Listening. Learning. Leading.®

Research Report
ETS RR–14-16

Simulate to Understand Models, Not Nature

Neil J. Dorans

December 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Simulate to Understand Models, Not Nature

Neil J. Dorans

Educational Testing Service, Princeton, NJ

Simulations are widely used. Simulations produce numbers that are deductive demonstrations of what a model says will happen. They produce numerical results that are consistent with the premises of the model used to generate the numbers. These simulated numerical results are not empirical data that address aspects of the world that lies outside the model. In contrast, empirical data are central to the scientific method. When a simulation is substituted for the assessment of hypotheses with real data, a false sense of understanding can ensue and with it a biased perspective on the world. To illustrate the limitations of simulation and their proper role, examples are drawn from simulation studies about score equating.

Keywords Simulations; empirical investigations; deductive demonstrations; attribute substitution; score equating

doi:10.1002/ets2.12013

Simulations abound in our field, as can be attested to by perusal of mainstream journals. Simulations produce numbers that are deductive demonstrations of what a model says will happen. They produce numerical results that are consistent with the premises of the model used to generate the numbers. These numerical results are not empirical data that address aspects of the world that lies outside the model. Sometimes, simulations are deductive demonstrations that follow from a model that purports to describe observed data. For example, a simulation may use observed estimates of item parameters as the starting point for simulating test data. Sometimes the predictions are based on a model alone, in which case the demonstrative nature of the simulation is quite evident. For example, a factorial design may be used in which different levels of factors such as item difficulty and discrimination are crossed to produce different combinations of item parameter for data generation. The numbers generated from this factorial design are deductive demonstrations of the generating model. Whether the simulation uses real data as a starting point or is based on varying the values of the parameters of a model, the results of the simulation are consequences of the model. These results are not empirical evaluations of a substantive question.

In the next section, I note the centrality of empirical data to the scientific method. I contrast this data-based evidential approach with an approach to inquiry that substitutes simulated numbers for empirical data. Examples are drawn from the score equating literature to illustrate applications of simulation. I close with the recommendation to simulate to understand models but collect empirical data to understand the natural world.

The Distinction Between Empirical Science and Simulation

The scientific method refers to an approach to inquiry that is characterized by recursive components of observation, formulation of hypotheses, predictions generated from these hypotheses, and empirical evaluation of the predictions about the world. The scientific method requires the collection of empirical data that can be used to test predictions that are based on a model or hypothesis. This method of inquiry has been described in many ways. Here is how a Nobel Prize physicist described it:

> Science is a way to teach how something gets to be known, what is not known, to what extent things are known (for nothing is known absolutely), how to handle doubt and uncertainty, what the rules of evidence are, how to think about things so that judgments can be made, how to distinguish truth from fraud, and from show—Richard Feynman. (Krauss, 2011, p. 1)

Corresponding author: N. Dorans, E-mail: ndorans@ets.org

A critical aspect of the scientific method that distinguishes it from other modes of inquiry is the *collection of empirical data*, where possible via well-designed experiments, to test or revise hypotheses about nature. By nature, I mean the material world that is accessible to us either directly or indirectly via our senses or their extensions, such as calibrated scientific instruments. While test takers may invoke the supernatural for help, test takers and their performance on tests are part of the material world. The proper use of empirical data is essential for distinguishing truth from what might just be a hypothesis with an appealing narrative.

Scientific investigations tend to be time-consuming, multistage, multi-investigator processes fraught with many steps that are twists and turns dictated by the quest to fit the model to data. One of the major appeals of a simulation study is that it purports to possess the *truth*.

Ayer (1936) made a distinction between analytical propositions and empirical propositions. The evaluation of empirical propositions requires collection of empirical evidence that is directly observable by the senses or, in some cases, using calibrated scientific instruments. Empirical evidence may be collected via experiments or from observation studies. The distinction between empirical and analytical differs from the distinction between data collected experimentally or via naturalistic observation, both of which are empirical manifestations of the material world.

Analytical propositions are a mainstay of mathematics, where one starts with a set of axioms or assumptions and proceeds to prove or disprove the truth of propositions. A mathematical proof of a meaningful analytical statement, however, does not constitute empirical verification.

Simulations can make complex abstract models more concrete and easier to understand. Simulations of data from a particular mathematical model can inform us about the statistical properties of parameter estimates under the conditions of the simulation. As deductive demonstrations, they are very valuable for didactic purposes. They can demonstrate what the mathematical model means in terms of numbers and graphs, which are more readily understood than equations.

Simulations are very valuable for stress testing of models, checking how much the validity of deductions from a model are affected by violations of model assumptions. Tucker, Koopman, and Linn (1969) illustrated this approach to stress testing in the context of factor analysis procedures. Sinharay and Holland (2009, 2010) stress tested three curvilinear equating procedures, as will be illustrated in the section on simulations of anchor test equatings.

Simulations can have negative value, however, when the numbers generated by a simulation are mistaken for nature itself. One prominent example is the battle of the logistic models that proponents of the Rasch model and the three-parameter logistic (3PL) model have had over the decades. Using the Rasch model to generate data will demonstrate its parsimonious superiority to the cumbersome 3PL model. Likewise, using the 3PL model to generate data will yield results that demonstrate its superior fit over the miserly Rasch model. Has either simulation demonstrated which model works best with real data? No. Suppose a simulation uses a model of test-taker performance (e.g., how test takers perform when they run short of testing time) to confirm the validity of the model. When conducted correctly, simulations produce results that are logically consistent with the assumptions employed to generate the data. They should yield expected answers. They have not demonstrated the validity of the model, however.

If a simulation produces results that surprise the simulator, several possibilities exist. The author of the simulation may not have understood the model underlying the simulation well enough to derive the correct analytical propositions or combinations of premises and conclusions. The author may not have executed the simulation properly (e.g., may have made errors in coding). The author, perhaps in an effort to imitate reality better, may have made the simulation so complex that the results were complicated and difficult to interpret. The choice of the word *author* is intentional; simulators are authors of fiction. Some fiction mimics reality. Some fiction is fantastical. The best fiction often contrives particulars to illustrate universals.

For fields like chemistry, an old science with an extensive empirical base and well-developed theories, it is possible to collect data under controlled conditions and make precise testable predictions. Consider the ideal gas law. This law, which can be found in introductory chemistry books, is a single equation that relates volume (V) of a vessel, absolute pressure (P), absolute temperature (T), and the quantity of gas measured in number of moles (n),

$$PV = nRT,$$

where R is the ideal gas constant. The ideal gas law was based on empirical investigations. Over time, repeated application of the scientific method led to the development in 1910 of van der Waals equation,

$$P + \frac{n^2 a}{V^2}(V - nb) = nRT,$$

as a modification of the ideal gas law. This equation approximates the behavior of real fluids, taking into account the nonzero size of molecules via the term b and the attraction between them via the term a. If the ideal gas law served as the data-generating model for a simulation and the simulated results were compared to data collected under highly controlled conditions, we would see that the simulated numbers would be highly related to the observed data. Examination of residuals and revising the math should lead to something like the van der Waals equation. The key point is that real data are necessary to assess the empirical validity of hypotheses (Dorans & Walker, 2013).

In *The Signal and the Noise*, Nate Silver (2012) examined how well a variety of disciplines predict data in their domains. In some fields, rich empirical data are collected in naturalistic settings, as is the case with meteorology. Simulations based on mathematical models that are grounded in these rich empirical data produce fairly accurate predictions of short-term weather. As the time interval increases, prediction becomes more difficult as more and more uncontrolled variation enters the system, as noted by Silver in Chapter 4.

In contrast to weather prediction, simulations in other fields are based on models that are consistent with the theoretical preferences of the simulator, preferences that may not be rooted in solid empirical ground. For example, the credit rating agencies rated credit default swaps highly in part because they used simulation models for risk analysis that were based on flawed assumptions (Silver, 2012, Chapter 1). They assumed that the probabilities of default for mortgages packaged together were independent. Hence, the risk associated with a default on the collection of these mortgages was presumed to be very small. These independence assumptions enabled them to package together collections of high-risk mortgages and sell them as low-risk investments. They also assumed (hoped) that housing prices would continue to rise. When housing prices did not continue to rise, and defaults among high-risk homeowners turned out to be quite correlated instead of independent, these financial instruments became worthless and the financial crises of 2008 ensued. These financial companies understood the mathematics of risk but grossly underestimated the uncertainty of the gamble they were taking. Those who did recognize the folly of the independence assumption and the likelihood that housing prices would not rise forever made a fortune by shorting the market on these instruments, as described in *The Big Short* by Michael Lewis (2010).

Simulations are used widely in domains where science underpinnings are well established. Flight simulation is an oft-cited example. Instead of crashing planes and losing the lives of pilots in the process, changes in planes are subjected to extensive simulated exercises before the planes are released for use with live passengers. The effectiveness of space exploration owes much to careful simulations that were rooted in well-supported scientific theories about nature.

Likewise, pilot skill can be effectively taught and evaluated with simulators. I would hesitate, however, to let the skills at coping with crises situations rest on findings with a handful of pilots. As a means of eliminating the variability associated with pilot skill and judgment, suppose that the latest models of human behavior were used to simulate how pilots would react in an emergency. Based on these models, avatars could be substituted for real pilots in a flight simulator. This substitution is an example of a simulation that is not rooted in well-supported scientific theory.

Closer to home, consider the plethora of simulations that generate item or test data that are consistent with a particular mathematical model. These simulations may or may not have pertinence for reality as manifested in observed data because the numbers produced by the simulations are merely consequences of the model used to generate the data, and the model may or may not adequately describe reality. This statement is as true for the environment in which items and tests are administered to examinees as it is for the carefully controlled experimental world in which variation of P, V, and T led to the discovery of the ideal gas law or the data-rich world of meteorology or the speculative world of financing debt instruments. In the next section, I use score equating simulations to illustrate some limitations of relying on simulated data to resolve questions of substance.

Simulations About Anchor Test Equating Methods

The purpose of score equating is to produce, to the extent possible, scores from two or more tests that can be used interchangeably. The anchor test design is often employed to equate scores. In anchor test designs there are two populations, P and Q, with a sample of test takers from P taking test X and a sample from Q taking test Y. In addition, both samples take an anchor test. Table 1 represents the anchor test design. The symbol @ in the cell of the table indicates that the sample from the population denoted by the row took the test or anchor indicated by the column. Von Davier, Holland, and Thayer (2004) and Holland and Dorans (2006) called this the nonequivalent groups with anchor test (or NEAT) design. Kolen and Brennan (2004) and others referred to this as the common-item nonequivalent groups design.

Table 1 Design Table for the Anchor Test Design

	Test X	Anchor A	Test Y
Population P	@	@	
Population Q		@	@

Note: The symbol, @, in the cell of the table indicates that the sample from the population denoted by the row took the test or anchor indicated by the column.

The role of the anchor test is to quantify the differences in ability between samples from P and Q that affect their performance on the two tests to be equated, X and Y. The best kind of anchor for equating is a test that measures the same construct that X and Y measure. The anchor, A, is usually, but not always, a shorter and less reliable test than the tests to be equated.

The use of common items requires the use of assumptions to make up for the fact that X is never observed for test takers in Q and that Y is never observed for test takers in P. For this reason, several distinct methods of scaling and equating tests use the NEAT design. Three types of methods are usually studied: statistical approaches known as *poststratification*, psychometric approaches that make assumptions about *true scores*, and approaches that decompose the anchor test design into two single group designs and *chain* the two single group linkings though the anchor test. Each of these types of methods makes different untestable assumptions about the missing data. The Appendix describes linear versions of these three approaches.

Much research has been conducted on methods employed with the anchor test design in score equating. This research has been motivated by the divergence of results obtained by different methods when anchor test score distributions differ and correlations between the anchor and total tests diverge more and more from one. Dorans, Liu, and Hammond (2008) summarized several simulation studies that used data from the *SAT*® test, including Dorans (1990); Eignor, Stocking, and Cook (1989); Lawrence and Dorans (1990); Livingston, Dorans, and Wright (1990); and Wright and Dorans (1993). These studies varied in the way in which real data were manipulated to produce simulated samples of test takers.

Lawrence and Dorans (1990) used the verbal anchor to create differences from the reference or base population and the pseudopopulations, and the same verbal anchor was used to equate the tests. Under these circumstances, the poststratification methods did best, and the true-score methods did slightly worse than the chained method. Eignor et al. (1989) used an item response theory (IRT) model to simulate data and found that the weakest results were obtained for poststratification on the basis of the verbal anchor and that the true-score methods were slightly better than the chained method. The Livingston et al. (1990) study used SAT Math to create difference in populations, and the results for the poststratification method were not good for equating verbal scores.

In the studies cited previously, the methods that won (or lost) depended on how the data were simulated. Nearly 20 years after these studies, Sinharay and Holland (2009, 2010) demonstrated conclusively how simulation models can be used to produce a winner. Holland (2004) had long viewed the anchor test design as a missing data design (Braun & Holland, 1982; Holland & Dorans, 2006; Holland & Wightman, 1982; von Davier et al., 2004). His work with Sinharay illustrated how to simulate data that are consistent with the assumptions about missingness made by the poststratification equipercentile equating, chained equipercentile equating, and observed score equating based on an IRT model that makes assumptions about true-score relationships. Construction of the data based on poststratification was straightforward. Likewise construction of data consistent with the IRT model was relatively easy. For chained equating, Sinharay and Holland (2009, 2010) used a relatively complex procedure called raking (Bishop, Fienberg, & Holland, 1975), however, to fill in the missing data in a manner that is consistent with chained equating.

Having filled in the missing data in each of the three ways, Sinharay and Holland (2009, 2010) computed equating f unctions in the complete data and used these as target equating functions in the subsequent simulations. Then, they computed equating functions from the observed data with each of the three methods: poststratification, chained, and IRT observed score equating. They compared each of these observed equating functions with the target equating functions obtained in the complete data sets that were constructed to be consistent with the assumption underlying the three methods.

Two major findings resulted. First, the poststratification method worked best at reproducing the equating function in the population constructed under the poststratification assumptions, the chained method worked best in the population constructed under the chained assumptions, and the IRT observed score equating worked best with the population

constructed according to the IRT assumptions. In short, the method most consistent with the data used to construct the population was the winner.

The second point, which replicated what has been found elsewhere, was that the chained equating method either finished first (when the data were constructed according to the chained assumptions) or second (when the data were constructed according to either the poststratification assumptions or the IRT assumptions) in the three simulation contests.

Chained equating appears to be the winner. It often is in simulations. Why is it the winner? Its success may be due to its assumptions being the most consistent with the data. Or it may win because it tends to fall between a psychometric method that uses assumptions about true scores and observed scores and a statistical method that makes adjustments in accord with the degree of relationship between the anchor and the total test score.

Like many compromises, the chained approach won not because it was always correct, but because it was less likely to be as incorrect as the other approaches might be. The use of different simulation models in the studies by Sinharay and Holland (2009, 2010) demonstrated this point; even when it was the wrong model, the chained method was less incorrect than the other wrong model for the data simulated to be consistent with the correct model.

The Appendix presents analytical relationships among the chained linear equating, Tucker equating, and Levine observed score equating methods. Tucker equating is actually the linear form of the poststratification method. Chained linear can be thought of as a linear version of the chained method. Levine, like IRT observed score equating, uses assumptions about true scores. The Appendix describes why chained linear falls between the other two linear methods.

Simulate to Understand Models, Not Nature

The findings of Sinharay and Holland (2009, 2010) and the analysis in the Appendix demonstrated that simulations reveal what they should be expected to reveal. Whether creating pseudotests with artificial differences in difficulty or creating pseudopopulations on the basis of some variable, assumptions are made that should have predictable and significant effects on the outcomes of the simulation.

Empirical investigations are difficult to conduct and rarely lead to truths. In contrast, simulations are so easy to do. In addition, they come with a *known truth* that is often touted as an advantage of simulations. The numbers produced by simulations, however, are not empirical data; they are manifestations of analytical propositions that can be deduced from a generating model. The analysis of numbers generated by a simulation design by different methods enables one to make comparisons to the truth of the model but not to truth about whether one method's assumptions are more reasonable than another method's assumptions in real data settings. As Sinharay and Holland (2009, 2010) demonstrated, the truth of a simulation depends on the assumptions that went into it. The results are analytical consequences of the simulation. These consequences may have little bearing on reality. Simulations are valuable tools for demonstrations about the presumed model but not necessarily useful for investigating what happens with real tests given to real people.

Working with real data as a starting point for a simulation imbues the simulation with an aura of authenticity that it does not possess. The results of the Livingston et al. (1990) and Wright and Dorans (1993) studies illustrated this point most vividly. For example, pseudopopulations for SAT Verbal equatings were constructed on the basis of SAT Math score distributions. When the verbal anchor was used as the anchor for the SAT Verbal equatings, the poststratification methods tended to be outperformed by both Levine and the chained methods. When SAT Math was used as the anchor for the SAT Verbal equatings, the Levine and chained methods produced very biased results whereas the poststratification methods performed better than they had with the verbal anchor. In fact, the best equating results were obtained when poststratification was used with the SAT Math score, which is as expected because it was the math score that was used to construct the pseudopopulations. A bizarre simulation, stratification on the other test score, produced bizarre results, results that could be explained by the simulation model.

What does this pair of studies have to do with reality? They used the same real data (and lots of it). This fact might lead to the conclusion that each study had implications for practice. But that inference would be unwise. The manner in which the samples were constructed influenced the results. If the same findings had been obtained with a simulation study that generated the data from a model, they would have lacked the authenticity conferred to them by the fact that they used real data as a starting point. The use of the SAT database as the starting point imbued the studies with a face validity that may or may not have any meaningful bearing on reality. The unanswered question is this: How well do these artificial simulations reflect what really led to population differences?[1]

I used these selected equating studies to demonstrate that simulation studies are demonstrations of the implications of the model used to generate the simulated data. This is true when the simulated data are numbers deduced via the manipulation of variables in a mathematical equation. It is also true when simulated differences in tests or samples of test takers are induced via manipulation of real data. Simulations are analytical exercises involving the generation of numbers that are consistent with a set of assumptions. Simulations yield predictable outcomes; empirical investigations do not. Simulations cannot provide evaluations of the empirical validity of their predictions. Observable data from carefully designed and properly analyzed studies are needed for empirical evaluations.

To illustrate this latter point, the adequacy of equating methods can be evaluated directly without a simulation if one is willing to make a relatively weak assumption, namely that large sample equivalent groups equating is a solid criterion. Administering the same pair of tests, X and Y, with the same anchor, A, to random samples from two populations, P and Q, will provide data for assessing different equating methods. The populations can be subpopulations, for example, females (P) and males (Q), from a population, T, in which equivalent groups receive one of the two tests, X or Y, and the anchor, A. This large sample experiment could test several hypotheses pertinent to equating. First and foremost, does the linking between the tests meet the two most easily assessed requirements of equating mentioned by Holland and Dorans (2006): equal reliability of scores on X and Y in a common population, be it P, Q, or T, and invariance of linking relationships between X and Y across P, Q, and T. Second, how do the anchor test equatings approximate the equivalent groups equating, provided it is the same in both populations P and Q? This real data experiment would avoid the use of artificial tests. It would also avoid the creation of artificial populations. It simply requires a strong data collection, namely administration of tests X, Y, and A to populations P and Q.

Simulation as Substitution

In *Thinking Fast and Slow*, author Daniel Kahneman (2011) contrasted *system 1* thinking and *system 2* thinking. System 1 thinking is fast, automatic, frequent, emotional, stereotypic, and subconscious, relying on heuristics to arrive at conclusions. System 1 is engaged when we walk effortlessly through the woods listening to a favorite piece of music. In contrast, system 2 thinking is slow, effortful, infrequent, logical, calculating, and conscious. If we tried to multiply 147 by 59 as we walked through the woods, we would probably come to a stop and concentrate on the math problem as our system 2 took charge.

Because system 1 thinking is effortless, Kahneman (2011) noted that it is prone to biases such as stereotyping and overgeneralization. One common bias is attribute substitution, in which a simple question is substituted for a more difficult one. Simulations are so easy to do and have a "truth" embedded in them. In contrast, empirical investigations are often difficult to conduct and rarely lead to universal truths. It should make it obvious, therefore, that the substitution of a simulation for empirical investigation is suspect. According to Ayer (1936), analytic statements, such as the statements of logic and mathematics, are tautologies. Their truth can be analytically derived. Simulations are tautologies. Empirical propositions, in contrast, make assertions about nature. The validity of the proposition requires empirical verification. Substituting an analytical consequence for empirical data is not science.

When a simulation is used as substitution for the assessment of hypotheses about real data, a false sense of understanding can ensue and with it a biased perspective on the world of nature. While mistaking the results of a simulation study for empirical science in the field of education is unlikely to contribute to a global economic crisis, it could have an adverse affect on inferences about educational issues and practices.

Acknowledgments

The author appreciated receiving comments from Shelby Haberman, Samuel Livingston, J. R. Lockwood, Michael Walker, Gautam Puhan, Mark Reckase, Peter van Rijn, Sandip Sinharay, Jonathon Weeks and Rebecca Zwick on earlier versions of this paper. The opinions expressed herein are those of the author.

Note

1 In the process of simulating the data from real data, the answer to the question under investigation may change in a significant manner. For example, in the process of creating parallel pseudotests, the degree of correlation between the anchor and the total

test score is lowered because these tests are shorter and its effect of the simulation on the results is predictable: Poststratification will be most negatively affected, a model-based method such as Levine and IRT will be less negatively affected, and chained linear will be somewhere in the middle of the two.

References

Ayer, A. J. (1936). *Language, truth and logic*. London, England: Gollancz.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis. Theory and practice*. Cambridge, MA: MIT Press.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York, NY: Academic Press.

Dorans, N. J. (Ed.). (1990). Selecting samples for equating: To match or not to match [Special issue]. *Applied Measurement in Education*, *3*(1).

Dorans, N. J., & Walker, M. E. (2013). Multiple test forms for large scale assessments: Making the real more ideal via empirically verified assessment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology Vol. 3: Testing and assessment in school psychology and education* (pp. 495–515). Washington, DC: American Psychological Association.

Dorans, N. J. (2014). *Simulate to understand models, not nature* (Research Report No. RR-14-15). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12013.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement*, *32*(1), 81–97.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1989). Simulation results of the effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education*, *3*, 37–52.

Holland, P. W. (2004, January). *Three methods of linear equating for the NEAT design*. (Unpublished manuscript).

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education/Prager.

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating: A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 271–297). New York, NY: Academic Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Strauss and Giroux.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Krauss, L. W. (2011). *Quantum man: Richard Feynman's life in science*. New York, NY: W. W. Norton.

Lawrence, I. M., & Dorans, N. J. (1990). A comparison of several equating methods for representative samples and samples matched on an anchor test. *Applied Measurement in Education*, *3*, 19–36.

Lewis, M. (2010). *The big short*. New York, NY: W. W. Norton.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, *3*, 73–95.

Silver, N. (2012). *The signal and the noise*. New York, NY: Penguin Press.

Sinharay, S., & Holland, P. W. (2009). *The missing data assumptions of the NEAT design and their implications for test equating* (Research Report No. RR-09-16). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, *47*, 261–285.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer Verlag.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). Princeton, NJ: Educational Testing Service.

Appendix

Analytical Relationships Among Three Linear Equating Models

In linear equating, a transformation is found such that scores on test X and test Y are said to be equated if they correspond to the same number of standard deviation units above and below the mean in T, where T is the population in which the equating is performed.

The Tucker or poststratification linear equating method assumes that the regression of total score, Y, onto the equating or anchor test score, A, is linear and homoscedastic and that this regression, which is observed in the sample that took the test, Y, with A, also holds in the sample that took the test, X, with A. A similar set of assumptions is made about the regression of X scores on A scores.

The Levine linear equating method assumes that the true scores on Y and A are perfectly related and that the ratio of the standard deviation of true scores on Y to the standard deviation of true scores on A is the same in the observed group, Q, and the synthetic population, T, created from a mixture of the old and new form samples. In addition, it assumes that the intercept of the regression line relating true scores on Y to true scores on A is the same in Q and T. Further, it assumes that the standard error of measurement for scores on Y and A is the same for groups Q and T. A similar set of assumptions is made about true scores on X and A in the observed groups P and T.

Chained linear equating assumes that the mean/sigma linking relationship between A and X scores in P would be the same if it were observed in Q. Likewise, it assumes that the mean/sigma linking relationship that exists in Q between Y and A scores would be the same in P if it were observed there.

Holland (2004 cited in Dorans et al., 2008) made the simplifying assumption that the slopes of these three equating functions are equal. This assumption is met when scores on tests X and Y are essentially tau-equivalent, the anchor A measures the same constructs as scores on X and Y, and scores on A have the same reliability in P and Q. He derived the following expressions for their intercepts:

$$\text{Chained linear}: \quad \mu_{YQ} - B\mu_{XP} + \left(\sigma_{YQ}/\sigma_{AQ}\right)\left(\mu_{AP} - \mu_{AQ}\right), \tag{A1}$$

$$\text{Tucker}: \quad \mu_{YQ} - B\mu_{XP} + C_T\left(\sigma_{YQ}/\sigma_{AQ}\right)\left(\mu_{AP} - \mu_{AQ}\right), \tag{A2}$$

$$\text{Levine}: \quad \mu_{YQ} - B\mu_{XP} + C_L\left(\sigma_{YQ}/\sigma_{AQ}\right)\left(\mu_{AP} - \mu_{AQ}\right), \tag{A3}$$

where

$$C_T = (1 - w)\,\rho_{XAP} + w\rho_{YAQ}, \tag{A4}$$

$$C_L = (1 - w)\left(\rho_{XXP}/\rho_{AAP}\right) + w\left(\rho_{YYQ}/\rho_{AAQ}\right), \tag{A5}$$

and where B is the common slope for the equating of X scores to Y scores; μ_{YQ} and μ_{XP} are observed means on Y in Q and X in P, respectively; and μ_{AP} and μ_{AQ} are the means of A in P and Q, and the corresponding standard deviations are represented by σ terms.

In Equation A4 for the Tucker method, ρ_{XAP} and ρ_{YAQ} are the correlations of A scores with X and Y scores in P and Q, respectively. In Equation A5 for the Levine method, ρ_{XXP} and ρ_{AAP} are the reliabilities of A and X scores in P, and ρ_{YYQ} and ρ_{AAQ} are the reliabilities of A and Y scores in Q. In Equations A4 and A5, w is the weight assigned to P to create the synthetic population $T = wP + (1 - w)Q$. Note that w does not appear in the expression for the chained linear method in Equation A1.

If the correlation between anchor and total test score is 1, all three methods converge to the same equation. This is obvious in Equation A4 for Tucker equating, where the term C_T becomes 1. In Equation A5, a perfect anchor test and total test score correlation implies perfect reliability for X, Y, and A, and so C_L also becomes 1. From the perspective of the Tucker approach, chained linear equating assumes a perfect correlation, and from the perspective of the Levine method, which is rooted in classical test theory, the perfect correlation implies perfect reliabilities.

Holland (2004) demonstrated from the equations above that:

$$C_T < 1 < C_L. \tag{A6}$$

This inequality in Equation A6 means that Tucker equating will tend to adjust scores less than chained equating, which will adjust scores less than Levine equating. As the correlation between anchor and total drops, the Tucker procedure will adjust less and less, whereas the Levine procedure will adjust more and more. The sensitivity of the Tucker method to the correlation is apparent in Equation A4. Note in Equation A5 that the reliabilities of X and Y, presumed fixed, are divided by the reliability of A, presumed to vary with changes in the relationship between the total tests and the anchor administered

with them. In essence, the Tucker method discounts the information in the anchor as the correlation between total score and anchor score drops. In contrast, the Levine method makes bigger and bigger adjustments because differences on the anchor are viewed as attenuated more and more by the reliability of the anchor as lower anchor total correlations are attributed by the approach to lower anchor test reliabilities. Chained linear ignores all bivariate information and sets means and standard deviations equal.

Equation A6 has implications for interpreting the results of the simulation studies conducted by Sinharay and Holland (2009, 2010). As expected, chained worked best with data that are consistent with its missingness assumptions. As noted above, the chained equating method finished first or second (when the data were constructed according to either the poststratification assumptions or the IRT assumptions) in the three simulation contests. Given Equation A6, chained is expected to work better than poststratification with data constructed from a psychometric model, IRT in the Sinharay and Holland simulations, which views observed data as a fallible version of the underlying data, than does a statistical method that adjusts to the degree to which the covariate or anchor is trustworthy and relevant. Likewise, it is also expected to work better with data constructed to be consistent with the missingness associated with the poststratification method than a method that presumes that the low correlation between the anchor and the total is simply due to the fallibility of the data.

Suggested citation:

Dorans, N. J. (2014). *Simulate to understand models, not nature* (Research Report No. RR-14-16). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12013

Action Editor: Rebecca Zwick

Reviewers: Skip Livingston and Michael Walker