Rethinking Measurement in Higher Education Research

Kenneth D. Royal and Kelly D. Bradley

University of Kentucky

Paper presented at the 2008 Mid-Western Educational Research Association

Abstract

Higher education research has long relied on Classical Test Theory (CTT) principles, despite compelling arguments made by measurement theorists that suggest CTT techniques make a number of erroneous assumptions. Many measurement theorists argue Item Response Theory (IRT) techniques overcome many of CTTs deficiencies and lead to more valid and reliable results. Despite the popularity of IRT models in fields such as psychology, medicine and business, the transfer and implementation of this knowledge to the higher education literature has yet to occur. The purpose of this study is to call attention to measurement issues in higher education research by examining the extent to which CTT approaches have been employed in higher education research. A meta-analysis of quantitative research published in the widely regarded top three, premiere higher education journals over a five year period were examined, and the extent to which the research appearing in those journals utilized a CTT or IRT approach were identified.

Rethinking Measurement in Higher Education Research

> "The lack of attention to measurement issues is one of the major deficiencies in the higher education research literature" (Smart, 2005, p. 470).

John Smart, editor of *Research in Higher Education*, reflected upon his long career in academe ranging from his experience as a doctoral student, a higher education scholar, and as an experienced editor for various scholarly publications in his "Perspectives of the Editor" article. In the article, Smart outlined what he believed to be attributes of exemplary manuscripts that employ quantitative analyses. He discussed the paramount importance of measurement in quality research and stated "Exemplary manuscripts… use measures that have established psychometric merit, and they provide evidence of the reliability and validity of those measures. Such attributes are rarely evident in the higher education research literature" (Smart, 2005, p. 470). He went on to posit that a number of higher education researchers possess strong statistical skills, but few are actually trained in measurement.

Hutchinson and Lovell (2004) offer support for Smart's comments in a meta-analysis of the methods employed in the three premiere higher education journals (*Journal of Higher Education*, *Research in Higher Education*, and the *Review of Higher Education)* "The methodologies showcased in the three journals… suggest that higher education researchers possess fairly strong methodological skills in statistical analyses, but somewhat limited training in measurement" (p. 398). The authors go on to address the types of analyses performed and provide counts and frequencies for the various techniques. Hutchinson and Lovell found nearly all quantitative analytical techniques incorporated a classical test theory (CTT) approach. This suggests a great deal of previous research may have ignored the principles of sound measurement and hastily analyzed data without great concern to measurement.

In response to both Smart and Hutchinson and Lovell's conclusions, a need surfaces to call attention to issues of measurement and expose deficiencies of training and/or practice in the current higher education arena. Conducting quantitative research without proper attention to measurement is problematic because measurement is a fundamental component of quality research. Measurement issues should be adequately addressed before any analyses are performed. Although the CTT approach has its strengths and purposes, an Item Response Theory (IRT) approach may be more appropriate for many quantitative studies, especially those that employ survey research techniques. Accordingly, this study will call to light many of the assumptions of the CTT approach and will make an argument for IRT. A meta-analysis of quantitative higher education literature published in the three premiere higher education journals over a five year period will be examined, and the extent to which the research utilizes a CTT or IRT will be identified.

Applications and Assumptions of Classical Test Theory

Classical Test Theory (CTT) was introduced in 1904 by Charles Spearman. "CTT is based upon conceptual models in which relations among constructs are theorized… from theories ground in previously published literature. Once a conceptual model of the relationships among different variables has been established, a measurement model can be constructed" (Embretson and Hershberger, 1999, p. 5). Generally, CTT is used to examine a group of individuals' responses to a test. As suggested above, a mathematical model is then applied to fit the data.

CTT is often criticized for several important reasons: First, all measurement units are considered equivalent (Becker, 2001). Second, error across measurement units is independent and uncorrelated (Becker). Other cited disadvantages of CTT include the argument that it is sample dependent and requires larger samples and/or test items (Bond & Fox, 2001; Bunderson,

2000; Hambleton, Swaminathan, & Rogers, 1991), as well as its use of test-retest reliability. Richter, Werne, Heerlein, Kraus, and Sauer (1998) argue test-retest reliability is problematic due to timing issues, meaning there is too much time between initial and follow-up administration which might lead to an underestimation of measures. Likewise, too little time between initial and follow-up administration might lead to an overestimation of measures.

CTT approaches make three major, erroneous assumptions within survey research. Briefly outlined, the assumptions are:

- There are equal distance between units of measurements
- Each item is of equal importance
- Scales are interval (whereas they are actually ordinal)

Take for instance a typical Likert-type scale with response options “Strongly Disagree” (SD), “Disagree” (D), “Agree” (A), and “Strongly Agree” (SA). In theory, the distance between SD and D would be equal to the distance between D and A, and so on. This concept is illustrated below:

SD	D	A	SA

In reality, the psychometric proximity between responses can vary considerably depending upon the content of the survey, the way items are phrased, etc. An actual response scale might look something like this:

SD	D	A	SA

To demonstrate the notion of items having unequal importance, consider an example provided by Bond and Fox (2001). A sample of grade school children were asked two questions: 1) I am so afraid of computers I avoid using them.

2) I am afraid that I will make mistakes when I use my computer.

Clearly, the first item demonstrates a greater fear of computers. Given these qualitative differences exist among survey items, why should survey researchers treat each item of equal importance?

Further clouding quality survey research is the notion of interval scales of measurement. Hays (1988) writes:

> The problem of measurement, and especially of attaining interval scales, is an extremely serious one for the social and behavioral sciences. It is unfortunate that in their search for quantitative methods, researchers sometimes overlook the question of level of measurement and tend to read quite unjustified meanings into their results. …However, the core problem of level of measurement lies outside the province of mathematics and statistics (p. 71).

As Hays suggests, interval scales are not possible in the human/behavioral sciences. Data appearing on Likert-type survey are actually ordinal in nature. Typically, the problem of treating ordinal data as interval is further compounded when researchers apply linear statistical techniques to these nonlinear data. Ordinal data is qualitative. It is inappropriate to apply quantitative techniques to qualitative data without proper treatment of the data. Fortunately, there is a technique than can overcome many of the deficiencies of the CTT approach and its assumptions in survey research.

Argument for Item Response Theory and The Rasch Measurement Model

Some sixty years after Spearman introduced CTT, scholars began to re-examine CTT and its assumptions and began to develop new models with stronger theoretical underpinnings. "CTT does not invoke a complex theoretical model to relate an examinee’s ability to succeed on a

particular item. Instead CTT collectively considers a pool of examinees and empirically examines their success rate on an item" (Fan, 1998, p. 358). The more theoretically sound solution came to be known as Item Response Theory (IRT). Bond and Fox (2001) define IRT as “a relatively recent development in psychometric theory that overcomes deficiencies of the classical test theory with a family of models to assess model-data fit and evaluate educational and psychological tests” (p. 231).

There are a number of differences between IRT models and CTT. According to Fan (1998), CTT focuses on test level information whereas IRT focuses on item-level information. In other words, IRT focuses on the interactions between individual persons and items, as the model suggests, each affects the other. Fan suggests IRT models assume a single trait is responsible for the subject’s response to a particular item. Another major difference is CTT assumes test-takers have both observed and true scores, where the observed score is an estimate of the true score plus or minus measurement error (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985). IRT, on the other hand, assumes the person’s ability is independent of the content of a test, and the relationship between the probability of choosing the correct answer and the ability of the person can be modeled differently depending on the content of the test (Hambleton, et al., 1991). This explains why IRT models generally assume unidimensionality, or the notion that test items measure a single trait.

One model, in particular, is extremely useful for analyzing survey data, the Rasch model. One of the fundamental benefits of Rasch measurement is it overcomes the aforementioned assumptions many researchers make in survey research. Rasch measurement calibrates scales to determine the psychometric distance between response options. Recall, raw scores are not measures. The Rasch model converts raw scores to their natural logarithm and places them along

a ruler, thus allowing these measures to become truly interval. Bond and Fox (2001) say "the Rasch model treatment of Likert scale data is intuitively more satisfactory and mathematically more justifiable than the traditional 'allocate 1 2 3 4 5 and add them up' approach [of CTT]" (p. 71). The assumption of equal importance is overcome by controlling for both persons and items on the same metric. Further, with regards to missing data, CTT principles would involve throwing out the data and treating as "missing". The Rasch model, on the other hand, allows researchers to use all remaining data even if certain items are missing.

Bradley and Sampson (2005) have eloquently summarized other advantages of Rasch measurement stating:

> Whereas the classical model produces a descriptive summary based on statistical analysis, it is limited, if not absent, in the measurement capacity. In contrast, Rasch measurement tackles many of the deficiencies of the classical test model in that it has the capacity to incorporate missing data, produces validity and reliability measures for person measures and item calibrations, measures persons and items on the same metric, and is not dependent on the particulars of the sample. Applications of the Rasch model allow the researcher to identify where possible misinterpretation occurs and which items do not appear to measure the construct of interest, while producing information about the structure of the rating scale and the degree to which each item contributes to the construct. Thus, it provides a mathematically sound alternative to traditional approaches to survey data analysis (p. 13).

Further, the Rasch model requires researchers to ensure data to model fit and rating scale functioning before any analyses occur.

As an additional testament to the Rasch model's strength, Curtis & Keeves (1999), Peck (2001), Waugh (1999) and Wright and Masters (1981) concur the Rasch model is the only IRT model that adheres to the seven principles of true measurement (as stated below).

- Each item should function as intended;
- Each item can be positioned on a common scale;
- The scale should be an interval one;
- Each person can be located along the same common scale used for items;
- The responses should form a valid response pattern for each item;
- Estimates of precision must be available for all scale measures;
- Each item should retain its meaning and function across individuals and groups (Curtis & Keeves; Wright & Masters);

Measurement in Higher Education

Despite the praise bestowed by many measurement theorists on the Rasch measurement model, the dissemination of this powerful technique to other academic fields has been relatively slow. Historically, educational psychology has been at the forefront for the use of Rasch measurement, as the theory originated from psychometrics. Increasingly, the use of the Rasch model is becoming more and more popular in health-related disciplines, market research and education. The question remains to what extent is Rasch measurement used in the higher education research arena.

Measurement and Graduate Training

As discussed in the introduction of this study, the majority of quantitative research in the higher education arena lacks sound measurement. Interestingly, however, there is an abundance of researchers skilled in statistical techniques (Smart, 2005; Hutchinson and Lovell, 2004).

Hutchinson and Lovell (2004), Lovell and Hutchinson (2003), Lovell, Hutchinson and Fairweather (1999), and Aiken, West, Sechrest and Reno (1990) argue the problem with measurement has largely to do with many higher education graduate programs' exclusion of measurement courses from the curriculum. Hutchinson and Lovell state:

> In the field of higher education, the inattention to measurement likely reflects a lack of appropriate measurement training as suggested by a survey of research requirements among higher education doctoral programs conducted by Lovell et al. (1999) and Lovell and Hutchinson (2003). Of the higher education programs responding to the survey, few required measurement courses, and most tended to require only introductory level, statistically focused courses (p. 398).

The authors go on to conclude a persistent link exists between the attention measurement issues are given in doctoral training programs and that of measurement issues discussed in the premiere higher education journals. Hutchinson and Lovell say "the lack of awareness about measurement issues in the three journals reviewed in the current study seems to mirror the general inattention to measurement in many doctoral training programs" (p. 398).

Methodologies Used in the Higher Education Literature

Inspired by Hutchinson and Lovell's analysis of higher education journals, it is useful to conduct a meta-analysis to determine the frequency with which studies in the top higher education journals incorporated either a CTT or an IRT approach. Understanding the frequency of these approaches would allow one to more closely examine the quality of measurement taking place in higher education research.

Similar to Hutchinson and Lovell's 2004 study, the meta-analysis was begun by choosing the three journals considered to be the most prestigious in higher education; the *Journal of*

Higher Education (JHE), the *Review of Higher Education* (RHE) and *Research in Higher Education* (RsHE). A timeframe of five years was arbitrarily chosen, and articles which spanned from 2003 to the summer of 2007 were analyzed. Each article was examined in detail and the analysis techniques employed were cited, as well as relevant information regarding the authors and the journal. Once the lists were generated, a code was provided for each technique according to whether it falls under the criteria of a CTT or an IRT approach. Counts and frequencies were then generated. The results were astounding.

Results, Conclusions and Implications

Results reveal 96.8% of articles published in the JHE, 97.6% of articles in RHE, and 97.4% of articles in RsHe incorporate a CTT approach (See Table 1).

Table 1

Frequency of CTT and IRT Applications in Higher Education's Top Journals

	Journal of Higher Education		*Review of Higher Education*		*Research in Higher Education*	
	Count	Percent	Count	Percent	Count	Percent
CTT approach	61	96.8	41	97.6	149	97.4
IRT approach	2	3.2	1	2.4	4	2.6

This indicates only two to four percent of the quantitative research published in the past five years in these journals incorporated a methodological approach based on some form of item response theory.

Taking this meta-analysis a step further, the number of instances in which Rasch measurement was employed in all published higher education literature was investigated. Performing a search in multiple databases spanning approximately 4,700 academic journals,

conference papers, etc., the terms “higher education” was entered, the connector “AND”, and “Rasch measurement” in open search fields with no limitations. Results yielded only 21 records. Revising the terms to produce maximum hits, the phrases “higher education” AND “Rasch” AND “measurement” were entered into the search. Only 67 hits were recorded. Of those 67 articles, the vast majority were published in educational measurement journals. Exclusively searching the three premiere higher education journals, the word “Rasch” was entered to detect the most hits possible. Results revealed a total of three articles published in 1993, 1994, and 2000, respectively. Collectively, the results of this meta-analysis suggest there is little doubt there is a significant lack of research rooted in measurement theory in the higher education literature.

Although measurement theorists have been arguing for some time now that the solution to many of CTT’s deficiencies can be alleviated by incorporating an IRT approach, particularly the Rasch measurement model (Andrich, 1978; Bond & Fox, 2001; Bradley & Sampson, 2005; Masters, 1982; Smith & Smith, 2004; Wright & Stone, 1979), the transfer and implementation of this knowledge to the higher education literature has yet to occur, at least in the mainstream higher education literature. Acknowledging Hutchinson and Lovell’s (2004) findings and implementing Smart’s (2005) suggestions could yield several important theoretical and methodological questions for researchers.

Additionally, this research is intended to challenge other higher education researchers to explore issues of measurement within their own research. Rasch measurement is not intended to take the place of statistics, but rather to complement the use of statistics. Utilizing a theoretically-sound and mathematically-just approach like Rasch measurement eliminates many assumptions researchers often make regarding methodological issues. Therefore, once proper

measurement takes place, statistical techniques can then be applied and the results will become more precise, and possibly more meaningful.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist, 45*(6), 721-734.

Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrica, 43*, 561-573.

Becker, G. (2001). Controlling decremental and inflationary effects in reliability estimation resulting from violations of assumptions. *Psychological Reports, 89*, 403-424.

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ; Lawrence Erlbaum Associates, Inc.

Bradley, K. D., & Sampson, S. (Spring, 2005). A case for using Rasch Rating Scale analysis to assess the quality of measurement in survey research. *The Respondent*, 12-13.

Bunderson, C. V. (2000). *Design experiments, design science, and the philosophy of measured realism: Philosophical foundations of design experiments*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. April 24-28.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston, Inc.

Curtis, D. & Keeves, J. (1999). *The Course Experience Questionnaire as an institutional performance indicator*. Paper presented at the HERDSA Annual International Conference, July 12-15, Melbourne, Australia.

Embretson, S. E. & Hershberger, S., (1999). Eds. *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational & Psychological Measurement, 58*, 357-380.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. New York: Sage publications.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth: Holt, Rinehart and Winston.

Hutchinson, S. R., & Lovell, C. D. (2004). A review of methodological characteristics of research published in key journals in higher education: Implications for Graduate research training. *Research in Higher Education, 45*, 4, 383-403.

Lovell, C. D., & Hutchinson, S. R. (2003). *Research training requirements in higher education doctoral programs*. Manuscript in preparation.

Lovell, C. D., Hutchinson, S. R., & Fairweather, J. R. (1999). *Graduate student research preparation in higher education: Too little, just right, or too much… Implications for our future*. Paper presented at the Annual Meeting of the Association for the Study of Higher Education. November, 1999, Miami, FL.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Peck, B. (2001). *Monocultural education for a multicultural population: The "ethnic disadvantage' issue revisited*. Unpublished doctoral dissertation, Murdoch University.

Richter, P., Werne, J., Heerlein, A, & Sauer, H. (1998). On the validity of the Beck Depression Inventory. *Psychopathology, 31*, 160-168.

Smart, J. C. (2005). Attributes of exemplary research manuscripts employing quantitative analyses. *Research in Higher Education, 46*, 4, 461-477.

Smith, Jr., E. V., & Smith, R. (Eds.). (2004). *Introduction to Rasch measurement: Theory, models and applications*. Maple Grove, MN: JAM Press.

Waugh, R. (1999). Approaches to studying for students in higher education: A Rasch measurement model analysis: *British Journal of Educational Psychology*, *69*, 63-79.

Wright, B. D., & Masters, G. (1981). *The measurement of knowledge and attitude* (Research memorandum No. 30). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Stone, M. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.