on simulation.

A number of terms are commonly used in simulations, and those terms are reviewed here. First, a *factor* in a simulation is akin to a factor in an analysis of variance (ANOVA) study: it is an independent variable, usually categorical, that is manipulated by the researcher. For example, if the simulation generates and studies small, medium, and large sample sizes, then sample size is a *factor* of study. *Conditions*, sometimes referred to as levels, are realizations of *factors*. Thus, if sample size is the *factor* of interest, "small" would be one of the *conditions*. Often, *factors* are *crossed*, meaning that each *condition* from each *factor* is realized. So, if a sample size *factor* (small, medium, and large) is *crossed* with a dimensionality *factor* (unidimensional vs. multidimensional), there would be a total of six *conditions* (small unidimensional, medium unidimensional, large unidimensional, small multidimensional, medium multidimensional, and large multidimensional). In sum, the language typically employed in experimental design carries over to simulation studies. Within each combination of *conditions*, often researchers simulate responses for *simulees*: simulated examinees. Finally, within each *condition*, there are often a certain number of *replications*. That is, for many research questions, it may be necessary to repeat the data generation process over multiple occasions—that is, *replications*—so that the empirical estimate of the sampling distribution (due to the simulation) of various statistics of interest may be observed. Thus, the number of replications refers to the number of times the entire experiment is repeated. As will be discussed later, determining the number of replications in a simulation study may require careful consideration by the researcher. Also, note that the language described above may not be universal (e.g., factor analysis simulation studies would likely avoid the term factor to describe choices in the simulation design), but is widely used across simulation studies within the psychometric literature.

Knowing the true parameters creates the opportunity to examine the properties of the errors associated with a given estimator. Three common ways to summarize these errors are to compute the average deviance between the estimated and true values of a parameter, the standard deviation across estimated values, and the average squared deviance between the estimated and true values–in other words, by computing the *bias*, *standard error (SE)*, and *mean squared error (MSE)*, respectively, of the estimator. Typically, *bias* is taken as a measure of systematic error, *SE* measures random error, and *MSE* is a measure of total error. More specific details on simulation evaluation criteria will be provided later in the section discussing simulation design.

How Is Simulation Used in Educational and Psychological Measurement?

To understand how simulations are used in educational and psychological measurement, we surveyed articles from two of the field's most read journals that commonly publish simulation research, *Applied Psychological Measurement* and the *Journal of Educational Measurement*. Twenty percent of articles published between 2008 and 2012 were randomly sampled from each journal. Within this sample, articles utilizing any form of simulation were selected for inclusion; a total of 27 articles representing 60% (27/45) of those sampled used simulation, of which 41% (11/27) included a real data example. Articles that did not include simulation were mostly editorials, book reviews, or discussions of measurement software. The proportion of articles that use simulation demonstrates the importance of simulation to the field.

Articles that included simulation were coded by what was simulated, the number of replications, and how the results were communicated. These results are summarized in Table 1 and show that a majority (67%) simulated IRT parameters based on (i) assumed distributions, (ii) observed distributions estimated from real data, or (iii) a sampling of empirical parameters that were treated as true parameters. The distributions used to simulate item parameters, particularly *a* parameters, varied across studies. Some studies generated *a* parameters and *c* parameters (if used) through a log normal distribution while others used either a uniform or normal distribution with fixed minimum and maximum values. Difficulty parameters were most often simulated as normal. Additionally, it was common for studies to sample each parameter independently; however, because item parameters tend to be correlated in practice, using a multivariate distribution to account for the covariance between parameters (as was done in Baldwin, 2011, for example) can improve generalizability in some cases.

Regarding the types of research that utilized simulation, it was not surprising to see IRT model fit at the top of the

Table 1. Summary of Survey on Simulation Articles

Survey Indices	*n*	Percent of Articles
Types of Data Simulated		
IRT Parameters	18	67
Raw/True Scores	5	19
Response Times	2	7
Other	3	11
Type of Research Utilizing Simulation		
IRT Model Fit	6	22
Equating/Linking	4	15
CAT	3	11
Cognitive Diagnostic Modeling/Q Matrix	2	7
Cheating	2	7
Differential Item Functioning (DIF)	2	7
Dimensionality	2	7
Reliability	2	7
Other	4	15
Number of Replications		
499 or Fewer	9	33
500–999	4	15
1,000–4,999	5	19
5,000–10,000	3	11
None	6	22
How Results Were Analyzed/Presented		
Bias/RMSE	14	52
Type I/Power	10	37
ANOVA	2	7
Model Fit	2	7
Other	4	15

list. The "Other" category comprised an assortment of studies including missing data, generalizability theory, and regression errors. Thus, simulation is a popular technique across a broad range of topics in the field of educational measurement.

Most studies used fewer than 500 replications, though that does not necessarily suggest that the number used was sufficient in all cases. Six of the 27 articles (22%) simulated a large amount of examinee data, but did not repeat the process across replications. Typically, authors did not provide justification for the number of replications they elected to run, and those who did often cited reasons other than the sampling properties across replications of whatever they were investigating. For example, Wang and Jin (2010) explained that "[o]nly 30 replications were completed because each replication required more than 24 hrs of computer time" (p. 54). Using those time estimates, it would have taken a month to run the 30 replications or 1 day using 30 computers. More generally, this suggests that resource limitations may impose an upper bound on the number of possible replications. It is therefore essential that researchers determine how many replications are needed to support the intended inferences and or claims during the study design phase. If the required number is too great, it may be better to change the study design rather than risk reporting chance results.

With respect to presenting simulation results, approximately half the studies surveyed (52%) compared bias and root mean square error (RMSE) to assess how well simulated data results matched with known values. Similarly, studies often used Type I error/power rates as the criterion in determining whether or not the simulation detected a true effect. A less common approach was ANOVA, which, although not applicable in all situations, allows the researcher to determine the amount of variance explained by each of the manipulated simulation factors and thus their relative impact and interactions.

Principles and Recommendations in Design, Analysis, and Communication

A researcher conducting a simulation study should expect reviewers to question the rationale for using simulated data, the logic behind the design choices and analysis methods, and especially the generalizability of the findings. It can be a big leap to generalize from simulated data to real-world consequences. Results from simulation research are limited in the same way that results from real data are limited based on the particular study from which they were derived (e.g., how the sample was selected, quality of the data, study design, analysis methods). Thus, the best way to ensure that results from simulated data have value is to make thoughtful choices about the alignment between the study design and purpose for the research.

Harwell et al. (1996) previously outlined the basic steps taken in an IRT simulation study: (i) specifying a research question, (ii) delineating conditions, (iii) choosing an experimental design, (iv) generating data, (v) estimating parameters, (vi) comparing true and estimated parameters, (vii) replicating the procedure a specified number of times, and (viii) analyzing results using an appropriate method based on the design. Here, we offer suggestions for many of these steps as well as how to summarize and present results.

Specifying how the above steps will be executed helps to clarify how the research questions can be answered using simulation. Additionally, writing out the protocol beforehand makes the methods section of the final paper easier to write and easier to understand for readers. As with other kinds of scientific studies, enough detail should be provided so that another competent researcher could replicate the findings. Principles of good experimental design carry over to simulation studies too, including clearly stating the factors that are varied and the values for each condition. Typically, conditions are selected that are both realistic and meaningful.[1] For instance, comparing between sample size conditions of $n = 1$ and $n = 1{,}000{,}000$ would show a large effect, but in many cases neither value would be considered realistic or meaningful.

Still, the core of simulation is, well, simulating data. This aspect of the study requires clear explanation for readers, including the type and version of software used. This information can be important for readers who wish to replicate aspects of a study because in some cases changes in software can affect formatting of output or possibly the calculation of certain quantities. Despite this, the type of software used let alone the version was rarely included in the simulation articles surveyed. For some large or computation-intensive simulations, providing information about the computer(s) (e.g., operating system, processing speed, RAM) lets readers know the resources needed to replicate or extend the study. Lastly, describing the methods to verify simulation code is critical given that an erroneous keystroke or two in writing simulation code can render the results worthless. Assuming the results look reasonable, one way to ensure that the data were generated correctly would be to compare against an independent set of code written by another researcher or

retrieving known results. Perhaps more feasible, simply ask a colleague to review a set of code for completeness and quality, and then publish the code along with the simulation study to allow interested readers the opportunity to examine the code.

Researchers should also keep a record of the random number of seeds set for each condition to provide upon request to other researchers interested in replicating their findings. The seed is an integer that sets the starting point for generating random numbers, similar to an experimenter using a random number table to assign participants to conditions. Specifying the seed before generating data allows the simulation results to be reproduced, as if another experimenter had access to the same participants and used the same random number table to assign them to each condition. However, knowing the seed does not guarantee that someone else can reproduce the results—it could depend on the particular machine or some other quasi-random process (for instance, the infamous RANDU random number generator that was not random; Marsaglia, 1968). Researchers have argued that the quality of simulation is directly tied to the quality of the random number generator (Brooks, Barcikowski, & Robey, 1999; Harwell et al., 1996), though this may not apply to modern random number generators. Additionally, others have suggested even reporting the seeds for each condition (Burton, Altman, Royston, & Holder, 2006), though this is probably an unnecessary use of journal space as researchers should be able to produce similar results under the same simulation conditions.

As described, when parameters are sampled from distributions, the distributions from which they are sampled can vary both within and across studies. Most often, proficiency is sampled from a normal distribution unless skewness is a factor the investigators wish to vary. In the context of the unidimensional 3PL IRT model, item discrimination parameters are often simulated as log-normal (the default prior from BILOG-MG; Zimowski, Muraki, Mislevy, & Bock, 2003); difficulty parameters are often sampled from normal distributions; lower asymptote parameters are seen to be sampled from beta, log-normal, logit-normal, uniform, or normal distributions with minimum and maximum values specified. Additionally, when item difficulty parameters are normal, discrimination parameters are log-normal, and lower asymptote parameters are logit-normal, it makes it easy to draw them from a multivariate normal distribution. Interested readers are directed to Mooney (1997) for a detailed description and comparison of various probability distributions from which parameters can be sampled. To make simulations more realistic, distributions are often specified based on the analysis of an empirical data set. Other times, researchers sample from distributions they believe to be realistic or vary the distributions' characteristics as a study factor to explore how these characteristics affect various outcomes of interest. Within each condition, a design choice has to be consciously made as to whether the same or independent data sets will be used for each outcome measure under study. If testing multiple outcomes on the same data set generated within a condition, an appropriate paired (repeated-measures) analysis technique should be used to control for Type I error. Of course, independent data sets can be generated for each outcome per condition, understanding that differences may arise due to sampling variability, particularly for smaller numbers of replications.

After selecting the data generating distributions, if relevant, one needs to consider the number of replications to include. This is one of the most important decisions to be made, as simulations are based on sampling theory and thus require purposeful thought as to the number of replications to ensure the demonstration has appropriate statistical properties. Simulations "are really statistical sampling experiments with an underlying model whose results are used to address research questions" (Harwell et al., 1996, p. 102). Additionally, replications provide an estimate of the stability of the simulation itself. Since the simulated data are dependent on comparisons to a random variable, the same code will produce different results if a random number seed is not used. Thus, the results of interest are influenced by the variability introduced by the manipulation performed by the researcher (say, for example, the type of estimator used) along with sampling variability.

More replications are always better in terms of producing a more accurate and reliable estimate of the parameters of interest. However, the question here is how much is enough? With too few replications, the desired analysis may not be robust enough. With too many replications (though of lesser concern), the simulation becomes an inefficient use of computer resources without adding much value to the analysis. Ultimately, the number of replications needed depends on the research questions being asked.

Harwell et al. (1996) suggested a minimum of 25 replications for IRT studies, but computing efficiency has improved a great deal since that recommendation and it would be difficult to justify due to the range of studies to which simulation techniques are currently applied in the field. Instead, one must make a realistic determination based on the practical constraints of time per replication (which is based on model complexity and number of factors coupled with hardware and software constraints) compared to an ideal number of samples. It is good practice to run code with a few replications, time how long it will take, and then extrapolate running time based off of the per-replication estimate.

One way to figure out an appropriate minimum number of replications is to consider the *SE* of the mean[2] (σ_M) for the parameters being estimated:

$$\sigma_M = \frac{\hat{\sigma}}{\sqrt{n-1}}, \tag{1}$$

where $\hat{\sigma}$ is the standard deviation of the estimated parameter across replications and n is the number of replications. In simulation, the mean across replications is similar, conceptually, to a sample mean across n participants and is subject to sampling error. Applying sampling theory, the distribution of sample means should approximate a normal distribution with σ_M as the standard deviation. This quantity allows us to specify a confidence interval around the mean based on the number of replications. For instance, say the standard deviation for an estimated parameter is .35 and you want to be 68% confident that any estimate of the mean is within ± 1 decimal point of the value that would be obtained if the entire population of replications were used. This means that the approximate spread corresponding to σ_M should be $\leq .1$, and thus $\sigma_M \leq .05$. Substituting the criterion of .05 into (1) we get

$$n = \left(\frac{\hat{\sigma}}{\sigma_M}\right)^2 + 1 = \left(\frac{0.35}{0.05}\right)^2 + 1 = 7^2 + 1 = 50.$$

Thus, a minimum of 50 replications would be enough to have sufficient precision out to the first decimal. If precision out

to the second place was desired ($\sigma_M = .005$), the required number of replications spikes to 4,901. Thus, the number of replications is a function of the desired precision and variation between replications for the estimated parameter.

When analyzing results, a determination must be made on how the estimated and true parameters are to be compared. Researchers often use measures such as correlation, coverage, bias (Equation 2), mean absolute difference (MAD, Equation 3), SE (Equation 4), mean square error (MSE; Equation 5), and RMSE (Equation 6).

$$\text{Bias} : \frac{\sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_{True}\right)}{n}, \tag{2}$$

$$MAD : \frac{\sum_{i=1}^{n} \left|\hat{\theta}_i - \theta_{True}\right|}{n-1}, \tag{3}$$

$$SE : \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{\theta}_i - \bar{\theta}\right)^2}{n-1}}, \tag{4}$$

$$MSE : \frac{\sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_{True}\right)^2}{n-1}, \tag{5}$$

$$\text{RMSE} : \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_{True}\right)^2}{n-1}}. \tag{6}$$

Above, n refers to the total number of replications used to estimate a parameter, θ_{True} refers to the true parameter, $\hat{\theta}_i$ refers to the estimated parameter for replication i, and $\bar{\theta}$ refers to the mean estimated parameter. Note that in this context the Greek letter theta is used to represent any parameter of interest (though within psychometrics, theta is conventionally used for a specific type of parameter). Also, given that the statistics in (3)–(6) are measures of dispersion, the denominator is $n - 1$ rather than n.

Bias provides a measure of the average distance between the estimated and true parameter. Positive bias occurs when the estimated parameter is greater than the true parameter and negative bias occurs when the estimated parameter is less than the true parameter. An unbiased estimator will have both positive and negative deviations from the true parameter value; however, these deviations across samples will average to zero. *SE* is the standard deviation of the estimated parameter across replications, a measure of random error. Accuracy of *SE* can be assessed by investigating the confidence interval coverage, or the number of replications in which a 95% confidence interval (2 *SE*s) around the estimated parameter contains the true parameter. The *SE* would be considered accurate if the confidence interval coverage rate was approximately 95%. However, if the estimated parameters were biased, then coverage would be less than 95% even if the *SE*s were accurate. MSE measures the average of the squares of the deviations between the estimated and true parameters and serves as an overall measure of error that combines both bias and variability: $MSE = bias^2 + SE^2$. For an unbiased estimator, bias is equal to 0 and MSE is the variance of the estimated parameter. RMSE is the square root of MSE, and thus for an unbiased estimator, RMSE is equal to SE, the standard deviation of the estimated parameter. A good check on RMSE is calculating MAD, which provides an absolute magnitude of the average distance between the estimated and true parameter. MAD includes both systematic and random error, similar to RMSE, and is much less sensitive to extreme sample values.

Selecting which statistics to compare, and how to compare them, depends on the research question, though bias, *SE*, and RMSE are the most common. It is also important to consider how simulations were evaluated in previous research, such that similar dependent variables can be applied to facilitate comparisons and allow results from your simulation to build upon previous research. Other comparisons can be picked based on the desire to see the amount of error in different formats, along with the direction of the error. However, when calculating error, it is important to ensure the parameters being compared are on the same metric by equating back to the true parameter metric, if necessary (Baker, 1990; Kolen & Brennan, 2004).

Power and Type I error rates are also commonly used to evaluate simulation results. In this context, a researcher may be interested in calculating the proportion of correctly rejected null hypotheses given some nominal level, or the proportion of hypothesis tests below a nominal level when the null is actually true. Additionally, in such a situation, it is important to control the nominal Type I error rates; otherwise, comparisons of power to detect an effect are confounded.

Communicating simulation results is just as important as generating the data from which they were derived because a poor table or graph will diminish the readers' ability to effectively draw the intended inferences. Many simulations report measures of error between various experimental conditions (e.g., bias, *SE*) in a table. However, this can make it difficult to infer differences between cells when the number of crossed conditions combined with the number of replications yields an unwieldy amount of summary data (e.g., see table 2 in Choi, Kim, Chen, & Dannels, 2011). In the context of helping the reader to interpret the results, a table should be designed for communication not data storage. Additionally, general rules for improving a table are applicable, such as (i) ordering rows and columns in a meaningful way, (ii) rounding unnecessary decimals, and (iii) including summary statistics across rows and columns (Wainer, 2000). Additionally, if a simulation has the form of a factorially designed experiment, it can be summarized and presented as you would in any such experiment. Thus, an inferential technique, such as ANOVA, may help test for whether and where any noteworthy differences between conditions exist. See table 2 in Sinharay (2010) and table 3 in Livingston and Kim (2009) for examples of simulation results presented in table form.

Graphical forms are almost always more interesting than a table display and can be a powerful way to draw attention to certain characteristics of a dependent variable. Common graphical displays in presenting simulation results are boxplots, histograms, density curves, scatterplots, and line graphs. A boxplot displays several key features of a distribution, including the quartiles and outliers and is particularly useful in highlighting differences between distributions for two or more groups of data. A histogram provides the count or percentage of individuals within equally spaced intervals. Thus, boxplots emphasize the center and spread of a distribution whereas histograms emphasize the distribution of values, making it easier to identify particular points with high frequency. A density curve is just a smoothed line version of

a histogram, in which the y-axis is not the frequency, but rather the proportion of the group within a range of values. Both scatterplots and line graphs highlight changes in values across the x-axis. Scatterplots provide the actual data points whereas a line graph represents the average across values and is particularly useful in simulation when comparing across multiple conditions. See figures in van der Linden and Wiberg (2010) and DeMars (2006) for examples of simulation results presented in figures. Also, for general guidance on how to present results, see *Visual Revelations* (Wainer, 2000), and for graphing in R, see *Graphical Data Analysis With R* (Unwin, 2015).

Illustration of Simulation Implementation

Any number of software programs or languages can be and are used to run a simulation study. Some researchers prefer computer programming languages such as C++, Java, or Fortran due to their speed and flexibility. R is an increasingly popular choice; R (R Core Team, 2015) is a free open-source statistics-specific programming language built on the S programming language (Becker, Chambers, & Wilks, 1988), with many free resources and code examples available online and strong graphics and visualization capabilities. For instance, new R users can learn/review basic programming from a free guide (http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf). Others use statistical packages such as SAS (SAS Institute, 2014), SPSS (IBM Corporation, 2013), or Stata (StataCorp, 2015). More specific psychometric software packages, such as flexMIRT (Cai, 2013), are able to simulate data as well. Additionally, Rick Wicklin (2013) has an entire book devoted to simulations in SAS; another text is devoted to R (Jones, Maillardet, & Robinson, 2009). Furthermore, resources have been written for specific approaches such as factor analysis (Tucker, Koopman, & Linn, 1969) and structural equation modeling (Bandalos, 2006; Muthén & Muthén, 2002), among others.

Often, a simulation is conducted by integrating more than one program. There can exist distinctions between programs used to generate data, to estimate models, and to facilitate the overall study. For instance, SAS uses the x command to cause Windows to execute a command, which can be used to call all three phases of BILOG-MG (Zimowski et al., 2003) to run a command file. Another approach would be to write a batch file that can issue a sequence of commands to execute a program (Gagné, Furlow, & Ross, 2009). Thus, multiple programs may need to work in concert to successfully implement a complete study. Such a set-up may be desirable when testing out a new software program, or when a statistical package does not have the ability to generate its own data. An advantage of R is that there are freely available, powerful functions from R packages that can easily be loaded into R to avoid involving other programs, which could help simplify and improve the efficiency of the simulation in some cases.

Some of the more common R packages with psychometric application are eRm (Mair, Hatzinger, Maier, & Rusch 2015), irtoys (Partchev, 2014), and ltm (Rizopoulos, 2006) for unidimensional IRT modeling; mirt (Chalmers, 2015) for multidimensional IRT modeling; mvtnorm (Genz et al., 2014) for generating multivariate data; plink (Weeks, 2010) and equate (Albano, 2014) for linking and equating; ggplot2 (Wickham, 2009) for producing graphs; psych (Revelle, 2015) for multivariate analyses, exploratory factor analysis, and scale construction; psychometric (Fletcher, 2010) for reliability and classical item analysis statistics; and lavaan (Rosseel et al., 2015) for confirmatory factor analysis, structural equation modeling, and growth curve modeling.

What follows is an illustrative scenario using the basic steps outlined earlier by Harwell et al. (1996) for conducting an IRT simulation. The accompanying R code is included in the Supporting Information and all steps are annotated to aid the reader in following how the code is working. Replication processing is accomplished by using loops: it would be extremely time-consuming and error-prone to run the simulation several times by updating the number of items in each condition or running each replication sequentially. Instead, the code is looped to allow the process to repeat the desired number of times. Once looping/macro processing is understood, it becomes clear that the code can be generalized to additional scenarios by varying conditions such as the number of simulees, the distributions used to sample ability and item parameters, or the IRT model estimating the data.

Realize that, as in any programming language, there are numerous ways to accomplish a single task. The code we provide may not be the most efficient, and is certainly not the only way to carry out the simulation. Run time for the code took approximately 2.5 hours based on R version 3.2.2, a computer with 2.60 GHz of processing speed, and 4 GB of RAM. After installing R packages irtoys and reshape, and updating the directory to which the output is saved, the sample code can be pasted directly into R to reproduce the simulation and summary results. In addition, the simulation can be easily adapted by updating the simulation conditions at the beginning of the code.

Illustrative Scenario

Testing Company X annually administers a high-stakes 300-item certification test to approximately 1,500 examinees. After the most recent administration, posttest survey feedback revealed that examinees were feeling rushed while testing, which suggests they may not be performing to the best of their ability. No performance evidence indicates that the test is speeded; nevertheless, the governance board overseeing the test mandates that examinees are provided more time to complete the test. Given that Testing Company X is unwilling to absorb costs associated with additional testing center seat time and the governance board is unwilling to pass along costs to examinees, you, the Psychometrician, suggest maintaining the same allotted total testing time, but reduce the number of items to allow more time per item. The governance board is amenable to this solution and requests that you provide a recommendation for a reduced test length.

Research Question(s)

To inform this recommendation, it would be worthwhile to explore the impact of reducing the tests' length on score precision, particularly around the passing standard. Aside from the primary pass/fail inference of the test, scores at various other points are of interest, such as at the low end to identify at risk examinees or at the high end to award honors/distinctions. Thus, there are two research questions:

1. What is the impact on classification accuracy as the number of items is reduced?
2. How does reducing the number of items affect score precision across the score scale?

Delineating Conditions

The simulation will manipulate one factor, test length. One way to create different test length conditions would be to reduce the current length of the test in evenly spaced intervals of 10 items, say from 300 to 200. However, a more realistic approach would be to consider how the exam is built and presented to examinees. For this scenario, though the test is composed of 300 items, they are administered in 6 blocks of 50 items, with 60 minutes provided to complete each block. Thus, examinees are managing their time at the block level and it would make sense to consider test length conditions that maintain a similar number of items per block. Additionally, the total allotted amount of time to complete each block should be a reasonable number to communicate to examinees (e.g., time in full minute increments). Given the current allotted time per item (72 seconds; 60 minutes/50 items), decreasing a block by 5, 10, and 15 items would free up approximately 6, 12, and 18 minutes of testing time, respectively, per block. Thus, the test length conditions explored by the simulation will be 300, 270, 240, and 210 items.

Experimental Design

To facilitate generalizability of the results, it is important that the data produced by the simulation align with realistic conditions of the actual testing program. Currently, the test is completed by approximately 1,500 examinees so the sample size for the simulation is 1,500. Operationally, examinees are scored using the Rasch (1960) model:

$$P\left(U_{ij} = 1 \left| \theta_j, b_i \right.\right) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta - b_i)}},$$

where the probability of a correct response to item i for examinee j is a function of item difficulty b_i, and examinee ability θ_j. Thus, the simulation will score examinees using the same Rasch model.

Given that both research questions are interested in modeling the real data, ability and item difficulty parameters should be sampled from probabilistic distributions that reflect the empirical distributions, not just standard normal distributions of $N \sim (0, 1)$. Figure 1 shows the distributions of estimated examinee ability and item difficulty parameters from the most recent administration. The examinee ability distribution indicates that a high proportion of examinees score high on the test, with an approximate pass rate of 83%. Additionally, there exists a divergence in the means of the item difficulty and examinee ability distributions, which can be common in classification tests where items are partially selected based on their potential to maximize information and reduce measurement error at or near the passing score. To account for the shape and relative location of the ability and item difficulty parameters in the real data, both distributions are fit with a normal curve to be used to sample true parameters for the simulation.

Alternatively, if this was not a didactic exercise and readers could access the real data, we could have instead sampled the empirical parameters and treated them as true for the purposes of simulation. Assuming there exist a sufficient number of examinees and items in the real data to approximate the population parameter distributions, sampling the empirical parameters would better capture the subtleties of the real data and increase generalizability of the results.

Generating Data

Examinee ability parameters are first sampled from $N \sim (1.6, .5)$ and stored as an object to be reused for each replication. Item difficulty parameters, however, are resampled from $N \sim (.4, .8)$ within each replication. This process of generating data ensures that we will have a certain number of replications to stochastically model random fluctuation in the estimated ability distribution that would occur in practice if the same examinees were exposed to different parallel forms of the same test. If the sample were much smaller (e.g., 20 examinees instead of 1,500), then we would need to also resample examinee ability parameters across replications to ensure sufficient coverage of the ability distribution.

Using these sampled parameters, correct/incorrect responses are computed based on the Rasch generating model; the ability of each simulee and the difficulty of each item are inputted into the model to calculate the probability of a correct response, which is then compared to a random draw from a uniform distribution between 0 and 1. Correct responses are assigned when a simulee's computed probability is higher than this draw; otherwise, the response is coded as incorrect. The 0/1 response vectors are then restructured and formatted for estimating IRT parameters.

Estimating Parameters

For each test length condition ($k = 300$, 270, 240, and 210), examinee ability parameters are estimated through maximum likelihood in the R ltm package (Rizopoulos, 2006) associated with irtoys. Item difficulties are fixed to their true parameter values since we are only interested in estimating ability and this approach eliminates the need to equate the estimated ability parameters back to the original scale of the true ability parameters (i.e., each calibration would force the mean of the estimated ability parameters to have the same mean of 0). After ability parameters are estimated, they are then merged back with the original set of true ability parameters that was used to generate the item response data and saved while repeating the data generation and parameter estimation process for the next replication. Saving each replication frees up RAM to improve efficiency as well as protecting against unforeseen circumstances that could interrupt the simulation such as a computer crash, power outage, or a cat walking across the keyboard!

Comparing True and Estimated Parameters

After replicating the data generation and parameter estimation process, all outcome files are read back in and merged into a single object containing all simulated data. To ensure a meaningful comparison between true and estimated parameters, we need to first convert all ability parameters to the reported score scale ($M = 50$, $SD = 10$). For instance, if differences are observed between a particular pairing of estimated and true ability parameters, but both round to the same scale score, then those observed difference are of no practical value. Thus, the threshold for a meaningful difference is $\geq .5$ scale score points.

To address the first research question, changes in pass/fail classification accuracy are investigated by comparing true and estimated scale scores against the passing standard of 41 scale score points and each simulee is categorized as a true passer (TP), true failer (TF), false passer (FP), or false

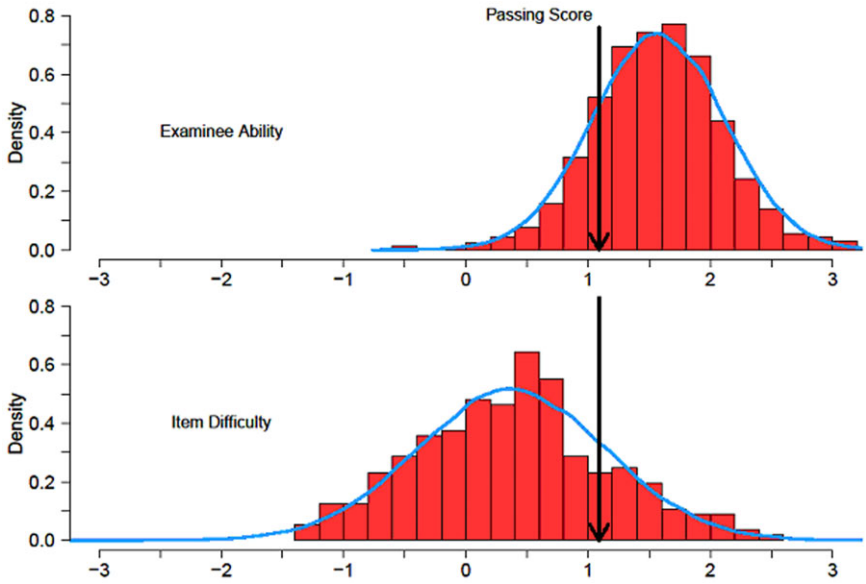


FIGURE 1. Empirical distributions of examinee ability and item difficulty parameters.

Table 2. Misclassification Results

	False Passer			False Failer		
Test Length	***M***	***SD***	***Rate***	***M***	***SD***	***Rate***
$k = 300$	29	5	1.9%	45	6	3.0%
$k = 270$	31	5	2.1%	48	6	3.2%
$k = 240$	33	5	2.2%	51	7	3.4%
$k = 210$	34	5	2.3%	55	7	3.7%

Note: *M* represents the expected number of errors across replications and the corresponding *Rate* was determined by dividing *M* by the total sample size ($N = 1{,}500$).

Table 3. ANOVA Results with Bias, SE, and RMSE as the Dependent Variables

Effect	***df***	***Mean Square***	***F***	η_p^2
Bias	3	.38	3	< .00
Error	5996	.12		
SE	3	90.44	596	.23
Error	5996	.15		
RMSE	3	90.59	589	.23
Error	5996	.15		

failer (FF). A simulee is a TP if both their true and estimated scale scores are greater than or equal to the passing score. A simulee is a TF if both their true and estimated scale scores are less than the passing score. A simulee is a FP if their true scale score is less than the passing score, but their estimated scale score is greater than or equal to the passing score. A simulee is a FF if their true scale score is greater than or equal to the passing score, but their estimated scale score is less than the passing score. The total number of simulees under each classification type is then summed within test length condition and averaged across replications.

To address the second research question, bias, *SE*, and RMSE are compared by averaging across replications for each simulees' true scale score and for each test length condition. The only manipulated factor is the number of items, thus it is expected that a majority of the error will be additional variability of the estimated parameters, *SE*. However, maximum likelihood estimation is known to be outwardly biased at the extremes. Thus, it would be useful to check if the relative locations of the ability and item difficulty distributions contribute to additional bias for the more reduced test lengths that have proportionally fewer items targeted at the upper and lower tails of the ability distribution.

Number of Replications

The data generation and parameter estimation process is looped for each condition and replicated 250 times. The number of replications was determined by running a small number of replications to estimate the standard deviation of the estimated parameters and then extrapolating based on an appropriate level of σ_M. For both research questions, we want to be 68% confident that the average estimated scale score and number of examinees misclassified is within 1 point of the value that would be obtained if infinite replications were run. After running 25 replications, the largest $\hat{\sigma}$ across quantities was 7. Substituting this $\hat{\sigma}$ and the criterion of $\sigma_M = .5$ into Equation 1, we get $n = 197$. Given that this is a minimum recommendation and to account for some imprecision in the $\hat{\sigma}$ estimate, we rounded up and ran 250 replications.

Analyzing Results

Table 2 illustrates how the misclassification results can be presented, where *M* represents the expected number of errors across replications and the rates were determined by dividing *M* by the total sample size ($N = 1{,}500$). The main inference from the table is that misclassification errors increase as the test is shortened, more so for FFs than FPs. Providing the expected number of examinees misclassified helps communicate the practical implications for reducing the test length (e.g., how concerning is passing two additional examinees whose true ability is less than the minimum standard?). The false fail rate is slightly higher compared to that of FPs, though not surprising given that there are more

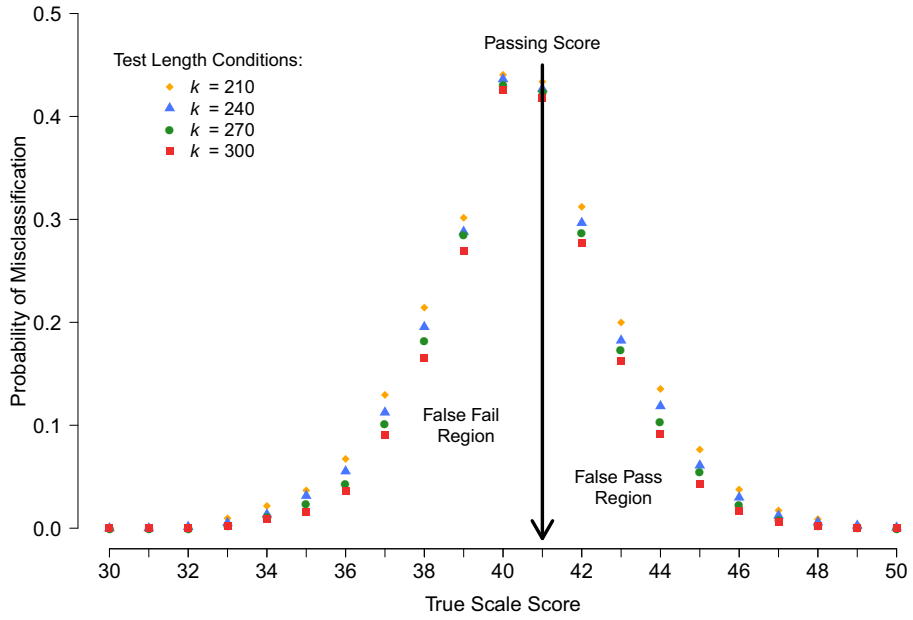


FIGURE 2. Probability of misclassification across the score scale for each test length condition.

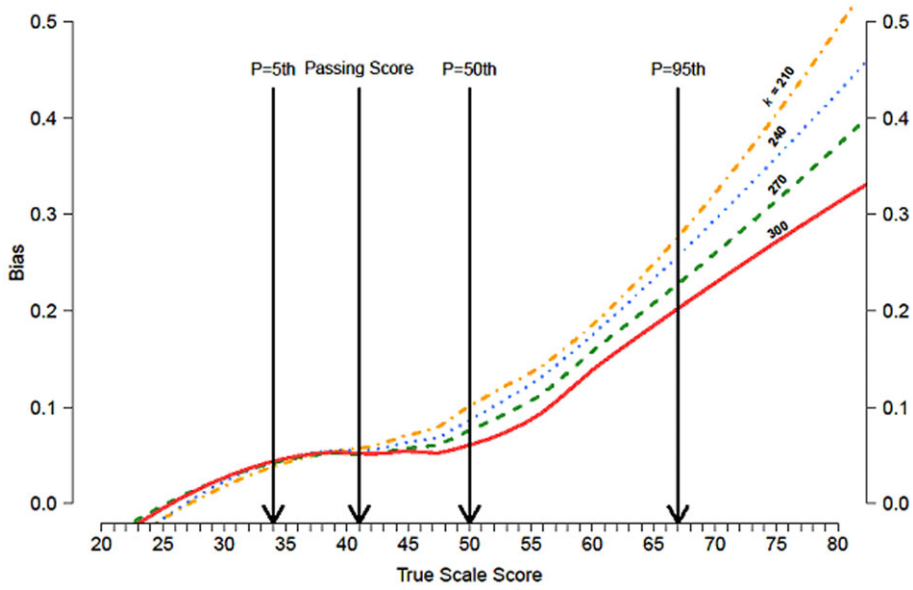


FIGURE 3. Bias results across the score scale for each test length condition.

examinees in the empirical distribution whose true ability is slightly above the passing standard and thus a greater likelihood for a false fail than a false pass. Additionally, the standard deviation for the expected number of misclassifications increases as the test is reduced, which is better illustrated in Figure 2. With the decreasing score precision associated with reducing the tests length, uncertainty around the passing score increases and, accordingly, the misclassification probability increases for an examinee whose true score is near the passing standard. Thus, the added imprecision as it relates to misclassifications is mainly contributed to by examinees whose true score is already around the passing standard. Both Table 2 and Figure 2 communicate similar information; the more detailed focus of the table highlights the practical consequences for each test length condition whereas the plot indicates where the loss in classification accuracy occurs on the score scale.

Instead of summarizing the bias, *SE*, and RMSE results in a dense table of numbers, Table 3 illustrates how the simulation results can be summarized as ANOVA's to highlight the main effects. Table 3 indicates that bias has relatively no overall effect while *SE* and RMSE demonstrate medium to large effects. Figure 3 further explores the bias results and lends support to the outward bias in the maximum likelihood estimation. However, as suggested by the ANOVA result, differences between test length conditions are negligible, particularly within the range of the score scale containing a majority of simulees. More importantly, regardless of condition or placement on the score scale, bias was less than .5 so any differences would always round to the same scale score point.

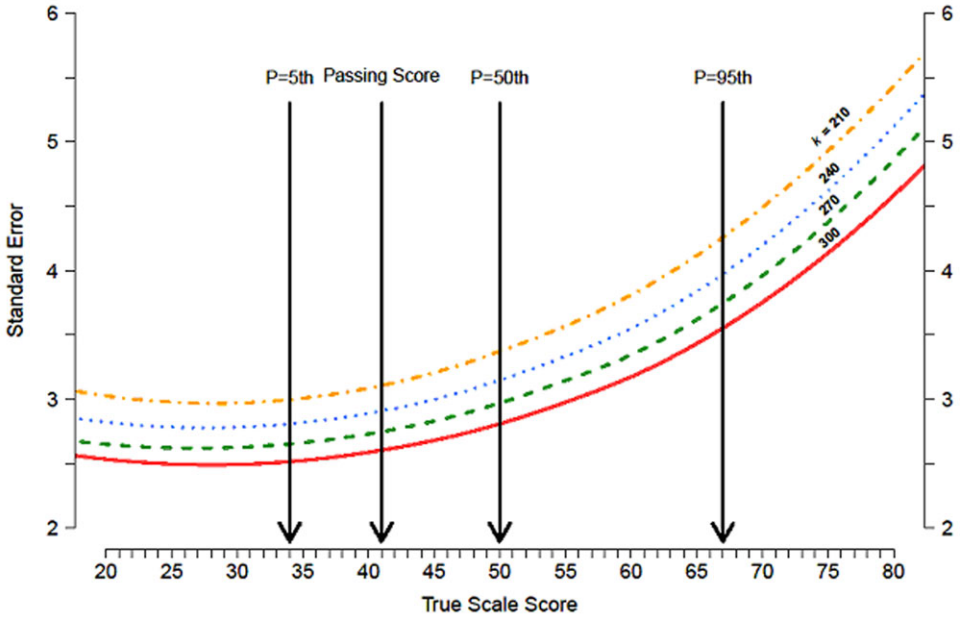


FIGURE 4. Standard error results across the score scale for each test length condition.

Given the bias results and that $RMSE = \sqrt{bias^2 + SE^2}$, it was unnecessary to plot RMSE. Figure 4 plots *SE* across the score scale. Differences between test lengths of 300 and 270 all appear to be less than .5 scale score points. Yet, differences between the current and 240 and 210 item test lengths would likely lead to additional variability in an estimated scale score. This result aligns with Table 2, which shows slightly larger standard deviations in expected misclassification counts for the 240 and 210 item test length conditions. Additionally, both bias and SE are at a minimum slightly below the passing score, which corresponds with Figure 1, which shows that the empirical item difficulty distribution was targeted slightly below the passing score—a useful image to present to those overseeing how the test is built in order to communicate the importance of properly aligning the item difficulty distribution to where precision matters most.

The results of the simulation do not necessarily point to one best answer for the research questions, as that depends on what characteristics of the testing program are most important to optimize. On one hand, the results suggest that the test could be reduced to 210 items without a substantial increase in imprecision. However, the more conservative psychometrician may heavily weigh the misclassification results and view any additional FPs as an egregious error. Additionally, an FF is a consequential error as well; the examinee will have a fail on their permanent record, will need to incur expenses involved with retesting, and could have financial implications if employment is contingent upon passing. Thus, interpretation of the results, simulation or real data, depends on the context, implications for those impacted, and how any changes would be implemented.

Summary and Conclusions

Simulation studies are a core part of research in the psychometric literature. However, few accessible resources are available for beginners interested in designing their own simulation study. This article outlines common simulation terms, design decisions, and possible outcome statistics to be employed. Additionally, computer code is discussed and shared so that readers can see how to implement basic building blocks commonly used in simulation. There are also many other uses for simulations not discussed in this ITEMS module, for example, simulating data for the purpose of generating a null distribution to evaluating model–data fit of real data. Thus, simulation is a broad analysis technique useful for individuals conducting psychometric research. It is hoped that the discussion here will also spur good simulation practices in future publications, regardless of the level of expertise of the researchers.

Although simulation has many practical uses, it is not a replacement for empirical study based on real data. The evolution of addressing a research question may begin with simulation, when data are either unavailable or insufficient for research purposes. Then, after some data have been collected, a resampling study with the potential for greater generalizability may be more appropriate where actual response vectors can be sampled with replacement to investigate various conditions. For instance, the illustrative scenario could be conducted as a resampling of real data where vectors of item scores are resampled for each examinee to create tests of different lengths. Lastly, with the advantage of utilizing findings from prior simulation and resampling studies, randomized controlled experiments would be well positioned to answer a specific research question.

Self-Test

1. Under what situations would it be preferable to simulate data instead of using real data?
2. Why are the following important considerations when designing a simulation?
   (a) Articulating the research question
   (b) Setting the random number seed
   (c) Selecting model generating parameters
   (d) Specifying the number of replications

3. Generate 2,000 ability parameters from $N \sim (0,1)$, 30 item parameters with difficulty from $N \sim (0,1)$ and discrimination from $\ln N \sim (0, .5)$, and create 0/1 response vectors from the 2PL model (Birnbaum, 1968):

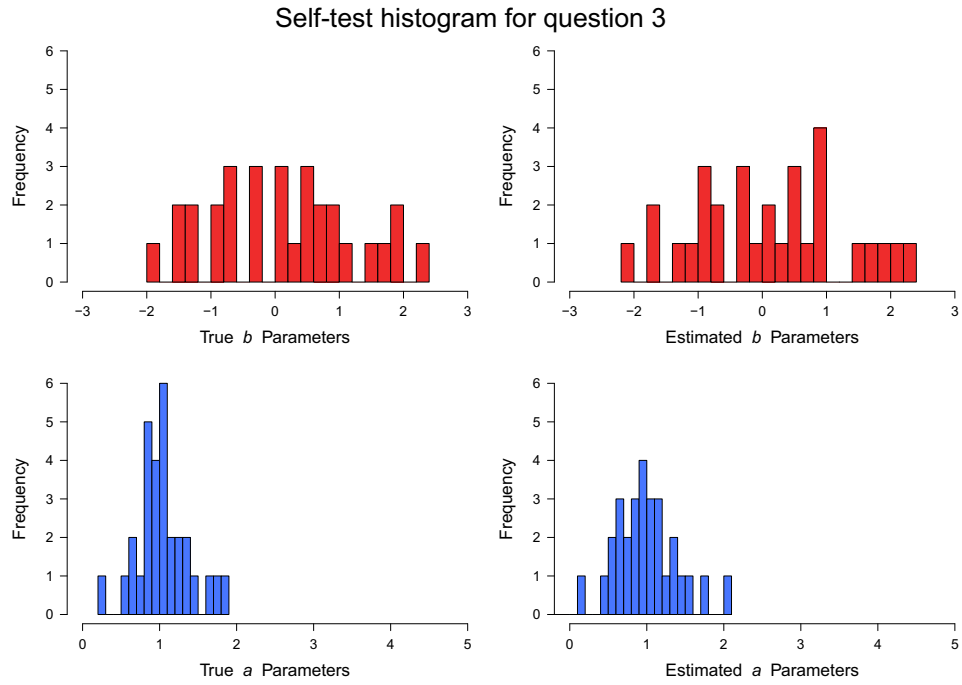
$$P\left(U_{ij} = 1 \left| \theta_j, a_i, b_i \right.\right) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}},$$

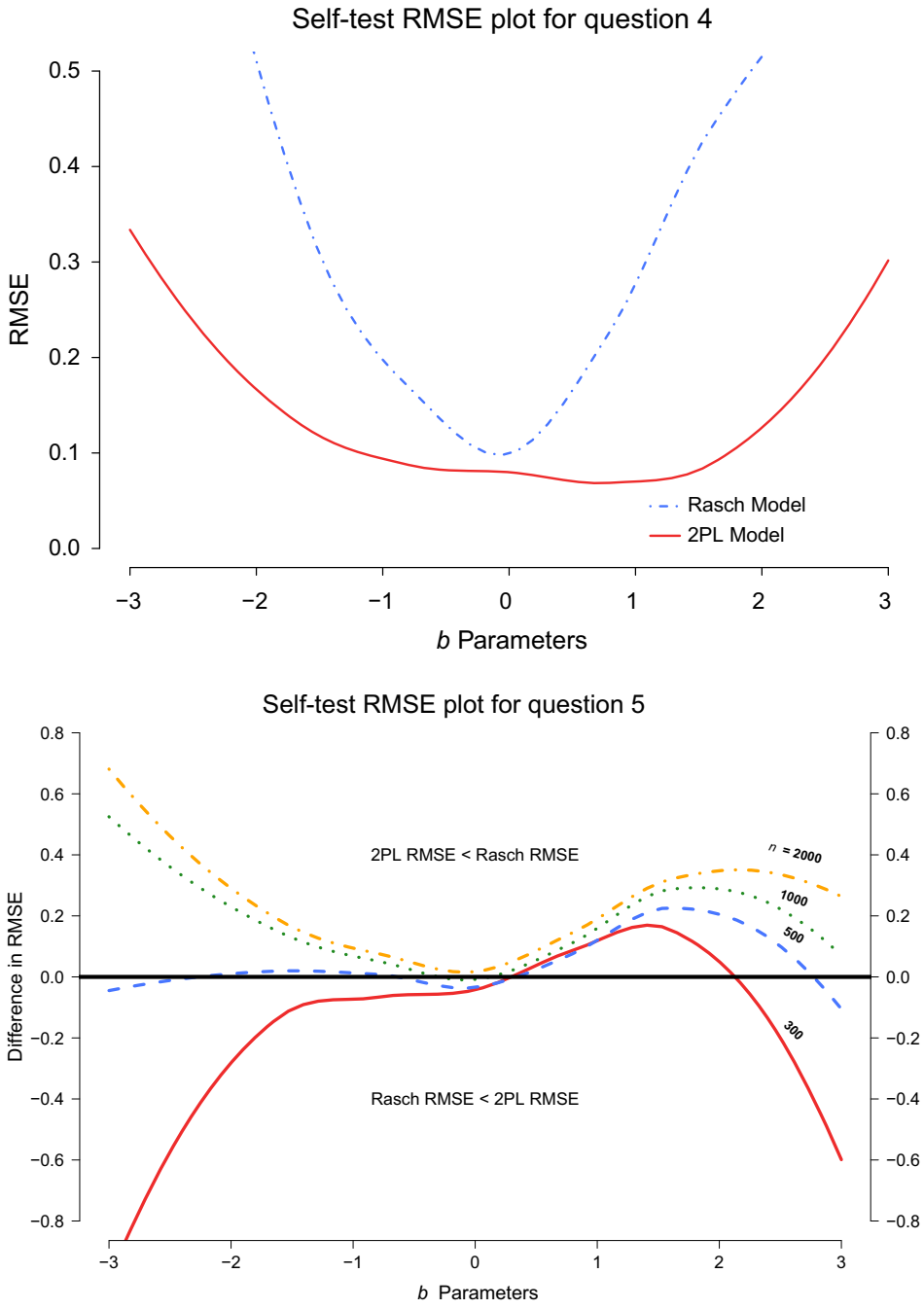
where the probability of a correct response to item i for examinee j is a function of item difficulty b, item discrimination a, and examinee ability θ. Next, using the ltm estimator in irtoys, estimate item difficulty and discrimination parameters. Display histograms to check that the distributions are similar between the estimated and true parameters.

4. Building on #3, compare RMSE in item difficulty estimates between the 2PL and Rasch models based on 2,000 ability parameters from $N \sim (0,1)$. In this situation, we are less concerned about the distribution of difficulty parameters and more interested in investigating differences at various difficulty levels. Generate an interval of item difficulties (–3 to 3 by .2) and pair each with a discrimination parameter sampled from $\ln N \sim (0, .5)$. Replicate 50 times, and plot results.
5. Add a loop to #4 to investigate the effect of sample size for the number of ability parameters (simulees) on the estimation of item difficulty between the 2PL and Rasch models. For sample sizes, $n = 300, 500, 1000$, and 2,000 plot the differences in RMSE results by IRT model for each sample size condition.

Self-Test Answers

1. Simulation studies can be preferable to real data studies in a number of cases, but are perhaps most fruitfully applied when it is useful to know the "true" parameter values when answering the research question or when various factors of interest can be manipulated in a manner not possible using an empirical data set.
2. (a) Articulating the research question: Articulating a clear research question in a simulation study serves the same importance as it does in a real data study: it forces the researcher to ensure that the exact design choices made align with the question being asked.
   (b) Setting the random number seed: Setting the random number seed enables a researcher to exactly replicate a string of random numbers, and thus reproduce the results of a particular analysis. It also allows another researcher to replicate the simulation study.
   (c) Selecting model generating parameters: Simulation studies are only generalizable to the extent that they approach reality. Depending on the purpose of a given simulation study, the results may only provide actionable results to data that look like the generating model used in the simulation.
   (d) Specifying the number of replications: The minimum number of replications in a simulation study takes careful consideration and should be based on the desired precision for an estimated parameter. This must also be weighed against practical time and computer resource constraints.





Acknowledgments

We wish to thank Derek Briggs, Holmes Finch, Peter Baldwin, Howard Wainer, Mike Jodoin, Carol Morrison, Daniel Jurich, Amanda Clauser, and Sandip Sinharay for their wisdom and insightful comments on earlier versions of this ITEMS module.

Notes

[1]In some cases, unrealistic conditions can still be meaningful. For example, finding the lower or upper bound of a given statistic could require including the minimum or maximum *possible* value for a condition even when such a value is not likely to arise in practice.

[2]Also referred to as the "Most Dangerous Equation" (Wainer, 2007) for the dangers associated with interpreting the mean of any estimator in the absence of considering how sample size affects statistical variation.

References

Albano, A. (2014). Equate: Observed-score linking and equating. R Package Version 2.0-3. Retrieved October 15, 2015, from https://cran.r-project.org/web/packages/equate/index.html)

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement*, *14*(2), 139–150.

Baldwin, P. (2011). A strategy for developing a common metric in item response theory when parameter posterior distributions are known. *Journal of Educational Measurement*, *48*(1), 1–11.

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385–426). Greenwich, CT: Information Age.

Becker, R. A., Chambers, J., & Wilks, A. R. (1988). *The New S Language*. London: Chapman & Hall.

Birnbaum, A. (1968). Chapters 17–20 in F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Brooks, G. P., Barcikowski, R. S. & Robey, R. R. (1999). *Monte Carlo simulation for perusal and practice*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulations studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292.

Cai, L. (2013). *flexMIRT version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring [computer program]*. Chapel Hill, NC: Vector Psychometric Group.

Carlson, J. E. (2012). *Issues in comparability of anchor item sets in comparative studies with implications for research and practice*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Chalmers, P. (2015). mirt: Multidimensional Item Response Theory. R Package Version 1.10. Retrieved October 15, 2015, from https://cran.r-project.org/web/packages/mirt/index.html)

Choi, J., Kim, S., Chen, J., & Dannels, S. (2011). A comparison of maximum likelihood and Bayesian estimation for polychoric correlation using Monte Carlo simulation. *Journal of Educational and Behavioral Statistics*, *36*(4), 523–549.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data (ACT Research Report Series 97-4)*. Iowa City, IA: American College Testing Program.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145–168.

Dorans, N. J. (2012). *Simulations are deductive demonstrations not empirical experiments*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Fletcher, T. D. (2010). Psychometric: Applied psychometric theory. R Package Version 2.2. Retrieved October 15, 2015, from http://CRAN.R-project.org/package=psychometric)

Gagné, P., Furlow, C., & Ross, T. (2009). Increasing the number of replications in item response theory simulations. *Educational and Psychological Measurement*, *69*(1), 79–84.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Genz, A., Bretz, F., Tetsuhisa, M., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2014).mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9997. Retrieved October 15, 2015, from http://CRAN.R-project.org/package=mvtnorm)

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125.

IBM Corporation. (2013). *IBM SPSS statistics for windows, Version 22.0*. Armonk, NY: IBM Corp.

Jones, O., Maillardet, R., & Robinson, A. (2009). *Introduction to scientific programming and simulation using R*. Boca Raton, FL: Chapman & Hall/CRC.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Livingston, S. A. (2012). *Keeping it real*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, *46*(3), 330–343.

Mair, P., Hatzinger, R., Maier, M. J., & Rusch, T. (2015). eRm: Extended Rasch Modeling. R package version 0.15-5. Retrieved October 15, 2015, from https://cran.r-project.org/web/packages/eRm/index.html)

Marsaglia, G. (1968). Random numbers fall mainly in the planes. *Proceedings of the National Academy of Sciences*, *61*(1), 25–28.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science, Special Issue*, *15*, 125–130.

Mooney, C. Z. (1997). *Monte Carol simulations*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-116. Thousand Oaks, CA: Sage

Muthén, L. K., & Muthén, B. O. (2002). Teacher's corner: How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620.

National Council on Measurement in Education. (2012). *2012 Annual meeting program*. Madison, WI: NCME.

Partchev, I. (2014). irtoys: Simple interface to the estimation and plotting of IRT models. R Package Version 0.1.7. Retrieved October 15, 2015, from http://CRAN.R-project.org/package=irtoys)

Patz, R., & Junker, B. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Patz, R., & Junker, B. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Revelle, W. (2015). Psych: Procedures for psychological, psychometric, and personality research. R Package Version 1.5.8. Retrieved October 15, 2015, from https://cran.r-project.org/web/packages/psych/index.html)

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and Item Response Theory analyses. *Journal of Statistical Software*, *17* (5), 1–25.

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Barendse, M.(2015). Lavaan: Latent variable analysis. R Package Version 0.5-18. Retrieved October 15, 2015, from https://cran.r-project.org/web/packages/lavaan/index.html)

SAS Institute Inc. (2014). *Version 9.4 of the SAS system for Windows*. Cary, NC: SAS Institute Inc.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150–174.

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement*, *7*(4), 405–425.

StataCorp. (2015). *Stata statistical software: Release 14*. College Station, TX: StataCorp LP.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

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

Unwin, A. (2015) *Graphical data analysis with R*, Boca Raton, FL: Taylor & Francis.

Van der Linden, W., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement*, *34*(8), 620–640.

von Neumann, J., & Ulam, S. J. (1949). Various techniques used in connection with random digits. *Journal of Research of the National Bureau of Standards*, *12*, 36–38.

Wainer, H. (2000). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (2007). The most dangerous equation, *American Scientist*, *95*(3), 249–256.

Walker, M. E. (2012). *Toward more principled simulations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Wang, W., & Jin, K. (2010). Multilevel, two-parameter, and random-weights generalizations of a model with internal restrictions on item difficulty. *Applied Psychological Measurement*, *34*(1), 46–65.
Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*(12), 1–33.
Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer
Wicklin, R. (2013). *Simulating data with SAS*. Cary, NC: SAS Institute Inc.
Wilcox, R. R. (1988). Simulation as a research technique. In J. P. Reeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 134–137). New York, NY: Pergamon.
Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*(4), 1261–1295.
Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 [computer program]*. Chicago, IL: Scientific Software.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix R Code.R
Self-Test Q3 R code.R
Self-Test Q4 R code.R
Self-Test Q5 R code.R