2005-07-08

Scale Construction and Halo Effect in Secondary Student Ratings of Teacher Performance

Eric Paul Rogers
Brigham Young University - Provo

SCALE CONSTRUCTION AND HALO EFFECT IN SECONDARY STUDENT RATINGS OF TEACHER PERFORMANCE

by

Eric Paul Rogers

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Instructional Psychology and Technology

Brigham Young University

August 2005

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

Eric Paul Rogers

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Richard R Sudweeks, Chair
Date	Guy L. Dorius
Date	Paul F. Merrill
Date	Dennis Wright
Date	Stephen C. Yanchar

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Eric Paul Rogers in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

____
Date

____
Richard R Sudweeks
Chair, Graduate Committee

Accepted for the Department

____
Andrew S. Gibbons
Department Chair

Accepted for the School

____
K. Richard Young
Dean, School of Education

ABSTRACT

SCALE CONSTRUCTION AND HALO EFFECT IN SECONDARY STUDENT RATINGS OF TEACHER PERFORMANCE

Eric Paul Rogers

Department of Instructional Psychology and Technology

Doctor of Philosophy

The use of rating scales in the evaluation of secondary teacher performance has been called into question and widely criticized. Of particular concern has been the use of student ratings of teacher performance. A review of instruments and practices used in the rating process reveals serious design flaws that account for the criticisms leveled against the use of rating scales.

This study sought to address the limitations evident in previous rating efforts by utilizing a combination of design methodologies and measurement models including elements of Classical Test Theory (CTT), factor analysis, and Item Response Theory (IRT). The IRT model employed was the one-parameter logistic model also known as the Rasch model. Twelve scales were developed consisting of a total of ninety-two items. These scales were developed to facilitate student ratings of secondary level teachers of religion in the Church Educational System (CES) of the Church of Jesus Christ of Latter-day Saints (LDS).

In addition to exploring rating scale design methodology and scale performance, this study also examined a potential threat to the validity of decisions based on ratings referred to as *halo effect*. Using a variety of approaches to operationally define and estimate halo error, the extent to which male and female students exhibit differing degrees of halo in their ratings of teachers was examined.

The results of the study revealed that of the twelve teacher traits hypothesized in the design of the rating scales, only three met defensible criteria based on CTT and Rasch model standards: the Student-Teacher Rapport Scale (STRS), the Scripture Mastery Expectation Scale (SMES), and the Spiritual Learning Environment Scale (SLES). Secondary students were unable to meaningfully discriminate between all twelve traits. Traditional approaches to halo effect estimation suggest that males exhibited halo to a greater degree than females, whereas Rasch model approaches to halo effect estimation were less consistent. Considered together, however, the evidence suggests differential halo error by gender, with males exhibiting halo to a greater degree than females. The implications of these findings for teacher evaluation, instructional design, and future research efforts are also addressed.

ACKNOWLEDGMENTS

I express sincere appreciation to the members of my graduate committee and especial gratitude to my chair, Dr. Richard R Sudweeks, who has been, and continues to be, a remarkable mentor and friend; to Juan Henderson, who shared not only his office with me, but his intellect, his wisdom, and his love; to Dr. Raymond O. Kelly who lent considerable time and support to various aspects of this project and who continued to express confidence in me when my confidence in myself waned; to Dr. Pablo Riboldi who served as confidant, cheerleader, and sounding board; to Dr. Paul V. Johnson, Dr. Garry K. Moore, Mark Pugh, and Dr. Robert F. Elzey of the Church Educational System who encouraged me and gave institutional support to the project; to my other colleagues in the Church Educational System, especially Jessica Burton Kotter, Mollie Gulsby Turner, and Kenneth J. Plummer, whose friendship and talents have assisted and motivated me in countless ways; to my children, Katie, Lelsy, Elijah, Jacob, and Kinsey for their patience and love; to my wife, Kelly Ensign Rogers, for her inspirational love of learning and her dedication to great causes and to me; and finally to my Savior Jesus Christ, who is the Way, the Truth, and the Life.

Table of Contents

List of Tables

Table Page

List of Figures

Figure Page

List of Equations

Chapter 1: Introduction

Evaluating Teacher Performance

Evaluating teacher performance is a potentially frustrating process for both teachers and administrators. The sources of frustration lie in the complexity of the teaching process and in the tension between teachers and administrators over how teachers should be evaluated and how the results should be used. Wolf (1973) explains

> Teachers mistrust evaluation. They feel that current . . . techniques fall short of collecting information that accurately characterizes . . . performance. They perceive . . . rating as depending more on the idiosyncrasies of the rater than on their own behavior in the classroom. . . . Teachers see nothing to be gained from evaluation. (p. 160)

Management theory has shifted in recent decades toward a greater emphasis upon the value of the person over the organization (Drucker, 1998; Farson, 1996). This shift is reflected in approaches of educational administrators within the domain of teacher evaluation as well (Beerens, 2000; Danielson & McGreal, 2000; Peterson, 2000). Teacher evaluation is being increasingly viewed as the basis for professional growth and improved student learning, rather than simply grounds on which to make personnel judgments. However, in most instances practice lags behind the theory.

According to Danielson and McGreal (2000) current teacher evaluation practices are inadequate for a variety of reasons: (a) evaluation criteria are limited and outdated, (b) teachers and administrators share few values and assumptions about good teaching, (c) evaluation of performance lacks precision, (d) communication is hierarchical and one-way in nature, (e) evaluation practices do not differentiate between novice and experienced practitioners, and (f) administrators have limited evaluation expertise.

Peterson (2000) characterizes most teacher evaluation as “a principal’s report of teacher performance usually recorded on a checklist form, and sometimes accompanied by a brief meeting” (p. 18). Teachers question the validity of the decisions drawn from administrators’ checklists when the research bases for item selection are only vaguely addressed, if addressed at all. Peterson suggests that some school districts avoid this problem “by using open-ended comment topics (e.g., instructional practices)” (p. 19).

Use of Rating Scales in Evaluating Teacher Performance

Among the most vilified elements of teacher evaluation is the use of rating scales. According to Stanley and Popham (1988), “No single idea or concept has been more detrimental to successful teacher evaluation than the rating scale” (p. 24). Danielson and McGreal (2000) explain that “most evaluation systems depend on a single dichotomous scale, such as ‘satisfactory,’ ‘needs improvement,’ and the like” (p. 4). Criticizing the lack of objectivity and specificity inherent in a dichotomous rating, Danielson and McGreal suggest that rating scales with additional response categories (e.g., needs improvement, satisfactory, and outstanding) also suffer serious shortcomings including a lack of agreement on what constitutes each level of performance and the absence of “the equivalent of the ‘anchor papers’ or ‘benchmarks’ used in evaluating student work against rubrics” (p. 4).

One of the most controversial uses of rating scales within the domain of teacher evaluation is student ratings of teachers. Many teachers acknowledge the value of student feedback to improve instruction. However, most challenge the use of student ratings in the making of personnel decisions (Marsh, Overall, & Kesler, 1979). Frequently student rating forms include questions about areas of teacher performance that students are not in a position to judge reliably. Such concerns about the reliability of the data gathered using student ratings and the validity of decisions based on those ratings contribute to the negative view of student ratings of teacher performance.

These concerns have resulted in an inaccurate view of rating scales in general; a view that has caused many to dismiss rating scales in the evaluation of teacher performance. However, rejecting rating scales as a valuable aid in teacher evaluation is like rejecting a screwdriver because it does not saw well. The design, manufacture, and use of a tool will determine its utility. If, however, a screwdriver is used to make a precision cut in a fine piece of hardwood the user will, no doubt, be sadly disappointed. Is it the fault of the screwdriver? No. The fault lies in the poor choice and use of the tool by the user. If rating scales are to be used effectively in teacher evaluation they must be carefully designed, developed, and implemented in a manner tailored to specific evaluation tasks. In teacher evaluation settings the instrument of evaluation is really the rater not the rating scale. The human being who observes, interprets, and judges the relative adequacy of the teacher’s performance is the instrument and the rating scale is simply a job aid that directs and focuses the observer’s attention and provides a convenient basis for recording the ratings.

The rejection of rating scales used with students as judges eliminates perhaps the richest source of information about teacher performance in the classroom. Students see and experience their teacher on his good days as well as his bad. They see how he responds to questions and challenges, and how he reacts under pressure. Although their judgments should not be the sole basis for judging teachers, their views should certainly be considered.

Research Questions

The purpose of this study was twofold. The first purpose was to construct student rating scales to measure teaching performance that constructively address the pitfalls described above. In this study, the rating scales were tailored to the unique objective and commission of religious educators in the Church Educational System (CES) of The Church of Jesus Christ of Latter-day Saints. Existing instruments designed to measure teaching performance do not capture the unique definitions and methods promoted by CES. For example, instruments developed for teacher evaluation in the public schools and in higher education do not address competencies such as (a) effectively teaching the gospel of Jesus Christ, (b) teaching by the Spirit, (c) effectively using scripture study and teaching skills, or (d) preparing young people for effective church service. Yet these competencies are central to religious education as defined by CES. Additionally, previous instruments developed and implemented within CES primarily reflect teachers' rapport with students and have not effectively captured the broad range of competencies constituting teaching performance (Elzey, 1998; Sudweeks, 1979). Furthermore, the inadequacies of teacher evaluation practice summarized by Danielson and McGreal (2000) are evident in CES evaluation efforts as well (Elzey, 1998; Howell, 1995; Maughan, 1994). The present study, therefore, applies a theoretical and operational framework, not applied in previous CES efforts, in a prescriptive and formative process of scale construction, refinement, and assessment.

The second purpose of this study was to compare and contrast various approaches to estimating *halo effect*. Although previous research reveals several sources of error in rating data including (a) severity or leniency, (b) halo effect, (c) central tendency, (d) restriction of range, and (e) lack of inter-rater reliability (Hoyt & Kerns, 1999; Landy & Farr, 1980; K. R. Murphy & Balzer, 1989; K. R. Murphy, Jako, & Anhalt, 1993; F E Saal, Downey, & Lahey, 1980), the scope of this study was restricted to the examination of halo effect. Broadly defined, halo effect refers to the tendency of a rater to attend to a global impression of the ratee, such as student-teacher rapport, rather than to distinguish differing levels of performance on separate dimensions, such as different aspects of teaching ability. Halo effect, therefore, constitutes a potential threat to the validity of conclusions drawn from rating scale data. Although halo effect has been investigated from a variety of perspectives, the research literature does not investigate *variable halo effect* between subpopulations such as males and females.

Therefore, the specific research questions addressed in this study are

1. What are the key areas of teacher performance valued by CES administrators, teachers, and students?

2. In what ways do students conceptualize these areas of valued teacher performance?

3. To what degree do the items derived from student conceptualizations function to produce reliable ratings from which valid conclusions may be drawn about teacher performance?

4. In what ways should the items and scales be revised to improve reliability and validity?

5. To what extent do male and female seminary students exhibit differing degrees of halo effect in their ratings of teachers?

Overview of Chapters

Chapter 1 has provided a brief introduction to key issues associated with teacher evaluation, student ratings of teachers, and the purposes of this study. Chapter 2 examines the research literature relevant to the purposes of the study outlined in Chapter 1. Chapter 3 describes the method used in obtaining answers to each of the research questions. Chapter 4 presents the results of the various analyses. Finally, Chapter 5 discusses conclusions supported by the findings and recommendations for further research.

Chapter 2: Review of Literature

Measurement Practices

Ancient Examples

The problems with measurement of teacher performance are not new. Nor are the problems unique to education or the social sciences. Humans have struggled with evaluative judgments based on measurements from the origins of their existence. DeVellis (2003) summarizes the historical origins of measurement pointing first to ancient examples such as Proverbs 11:1: “A false balance is an abomination to the Lord, but a just weight is a delight.” Duncan (1984) refers to the writings of Aristotle in which officials were charged with policing weights and measures. Anastasi (1968) suggests that the Socratic method utilized in ancient Greece constituted a form of psychological testing. Citing an essay by DuBois in 1964 that was reprinted in Barnette (1976), DeVellis notes that the Chinese administered civil service tests as far back as 2200 BCE. The importance of accurate measurement in antiquity is also noted by Wright (1999) who explains that seventh century Muslims ascribed great significance to the unit of measurement referred to as the “weight of seven” upon which taxation was based. Wright also notes that some attribute the rise of the peasants during the French Revolution, at least in part, to unfair measurement practices.

Convergence of Statistical Methods, Mental Testing, and Psychophysics

Measurement and the evaluation of measurement practice has a long history. However, modern approaches to human ability measurement are generally considered to have their origins in the latter half of the nineteenth century. DeVellis (2003) cites the work of Sir Francis Galton who extended the work of his cousin, Charles Darwin, to systematic observation of human differences. Karl Pearson, a colleague of Galton's, developed "the mathematical tools, including the product-moment correlation coefficient bearing his name, needed to examine systematically relationships among variables" (DeVellis, 2003). Like Alfred Binet (1905), who created mental ability tests in France at the beginning of the twentieth century, many of the early contributors to formal psychological measurement theory and practice shared a keen interest in measuring human intelligence.

During this period, developments in statistics and mental testing converged with developments in psychophysics. S. Stevens argued that people could make ratio judgments about sound intensity and claimed that this ratio property enabled the data from such measurements to be subjected to mathematical manipulation (DeVellis, 2003). Stevens is credited with classifying measurements into nominal, ordinal, interval, and ratio scales. However, Stevens would credit Louis L. Thurstone (1927a, 1927b, 1927c, 1927d, 1931, 1935), who was working on the mathematical foundations of factor analysis, as the first to apply psychophysical scaling to social stimuli (Duncan, 1984).

Psychometrics as a Methodological Paradigm

Duncan (1984) argues, further, that the convergence of developments in statistical methods, mental testing, and psychophysics spawned a new methodological paradigm for measurement in the social sciences. He bases this argument upon the widespread acceptance and use of the concepts of reliability and validity, the popularity of factor analysis in social science research, and the use of the psychometrics paradigm in the development of scales designed to measure a diverse array of psychological and social phenomena.

Within the field of psychometrics a variety of approaches to conceptualizing and applying measurement has emerged out of the convergence of these disciplines. An overview of these approaches is necessary in order to understand the logic underlying the application of measurement models in the present study.

Measurement Models

Tests are designed to measure a variety of traits such as knowledge, attitudes, and performance and a broad range of personality variables and characteristics. Measurement models provide a foundation on which to relate performance on a test to the actual trait of interest.

Classical Test Theory

Based on the early pioneering work of Spearman (1907; 1913) and later Gulliksen (1950), classical test theory (CTT) has been the predominant paradigm for test development for the majority of the twentieth century. Under CTT a respondent's *observed score* on a test is perceived as consisting of a linear combination of *true score* and *error score*. This theoretical framework is frequently expressed as Observed Score = True Score + Error.

Lord and Novick (1968) articulate the relationship as $X_{ij} = \theta_i + \varepsilon_{ij}$. In this expression X_{ij} is the observed score of person i for the j-th measurement protocol (e.g., test of knowledge, attitude survey, or performance rating). The constant θ (theta), or true score, is the true, but unknown, value of person *i* on the characteristic being measured, and ε_{ij} is the error in the observation. CTT defines true score for the *i*-th individual on the *j*-th measurement as the expected value of the observed score, $\theta_i \equiv E\{X_{ij}\}$. In other words, true score is equivalent to the mean of repeated and independent observations of X_{ij} under identical conditions. Although such repeated and independent observations cannot be carried out in practice, this conceptualization permits estimation of the precision of observed scores. This precision is referred to as *reliability*.

In CTT the reliability, or precision, of test data is defined as the squared correlation coefficient for observed and true scores, $\rho^2_{x\theta}$. In general discourse, $\rho^2_{x\theta}$ is referred to as the reliability of X. Because $\rho^2_{x\theta} = \sigma^2_\theta / \sigma^2_x$, reliability can be considered a measure of the amount of observed-score variance, σ^2_x, that is accounted for by true-score variance, σ^2_θ. Therefore, $\rho^2_{x\theta}$ expresses how precisely the observed score reflects the unknown true score.

Although this model has proven tremendously useful over the years, and continues to be used in many practical testing settings, CTT has several inherent shortcomings. Explaining these limitations, Embretsen and Herschberger (1999) write

1. True score applies only to items on a specific test or to items on a test with equivalent item properties.

2. Although the model specifies two separate independent variables for a person (person true score and person error), these independent variables are not really separable for an individual score.

3. Item properties are not linked to behavior. The omission of item properties from the model requires that they be justified elsewhere, such as by their impact on various group statistics such as variances and reliabilities.

CTT person ability estimates are *test-dependent*. On the other hand, item difficulty estimates are *sample-* or *group-dependent*. Additionally, CTT assumes that the standard error of measurement, which is a function of reliability and variance, is the same for all examinees of varying abilities. This assumption is untenable given that test scores are not equally precise measures for individuals with different abilities. Finally, CTT is test oriented rather than item oriented. CTT provides no basis on which to predict performance of an individual with a given ability on a given item. This limitation is significant given the many testing applications where the intent is to discriminate effectively between individuals in a particular ability group.

Citing the work of several researchers, Hambleton, Swaminathan, and Rogers (1991) note that, in addition to the limitations outlined above, CTT fails to provide satisfactory solutions for test design (Lord, 1980), detecting item bias (Lord, 1980), adaptive testing (Weiss, 1983), and test equating (Cook & Eignor, 1983, 1989). Given the limitations of CTT, Hambleton Swaminathan, and Rogers (1991) explain that psychometricians have attempted to develop alternative theories and models for mental measurement with the following features "(a) item characteristics that are *not* group-dependent, (b) scores describing examinee proficiency that are *not* test-dependent, (c) a model that is expressed at the item level rather than at the test level, (d) a model that does *not* require strictly parallel tests for assessing reliability, and (e) a model that provides a measure of precision for each ability score" (p. 5). The family of models that attempt to provide these desirable features are categorized under the umbrella of Item Response Theory.

Item Response Theory

Item Response Theory (IRT) involves the study of individuals' responses to the items in a test or questionnaire based on assumptions concerning the mathematical relationship between the ability or latent trait being measured and an individual's response. Hence, IRT is sometimes referred to as Latent Trait Theory. The family of IRT models includes the one-parameter logistic model (1-PL Model) in which the only item parameter modeled is the *difficulty* of the item, the two-parameter logistic model (2-PL) in which both item difficulty and item *discrimination* are modeled, and the three-parameter logistic model (3-PL) in which item difficulty, discrimination, and *pseudo-chance-level* parameters are estimated. These terms will be elaborated later in this chapter. The Rasch Model is mathematically equivalent to the 1-PL model, but was developed independently of the other IRT models by a Danish mathematician named Georg Rasch (1960). Before exploring the definitions of these parameters and the characteristics of IRT models, I will provide a brief summary of their historical development.

Origins of Item Response Theory models. Embretson and Reise (2000) identify two separate lines of development in IRT—one in the United States and one in Europe. In the United States during the 1950s Allan Birnbaum and Frederick Lord were developing IRT. Lord's (1953) "The Relation of Test Score to the Trait Underlying the Test" and Birnbaum's research, documented in three U.S. Air Force technical reports (Birnbaum, 1957, 1958a; 1958b), constituted the early formalization of IRT in the U.S. Birnbaum's writings were relatively unknown until 1968 when Lord and Novick (1968) published *Statistical Theories of Mental Test Scores* that contained much of Birnbaum's material.

Lord was employed at Educational Testing Service (ETS). An ongoing seminar at ETS included a number of psychometricians who, with Lord and Novick, later made significant contributions to psychometric methods. R. Darrell Bock, for example, explored the development of algorithms for estimating IRT parameters. Bock and several of his graduate students implemented their algorithms for IRT parameter estimation into computer programs such as BILOG, TESTFACT, MULTILOG, and PARSCALE. The marginal maximum likelihood method for estimating IRT parameters developed by Bock is now considered state of the art (Embretson & Reise, 2000).

Meanwhile, Georg Rasch was developing a psychometric model in Denmark that was applied to testing needs in the Danish military. Rasch's (1960) *Probabilistic Models for Some Intelligence and Attainment Tests* articulated this model and focused attention on important scientific properties of measurement models. For example, Embretsen and Reise (2000) write that Rasch "noted that person and item parameters were fully separable in his models, a property he elaborated as *specific objectivity*" (p. 6). Rasch's work was elaborated upon by Anderson (1970, 1972) and Fischer (1973, 1974) in Europe. Although influential among psychometricians in Europe, Fischer's work had limited exposure in the U.S. because it was written in German.

Inspired by a series of lectures by Georg Rasch at the University of Chicago, Benjamin Wright, a professor of education, saw considerable promise in Rasch's conception of *objective measurement*. Wright's interest resulted in a series of dissertations that have contributed significantly to Rasch model research. These studies include the work of David Andrich (1978), who developed the Rasch rating scale model; Geoffrey Masters (1982), who developed the Rasch partial-credit model; as well as

Graham Douglas (Wright & Douglas, 1977) and Mark Wilson (1989). Other significant contributors to the development and application of Rasch models in the U.S. include John M. Linacre, the author of WINSTEPS (Linacre, 2004b) and FACETS (Linacre, 2004a), computer applications that produce Rasch model parameter estimates and analyses; Richard M. Smith, the founding editor of both the *Journal for Outcome Measurement* and the *Journal of Applied Measurement*, journals that have provided a forum for the publication of theoretical and applied studies related to measurement in general and Rasch models in particular; and Trevor G. Bond and Christine M. Fox (2001) whose popular *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* will soon be released in a second edition.

Having briefly summarized the origins of IRT models, attention will now be turned to an examination of the basic principles of IRT and to the mathematical conceptualizations for several members of the IRT family. Understanding the similarities and differences between these models, as well as their underlying properties and assumptions, will allow the reader to better grasp the reasons for the choice of models applied in this study.

Basic concepts in Item Response Theory. As we begin to explore the characteristics, mathematical representations, and assumptions, of the various IRT models it is important to comment on the varied and sometimes confusing notation used in IRT literature. Person ability is represented as θ in most general IRT sources, but in Rasch model sources person ability is represented as *B*. This is potentially confusing because general IRT sources represent item difficulty with lowercase *b*, whereas Rasch sources represent item difficulty with D_i. This literature review uses both sets of notation to reflect the conventions followed in the respective bodies of research.

IRT models predict the performance of a person of a given ability to a test item of a given difficulty relating to the trait of interest. For example, a person with low ability will have a low probability of answering an item of average difficulty correctly, whereas a person with high ability will have a high probability of answering an item of average difficulty correctly. A person of average ability may have a .50 chance of answering the item correctly. IRT represents these probabilities as a regression line with person ability and item difficulty both represented on the *x*-axis on the same metric, and probability of success represented on the *y*-axis. The resulting regression line appears as an s-shaped curve or *ogive*. In IRT terms, this line is referred to as an *item characteristic curve* (ICC). A one-parameter model item characteristic curve (ICC) is displayed in Figure 1.

One-parameter logistic or Rasch model. The characteristics of the ICC are the parameters estimated in IRT models. In the 1-PL and Rasch models (hereafter referred to exclusively as the Rasch model) difficulty is the estimated parameter. The difficulty of an item is represented by the point along the *x*-axis at which a person of average ability would have a .50 probability of succeeding on that item. That point is shown in Figure 1 by the vertical arrowed line.

The ICC for the Rasch model is given by Equation 1:

$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i \qquad (1)$$

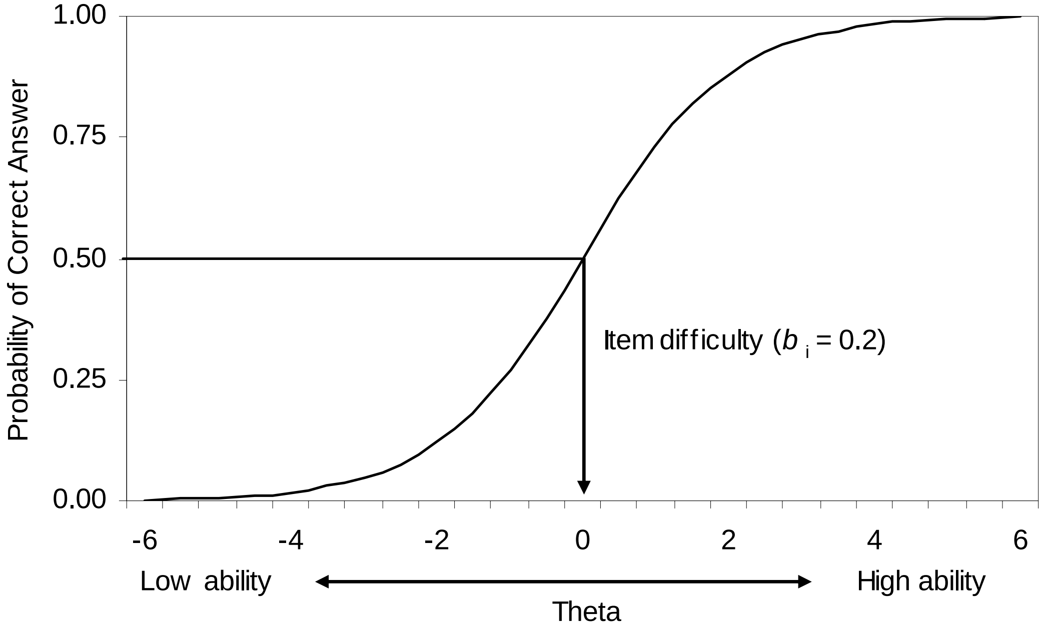


Figure 1. Item characteristic curve.

where P_{ni} is the probability of person *n* with ability B_n succeeding on item *i* which has difficulty level *D*. It is an s-shaped curve with values between 0 and 1 over the ability scale.

Figure 2 illustrates three Rasch model ICCs. The *x*-axis scale of the item ICC is expressed in *logits* in the Rasch model. A logit, or log odds unit, is the natural log transformation of the odds ratio. The odds ratio is the probability of answering correctly over the probability of answering incorrectly. Because of this conceptualization the resulting θ and *b* scales possess an equal-interval property. The 2-PL and 3-PL models must define the scale differently and in so doing fail to achieve the desired equal-interval scale critical to fundamental measurement. This argument will be further developed later in this chapter.

Two-parameter logistic model. In the 2-PL model both difficulty and discrimination are estimated. An item's discrimination parameter is represented by the slope of the ogive at the point of inflection. A steeper slope translates into increased discrimination, whereas a flatter slope means the item is less discriminating. The ICC for the 2-PL model (Figure 3) is given by Equation 2:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (2)$$

where

D is a scaling factor (1.70) introduced to make the logistic function as close as possible to the normal ogive function, and

a_{i} is the item discrimination parameter.

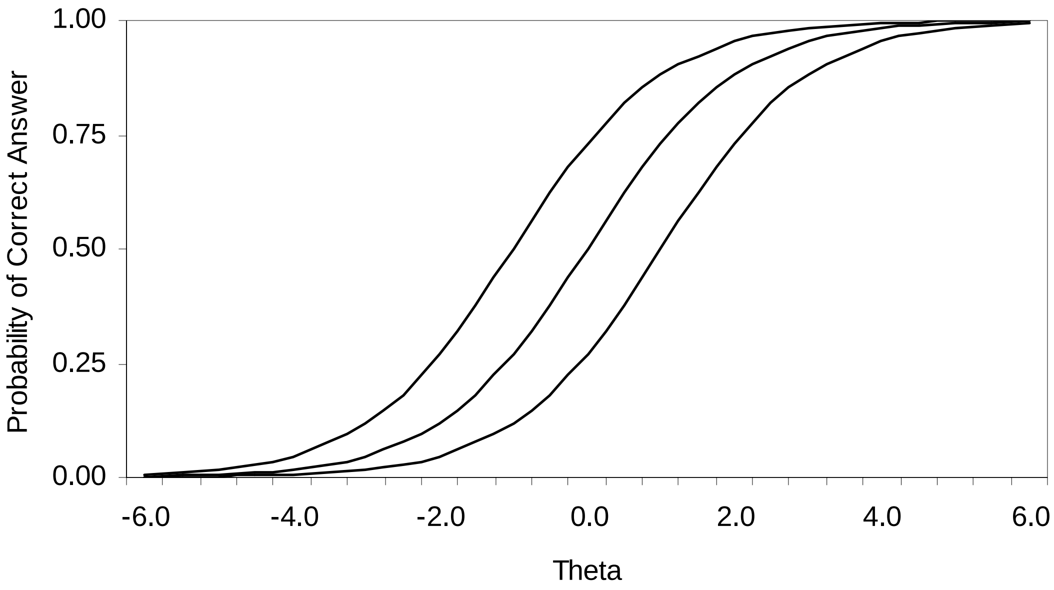


Figure 2. Three Rasch model ICCs with the same discrimination parameters but different difficulty parameters.

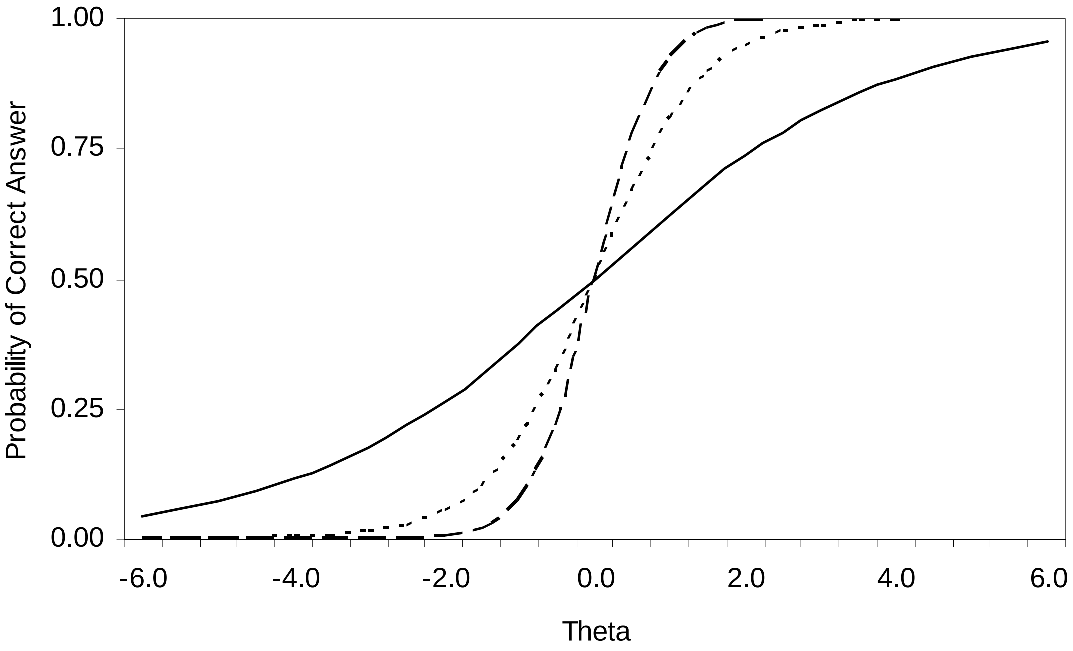


Figure 3. Three 2-PL model ICCs with the same difficulty parameters but different discrimination parameters.

Three-parameter logistic model. In the 3-PL model the difficulty, discrimination, and pseudo-chance-level parameters are estimated. The pseudo-chance-level parameter can be understood in the context of a multiple choice question. The value of this parameter describes the probability that a person with no ability will answer the item correctly by selecting the correct response exclusively by chance. The value of this parameter is the lower asymptote of the ogive (where the regression line intersects the *y*-axis when the *y*-axis is at negative infinity) in the 3-PL model ICC. The ICC for the 3-PL model (Figure 4) is given by Equation 3:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{3}$$

where

c_{i} is the pseudo-chance-level parameter.

Item Response Theory assumptions. Each of the IRT models is based on specific assumptions. Two assumptions are common to the 1-PL, 2-PL, and 3-PL IRT models: *unidimensionality* and *local independence*. Unidimensionality is the assumption that only one trait is being measured by a given set of items. Although this assumption can never be fully satisfied in a strict sense because of the many extraneous factors that are likely to influence examinees’ responses (e.g., differences in testwiseness, test anxiety, reading ability), it is possible to achieve essential unidimensionality which means that a single dominant trait accounts for most of the variance in a set of items. Local independence, on the other hand, is the assumption that the probability of correctly answering one item is not dependent upon correctly answering another item in the same test. To the degree

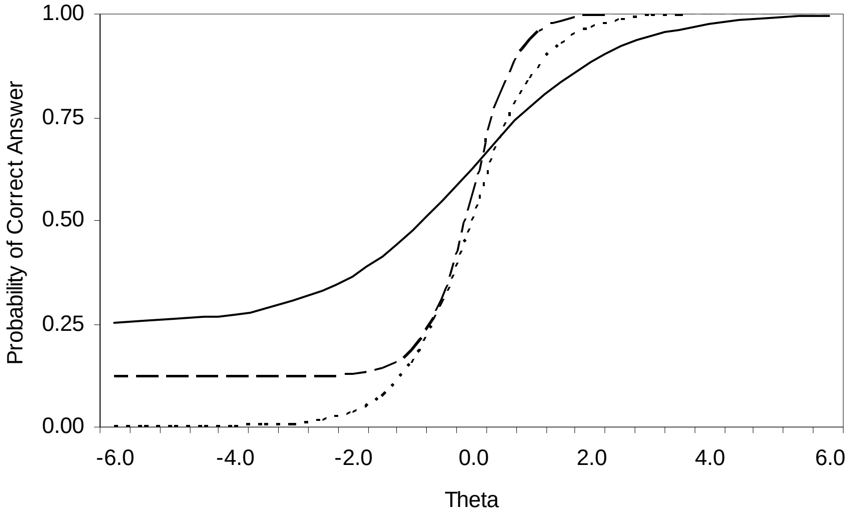


Figure 4. Three 3-PL model ICCs with the same difficulty parameters but different discrimination and pseudo-chance parameters.

that the model fits the data, then a number of desirable measurement qualities are achieved: person ability estimates are not test-dependent and item difficulty estimates are not sample-dependent. In other words, person ability estimates would not change significantly based on the administration of different items measuring the same trait, and item difficulty estimates would not change significantly if a different group of persons with a different range of ability responded to the items. This quality is referred to as *invariance* (Hambleton et al., 1991).

The concept of model fit is viewed differently by proponents of the 2-PL and 3-PL models than by the proponents of the Rasch model. Rasch model proponents seek to write items and scales that produce responses that fit the model because the model possesses the qualities of invariance and equal-interval scaling. On the other hand, proponents of the 2-PL and 3-PL proponents models seek to fit the model to the data. This difference in philosophy is attributable, in part, to the disciplines from which the originators of these models come. Proponents of the 2-PL and 3-PL models typically come from the social sciences in which emphasis is placed on modeling data, whereas the Rasch model grew out of a measurement philosophy based on efforts in the physical sciences to adhere to principles of fundamental measurement. By demonstrating adherence to the assumptions of unidimensionality and local independence, IRT models overcome many of the limitations of CTT.

Unfortunately, however, many psychometricians and quantitative social scientists have ignored measurement theory and have failed to address the assumptions of fundamental measurement. Stone (2004) explains

> We do not seek models to fit data. Rather we construct data to fit the model which is necessary in order to construct measures. Social science literature often speaks of a need to find a model to fit a particular collection of data. But the problem is, in fact, entirely opposite. The challenge is to produce data good enough to fit a measurement model. Rasch models provide the means by which to construct measures from carefully produced data and to monitor the process to reach a best possible solution. (p. 215)

Fundamental Measurement

Sufficient care has not been taken in the social sciences to evaluate the assumptions that must be met in order for scores to qualify for mathematical manipulation. Bond and Fox (2001) argue that “quantitative researchers in the human sciences are too narrowly focused on statistical analysis, and not concerned nearly enough about the quality of the measures on which they use these statistics” (p. 1). Too often researchers assume that the mere assignment of numerical values to objects suffices for the purposes of measurement. The practice of summing numbers obtained by performance on tests, treating the total as if it were a measure, and applying statistical analyses to those scores is indefensible. Treating ordinal data as if it were interval data ignores the necessity of an additive structure for mathematical manipulation (Michell, 1997). In addition to *additivity*, fundamental measurement requires *objectivity*. Objectivity posits that the measurement ascribed to a person be independent of the observer. Quoting Karabatsos (1999), Bond and Fox (2001) note

> The properties of extensive measurement are the logical rules underlying explicit physical measurement. Obviously, in situations where measurements are not

> directly observable, such as achievement or intelligence, the structures are not explicit. *But this does not mean that structure cannot be used to verify the measurement of latent traits.* The theory of additive conjoint measurement makes fundamental measurement a possibility in the social sciences, where all observations are ordinal. Within the psychometrics framework, this theory proves that when the axioms of independence and double cancellation are satisfied, then the numbers assigned to persons and items represent a common linear (interval) style, measure on a single dimension. Furthermore, person measurement is independent of the items used, and item calibrations are independent of which persons they measure. (p. 195)

As noted earlier in this chapter, of the IRT models only the Rasch model possesses the attributes of conjoint additivity necessary for fundamental measurement. Continuing their quotation of Karabatsos (1999), Bond and Fox (2001) add

> There is strong support that almost 100% of the time, the parameters of the 2 PL and 3 PL violate interval scaling. On the other hand, the theoretical probabilities of Rasch models will always support a stable, interval scale structure. If the intention is to construct stable interval measurement, data should approximate uniform item-characteristic curves. The argument that 2 PL and 3 PL are advantageous because they are less restrictive alternatives to Rasch models (Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997) does not recognize the connections between linear measurement requirements and uniform ICCs. . . . This study is not needed to prove that Rasch models satisfy conjoint

> measurement axioms. However, it seems that the field of item response theory does not recognize the mathematical requirements of measurement. (p. 196)

In summary, a brief history of CTT and IRT models was presented. The inherent limitations of CTT were detailed and the resulting efforts to overcome these limitations by IRT were explained. The essential characteristics of additivity and objectivity to fundamental measurement were noted and the Rasch model was identified as the only IRT model that meets the requirements of fundamental measurement. Therefore, in addition to the application of CTT approaches to scale development and assessment, the Rasch model will be employed in the construction, refinement, and assessment of rating scales designed to measure teacher performance.

The mathematical models detailed in this chapter have thus far been expressed for dichotomous items (answers are either right or wrong). In this study the rating scale items are polytomous in nature (each item has more than two response categories). Therefore, the Rasch rating scale model (Andrich, 1978; Wright & Masters, 1982) will be employed. This model will be elaborated in more detail in the following section where item analysis is addressed.

Summated Rating Scales

Definition

Many scholars attribute the origins of summated rating scales to Rensis Likert (1932) who used this approach to assess attitudes. Hence, summated rating scales are frequently referred to as Likert scales. However, while all Likert scales are usually considered summated rating scales, the reverse is not true. For example, a semantic differential scale is a summated rating scale, but it is not a Likert scale. In the context of this study, the scales are considered both summated rating scales and Likert scales, but will simply be referred to as scales. Scales have the following characteristics:

1. A scale must contain multiple items that each produces responses with a numerical value that will be summed or averaged to create a single score describing a respondent's location on the underlying trait continuum.
2. Each item has no "right" or "wrong" answer as does a multiple-choice test.
3. Each item requires the respondent to respond to a statement or group of words (e.g., semantic differentials consisting of pairs of bipolar adjectives as items).

Scales have been used to measure hundreds of different variables (e.g., anxiety, autonomy, locus of control, self-efficacy, workload). The development and use of scales to measure variables of interest can be time consuming and costly. However, some variables, particularly in the social sciences, cannot be studied meaningfully in any other way. Scales are used because a single item cannot provide reliable information about complex variables. A well-functioning scale provides reliable data from which valid conclusions may be drawn. In order to promote reliable data and valid conclusions, an iterative process is followed in which (a) the purposes of the assessment is defined, (b) the construct or target variable to be assessed is clearly defined, (c) items and response formats are written and reviewed, (d) the scale is administered and response patterns are analyzed, and (e) variable definition and scale design are refined based on an iterative process of tryout and revision based on item analysis and other psychometric checks (DeVellis, 2003; Netemeyer, Bearden, & Sharma, 2003; Spector, 1992; Wright & Masters, 1982).

Development Methodology

Identify the intended purposes of the scale. The value of any scale is dependent on the degree to which the purposes of the assessment are fulfilled. Any assessment effort should begin, therefore, with a careful examination and articulation of the purposes of the scale. Content experts should review and approve the types of items, response formats, scoring procedures, and test administration procedures to ensure that each is based on the "purposes of the test, the domain to be measured, and the intended test takers" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Define the variable. Essential to the construction of meaningful scales is the careful definition of the theoretical abstraction that is to be measured. Typically a first step in the definition of a variable is a literature review in which special attention is given to "specific details of exactly what the construct has been described to be" (Spector, 1992). Where little empirical research has been done relative to a particular variable it is to be expected that the variable and the scale will evolve together. As the variable is clarified and refined, attention must be given to the dimensionality of the trait of interest. In the discussion of IRT, the assumption of unidimensionality was addressed. DeVellis (2003) reminds scale developers that in many cases similar items get at very different variables. By clearly defining a unidimensional variable the stage is set to generate an item pool that is sufficiently homogenous to provide meaningful data about the underlying variable.

Design the scale. Once the variable is defined, individual items are written that reflect the latent variable. In a sense, each item is a test of the variable that provides unique information. Somewhat paradoxically, the intent is to write items that are similar but different. The items are intended to be similar in that each item reflects the targeted variable, but different in that each item contributes unique information about some specific facet of the variable. In the initial stages of scale development it is wise to generate a large item pool. A large pool facilitates a degree of redundancy that will allow for a reduction in items based on empirical analysis of responses to items.

Rating scale items consist of two parts, the *stem* and the *response categories* or *anchors* to the stem. Numerous scholars (DeVellis, 2003; Dillman, 2000; Edwards, 1957; Fink, 1995; Kerlinger, 1986; Likert, 1932; Payne, 1951; Spector, 1992; Thomas, 2004; Warwick & Lininger, 1975) have identified guidelines for writing stems. A number of common guidelines are captured by Stone (2004), who recommends that scale developers

1. Avoid factual statements.
2. Do not mix past and present. Present is preferred.
3. Avoid ambiguity.
4. Do not ask questions that everyone will endorse.
5. Keep wording clear and simple.
6. Keep statements short and similar in length.
7. Express only one concept in each item.
8. Avoid compound sentences.
9. Assure that reading difficulty is appropriate.
10. Do not use double negatives.
11. Do not use *and* or *or* or lists of instances.

Additionally, the use of negatively-oriented items is recommended by some sources in order to avoid response sets. DeVellis notes, however, that there may be a price to pay for mixing positively and negatively worded items. He cites a number of instances in which negatively-oriented items performed very poorly (Currey, Callahan, & DeVellis, 2002; DeVellis & Callahan, 1993). When these items were rewritten with the word "not," however, the change in valence resulted in dramatically improved item performance. DeVellis' findings are supported by Yamaguchi (1997) and Linacre (2004a), who demonstrate that the rating scale categories function differently for negatively-oriented items. Therefore, negatively-stated items should be analyzed separately.

The response options for an item should be directly relevant to the idea asserted in the stem of the item. The number of response categories should not exceed the levels of gradation between which respondents are able to meaningfully distinguish. Frequently scale developers include many response options believing that the more options the more information will be obtained. In practice, however, this is seldom achieved; only a few of the response categories are actually used. Therefore, it is important to anticipate the way in which respondents will perceive the proposed response categories and only provide those categories that respondents will meaningfully differentiate. Stone (2004) suggests that four options are usually enough, but strongly recommends that researchers analyze their own items and response patterns to make judgments about the useful range of responses. The categories must define, in a clear and unambiguous manner, gradations that are meaningful to each respondent. Additionally, the terminal points of the scale categories should be reasonable. For example, *never* and *always* may not be reasonable terminal points for many scale items. The application of these guidelines provides a useful foundation on which to build scale response options that will function as intended. Wright and Stone (2003) and Linacre (2004a) articulate specific guidelines for evaluating the degree to which response options function as intended. These will be elaborated in the discussion of item analysis.

Finally, it is critical that the generated items and response categories be reviewed and validated as reasonable indicators of the trait of interest by individuals who are qualified as content experts by virtue of their relevant training and experience. In a sense, this is an effort to ensure quality control. In the process of translating variable definitions into items it is possible to neglect or over-represent some aspects of the latent trait. By obtaining expert opinion on the alignment of items and response categories with the target variable, a necessary step is taken towards building a defensible validity argument. This portion of the validity argument is generally referred to in the literature as *content validity*. Content validity is concerned with the selection of items from a domain of potential items representing a given variable. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) state

> The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. (p. 44)

Pilot test the scale. Once the stems and response categories have been carefully crafted and validated, the next step is to test the items with respondents. This step requires consideration of sample characteristics. The sample used in the pilot test should be representative of the population in which the scale will be used. The size of the sample is dictated by the desired stability of person ability and item difficulty estimates. Table 1 identifies the sample sizes necessary to obtain specific confidence intervals for item calibrations using the Rasch model (Linacre, 1994). Additionally, if any kind of confirmatory factor analysis is planned as a means of obtaining evidence of construct validity, then sample size must be considered for this purpose also. For confirmatory factor analysis five times as many subjects as items is recommended (Bryant & Yarnold, 1995; Netemeyer et al., 2003).

Table 1

Sample Size for Stable Rasch Model Item Calibrations

Item Calibrations stable within	Confidence	Minimum sample size range (best to poor targeting)	Size for most purposes
± 1 logit	95%	16 – 36	30
± 1 logit	99%	27 – 61	50
± ½ logit	95%	64 – 144	100
± ½ logit	99%	108 – 243	150

With empirical data generated by the pilot administration, the researcher is prepared to begin to assess the psychometric properties of the scales. As mentioned previously a combination of classical and IRT approaches will be employed in this study.

Exploratory factor analyses. At this point in the scale development process, the purposes of exploratory factor analysis (EFA) are to reduce the number of items in each scale while maximizing explained variance and reliability; and assess the dimensionality of data produced by the scales. Factor analysis is typically used for this purpose. Factor analysis seeks to identify the fewest number of factors that account most economically for the variance in the response data. This is accomplished by an analysis of the item correlation matrix. Items that contribute little to explaining the variance in the data are targeted for deletion.

There is no definitive answer to the question of how many factors are supported by a given data set (Cudeck, 2000). The objective of EFA is to develop a model that makes sense and that reasonably describes the data. Several criteria are commonly used to determine how many factors are supported by the data. These criteria include (a) the Kaiser-Guttman criterion; (b) the scree test criterion; (c) the factor loading criterion; (d) the variance explained criterion; and (e) the a priori factor extraction criterion.

The Kaiser-Guttman criterion, also known as the Latent Root criterion, recommends that to be meaningful a factor must have an eigenvalue greater than 1 (Guttman, 1954; Kaiser, 1960). This is an absolute criterion as opposed to the relative criterion associated with the use of a scree test.

A scree test involves graphically plotting factors on the *x*-axis and their respective eigenvalues on the *y*-axis. The investigator examines the relative magnitude of each factor's eigenvalue. A line is created by connecting the plotted values. The scree test criterion involves identifying the point in the plotted line at which a break in magnitude occurs (Cattel, 1966). Factors with eigenvalues that are relatively larger than all others are retained, while those that are relatively smaller are excluded.

The factor loading criterion involves examining the magnitude of the loadings of individual items on a given factor. Items that load on a factor with a value greater than .40 are considered substantial (Netemeyer et al., 2003). Three or more items exhibiting substantial loadings on a factor is considered evidence of an independent factor.

The variance-explained criterion recommends that the number of factors extracted should account for 50-60% of the variance in the data and each individual factor should account for at least 5% of total variance to be meaningful (Hair, Anderson, Tatham, & Black, 1998; Netemeyer et al., 2003).

The a priori factor extraction criterion recommends that the number of factors to be extracted are those theorized by the developer of the scales prior to analysis. This approach allows the developer to identify cross-loadings of items that would identify those items as candidates for deletion (Floyd & Widaman, 1995; Hair et al., 1998).

In summary, the retention of scale items should reflect factors with eigenvalues greater than 1.0, factors with eigenvalues of a magnitude greater than other factors on the scree plot, factors that have at least three items with factor loadings greater than .40 (assuming a sample size of at least 5-10 times the number of respondents as items), and factors that align with the theoretical understanding of the underlying constructs that do not load substantially on multiple factors.

Classical reliability and item statistics. In addition to EFA several rules of thumb are commonly used to assess the reliability of scale data. These include (a) Cronbach's coefficient alpha greater than .80 (Bearden & Netemeyer, 1998; Clark & Watson, 1995), (b) inter-item correlations greater than .30 (Netemeyer et al., 2003; Robinson, Shaver, & Wrightsman, 1991), and (c) corrected item-total correlations greater than .50 (Netemeyer, Boles, & McMurrian, 1996; Obermiller & Spangenberg, 1998; Tian, Bearden, & Hunter, 2001). These estimates can be easily calculated using SPSS.

Confirmatory factor analyses. Confirmatory factor analysis (CFA) has become a common step in confirming factor structure and finalizing scale development. CFA is most commonly conducted using structural equation modeling software such as AMOS, EQS, LISREL, and SAS. In CFA an a priori factor structure is specified and tested. The null hypothesis—that the data fit the specified model—is tested using a chi-square statistic. Because this statistic is particularly sensitive to sample size, a variety of alternative goodness-of-fit indices are also calculated. Non-rejection of the null hypothesis is desired. Additionally, each goodness-of-fit statistic has a range of acceptable values. For example, some researchers consider an estimate of the root-mean-square-error-of-approximation (RMSEA) of .08 or less evidence of acceptable fit (Browne & Cudeck, 1993), whereas others recommend a more rigorous criterion of .06 or less (Hu & Bentler, 1999). Other indices, such as the comparative fit index (CFI), have a recommended acceptable fit value of .95 (Hu & Bentler, 1999).

The results of CFA, like EFA, provide evidence of unidimensionality and construct validity. They also point the scale developer to potential problems such as multidimensionality and violations of local item dependence that can guide the developer in the revision and improvement of items and scales.

Rasch model analyses. The rating scale model is an extension of the dichotomous model that reflects the probability of endorsing a particular response category rather than the probability of providing the correct answers. The dichotomous model presented in Equation 1 is extended and expressed as the rating scale model in Equation 4:

$$\ln\left(\frac{P_{nik}}{1-P_{ni(k-1)}}\right) = B_n - D_i - F_k \qquad (4)$$

where

P_{nik} is the probability that person *n*, on encountering item *i* would be observed in category *k*,

$P_{ni(k-1)}$ is the probability that the observation would be in category $k-1$,

B_n is the ability of person *n*,

D_i is the difficulty of item *i*,

F_l is the impediment to being observed in category *k* relative to category $k-1$.

The probabilities associated with the selection of a particular response category may be represented graphically in terms of *category probability curves* (Figure 5). Persons with the lowest ability are most likely to select category 1. By ability I mean the location of the person along the trait's continuum. Use of the word ability, however, does not make sense in many rating situations even though it is frequently used as a general term to describe the location of persons along the continuum. For example, the trait under

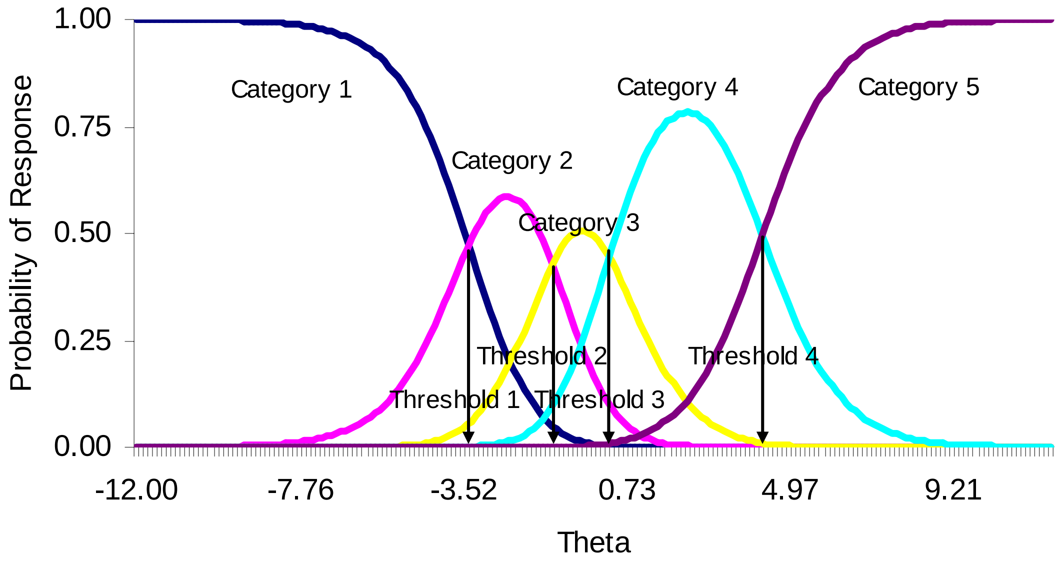


Figure 5. Category probability curves.

examination may not be an ability, but rather an attitude or disposition. The rater's response, represented by their choice of response categories, reflects their *endorsement* of a given item and therefore their attitude, disposition, or ability relative to the trait. As ability increases (moves to the right along the *x*-axis), the probability of selecting category 1 decreases and the probability of selecting category 2 increases. The location along the *x*-axis, where the category probability curves intersect, is referred to as a *threshold* or *step*. In Figure 5, where five response categories are represented, there are four thresholds. An analysis of the category probability curves allows the researcher to evaluate whether or not the response categories function as intended. Before response category function is examined in more detail, however, a step-by-step approach to the analysis of rating scale data will be outlined.

In a Rasch model item analysis, the response data are tested for their fit to the model to determine "what is measurable, decide which data are useful, and expose which data are not" (Wright, 1999). R. M. Smith (1998) proposes a best practice model for assessing the dimensionality and fit of rating scale data. R. M. Smith's guidelines will be noted and elaborated.

1. Examine the person separation reliability estimates. Because person ability estimates are on a linear scale they are suitable for the calculations of means and variances. Direct estimates of the modeled error variance for each estimate of a person's ability are provided by the Rasch model in the form of a standard error (SE). The SE quantifies the precision of every person measure. E. V. Smith (2004) explains that these individual SEs can be squared and summed to produce an average error variance for the sample that can be used in internal consistency reliability formulas (Equation 5). The

group estimate of reliability for persons, or person separation reliability, is obtained by subtracting the average person measurement error variance, MSE$_p$, defined in Equation 6

$$MSE_p = \sum_{n=1}^{N} S_n^2 / N \tag{5}$$

where S_n is the SE for each person measure from the observed variance among person, SD^2:

$$SA_p^2 = SD_p^2 - MSE_p \,. \tag{6}$$

Therefore, the person variance is adjusted for measurement error and represents the *true* variance in the person estimates. Person separation reliability is expressed as the ratio of adjusted (true) variance to observed variance and can be interpreted as the proportion of variance that is not due to error (Equation 7):

$$R_p = SA_p^2 / SD_p^2 = 1 - \frac{MSE_p}{SD_p^2} \tag{7}$$

In Linacre's WINSTEPS application this estimate is labeled "person reliability." Another estimate of reliability used by Rasch practitioners is the person separation index which is labeled "person separation" in WINSTEPS. Unlike the person reliability estimate that has a maximum value of 1.00, the person separation index, G_p, is not constrained by an upper boundary, but has a range of zero to infinity. Where a significant increase in reliability would not be evident in the person reliability estimate (e.g., an increase from 0.97 to 0.98) because of ceiling effect, it *would* be evident in the person separation estimate (e.g., an increase from 5.5 to 7.0). Equation 8 expresses G_p:

$$G_p = \frac{SA_p}{SE_p} = \sqrt{\frac{R_p}{1 - R_p}} \qquad (8)$$

where SE_p is the Root Mean Square Measurement Error (RMSE) and is equal to $(MSE_p)^{1/2}$. Larger R_p and the G_p values reflect a greater spread of persons along the variable being measured, a generally desirable characteristic of psychometric instruments.

Additionally, E. V. Smith (2004) suggests a transformation of R_p and G_p that results in another reliability index called *strata*. Based on the work of Wright and Masters (1982) and Fisher (1992), strata are defined as the number of statistically distinct levels of person ability that the items have distinguished. Levels are considered distinct if they are separated by at least three errors of measurement. Strata are calculated using Equation 9:

$$Strata = \frac{4G_p + 1}{3}. \qquad (9)$$

The relationships between these approaches to person separation reliability estimation are summarized in Table 2 from E. V. Smith (2004). R. M. Smith (1998) suggests that person reliability estimates are usually greater than .70. If they do not exceed .70, he suggests examination of person and item statistics to determine why the person reliability estimates are suspect.

2. Examine the item separation reliability estimate, which, like the person separation estimate, should be greater than .70. If it is not greater than .70 the reasons underlying this deficiency are examined. Item estimates are derived from the equations outlined in the previous section on person separation reliability by the replacement of the subscripts l, *L*, and *i* for n, *N*, and *p*.

3. Examine the person fit statistics. The fit statistics generated in the item analysis provide an empirical basis for assessing the underlying assumptions of unidimensionality and local independence that must accrue for invariant, equal-interval scaling. Rasch (1960) proposed a variety of statistical and graphical methods for analyzing fit. In the absence of computerized methods for calculating fit indices, however, these were not widely used. Wright and Panchapakesan (1969) developed an overall chi-square statistic based on Pearsonian chi-square that was used with the first computer programs that produced Rasch model item calibrations. Wright and Panchapakesan later demonstrated the use of this statistic summarized on an item by item basis (R. M. Smith, 2004). Providing a comprehensive review of item fit statistics, R. M. Smith (1991) demonstrates how the chi-square fit statistics may be transformed to simplify interpretation. This is accomplished by dividing the chi-square by its degrees of freedom resulting in a mean square with an expected value of 1 and a range of 0 and $+\infty$.

Table 2

Relationship Among Person Reliability, Person Separation, Ratio of True and Error Variance, and Strata

Person Separation	Person Reliability	% variance: Not Due to Error/ Due to Error	Strata
0.0	.00	0/100	1
0.5	.20	20/80	1
0.1	.50	50/50	1
1.5	.70	70/30	2
2.0	.80	80/20	3
2.5	.86	86/14	4
3.0	.90	90/10	4
3.5	.92	92/8	5
4.0	.94	94/6	5
4.5	.95	95/5	6
5.0	.96	96/4	7
5.5	.97	97/3	7
6.0	.97	97/3	8
6.5	.98	98/2	9
7.0	.98	98/2	9

The mean square is then transformed, usually with a cube-root, converting the mean square to an approximate *t*-statistic commonly known as a standardized fit index. This value is labeled ZSTD in the WINSTEPS software used in this study and will be referred to as the standardized *z* in this paper. If the data fit the model perfectly, the standardized *z* has a mean of 0 and a standard deviation of 1.

At what point does an item or person fail to fit the model? Some researchers define misfit in terms of a mean square value less than 0.6 or greater than 1.4, while others define misfit in terms of a standardized *z* greater than 2.0. Linacre (2002) conceptualizes the impact of misfit on measurement in Tables 3 and 4. The person fit statistics to be examined include both unweighted (outfit) and weighted (infit) mean square and standardized *z* values. The outfit statistic is based on the sum of squared standardized residuals. For every person, each standardized residual cell is squared and the string of those squared residuals, one for each and every item encountered by person, is summed and its mean found by dividing by the number of items to which the person responded as expressed in Equation 10:

$$outfit = \frac{\Sigma Z_i^2}{N} \,. \qquad (10)$$

Infit, on the other hand, is an information-weighted sum. Rasch statistical information lies in its variance, or the standard deviation (SD) of the observations squared (SD^2). Variances are larger for well-targeted observations and smaller for extreme observations. infit, therefore seeks to weight these variances accordingly. This is achieved by calculating the squared standardized residual value in the response string weighted by its

Table 3

Rasch Model Mean Square Fit Implications

Mean Square	Implication for Measurement
> 2.0	Distorts or degrades the measurement system. May be caused by only one or two observations.
1.5 – 2.0	Unproductive for construction of measurement, but not degrading.
0.5 – 1.5	Productive for measurement.
< 0.5	Less productive for measurement, but not degrading. May produce misleadingly high reliability and separation coefficients.

Table 4

Rasch Model Standardized Z Fit Implications

Standardized *Z*	Implication for Measurement
≥ 3	Data very unexpected if they fit the model (perfectly), so they probably do not. But, with large sample size, substantive misfit may be small.
2.0 - 2.9	Data noticeably unpredictable.
-1.9 - 1.9	Data have reasonable predictability.
≤ -2	Data are too predictable. Other "dimensions" may be constraining the response patterns.

variance and summed as in Equation 11:

$$infit = \frac{\Sigma Z_{ni}^{2} W_{ni}}{\Sigma W_{ni}} \quad . \tag{11}$$

The resulting infit statistic has the same distribution as the outfit statistic but is differentially weighted. Outfit emphasizes unexpected responses far from a person’s or item’s calibration, whereas infit emphasizes unexpected responses near a person’s or item’s calibration. Furthermore, negative standardized *z* values indicate less variation than modeled, whereas positive values indicate more variation than expected (Bond & Fox, 2001). Persons whose measures show significant misfit are generally excluded from the analysis generating item calibrations as long as the misfit is not systematic for a particular subpopulation of respondents.

4. Examine the item fit statistics. This includes the means and standard deviations of the item infit and outfit statistics, and item point-measure correlations. These statistics are arrived at in the same fashion as person fit statistics, but use item values in the calculations. The guidelines for assessing fit are the same for item fit as for person fit.

5. Examine unusual subpopulation differences (e.g., sex, ethnicity). This may be accomplished by comparing item calibrations generated by the subpopulations of interest. If a statistically significant difference in item calibrations by group can be demonstrated, then the items are working differently for the groups in question. This is evidence of

differential item functioning (DIF). In this case additional dimensions may be present in the data that threaten the assumption of unidimensionality and must be addressed.

6. Examine the threshold structure. Linacre (2004a) suggests several guidelines for analyzing threshold structure. Preliminarily, it is critical to investigate whether or not all items are oriented with the latent variable. Typically, items on the same scale employ the same response categories. However, in some instances items will have unique response options or reversed polarity (i.e., negatively oriented items that will be reverse scored). Before item analyses are conducted it is helpful to confirm the polarity of all items constituting a scale using point-measure correlations. For example, an item that should have been reverse scored, but was not, would likely show a negative point-measure correlation and can then be correctly scored to form an item orientation consensus. Once item orientation has been checked, the following guidelines are proposed:

   a. At least 10 observations of each category. If category observance is low, the resulting threshold calibration will be imprecisely estimated. This imprecision may result in instability. Therefore, at least 10 observations of each category are required for stable threshold estimates.

   b. Regular observation distribution. Optimal threshold calibrations are obtained with a uniform distribution of observations across all categories.

   c. Average measures advance monotonically with category. Remember that a person or item measure (calibration) is obtained by combining person ability and item difficulty $(B_n - D_i)$. It is essential that higher measure combinations

$(B_n - D_i)$ produce observations in the higher response categories and vice-versa.

d. Outfit mean square less than 2.0. If an outfit mean square value is greater than 2.0 then the unexplained noise in the data is greater than the explained noise. In other words, there is more misinformation than information.

e. Threshold calibrations advance. Increasing amounts of the underlying variable should result in increased probabilities of respondents producing responses in the higher categories of the scale. This applies to the scale as a whole and to individual categories. A violation of this guideline would be reflected in disordered category probability curves. For example, the probability of scoring a 2 is left of the probability of scoring a 1 on the theta continuum.

f. Ratings imply measures, and measures imply ratings. For example, an average expected score of 3 will correspond to a measure of 3, and a measure of 3 will correspond to an average expected score of 3. The *coherence* of average expected score and measure may be assessed using the *expected item score ogive*, which shows the relationship between average expected ratings (*y*-axis) and measures (*x*-axis). A measure of coherence is also provided in the WINSTEPS table detailing category structure statistics.

g. Threshold measures advance by at least 1.4 logits.

h. Threshold measures advance by less than 5.0 logits.

Linacre (2004a) summarizes the pertinence of each of these guidelines to the measurement process in Table 5.

Table 5

Summary of Threshold Guideline Pertinence

	Guideline	Measure Stability	Measure Accuracy (Fit)	Description of this sample	Inference for next sample
Pre.	Scale oriented with latent variable.	Essential	Essential	Essential	Essential
1	At least 10 observations of each category.	Essential	Helpful		Helpful
2	Regular observation distribution.	Helpful			Helpful
3	Average measures advance monotonically with category.	Helpful	Essential	Essential	Essential
4	outfit mean square less than 2.0.	Helpful	Essential	Helpful	Helpful
5	Step calibrations advance.			Helpful	
6	Ratings imply measures, and measures imply ratings.		Helpful		Helpful
7	Step difficulties advance by at least 1.4 logits.				Helpful
8	Step difficulties advance by less than 5.0 logits.	Helpful			

7. Check the variable definition by looking at the content of the items. Make sure that the implied difficulty approximates your understanding of what is being measured. This may be achieved by plotting persons and items along a vertical theta (ability/difficulty) continuum with persons on the left and items on the right. Referred to as a variable map, this graphic permits assessment of the alignment of persons and items; mean person ability and mean item difficulty should be relatively close to each other and items difficulties should span the range of person abilities. See Figure 6 for an example of a variable map. Notice in Figure 6 that the "M" on each side of the vertical continuum represents the mean person and item calibrations. "S" represents one standard deviation from the mean and "T" represents two standard deviations from the mean. Additionally, items that are theoretically more difficult should be higher on the vertical continuum than items that are less difficult. An additional graphical approach to assessing this guideline in the context of rating scale data is the general keyform. In Figure 7 the general keyform plots response category probabilities along the theta continuum for each item with the person distribution beneath the theta continuum. Notice in Figure 7 the mean (M) and standard deviations (S and T) for person measures represented below the *x*-axis.

8. Examine the principal component analysis (PCA) of the standardized residuals. This method extracts the first component, the item difficulty, and looks at the standardized residuals in the analysis. A limitation of PCA on raw scores lies in use of ordinal data with a parametric technique (PCA) that assumes at least an interval scale of measurement. Therefore, the PCA is conducted on the standardized residuals which possess the interval attribute. This analysis identifies characteristics shared in common among items. These common characteristics represent possible sub-dimensions within the

```
MEASURE                                 |                               MEASURE
  <more> --------------------- KIDS    -+- ACTS    --------------------- <rare>
    5                                X  +                                   5
                                     X  |
                                        |
                                        |
                                        |
                                        |
    4                                X  +                                   4
                                        |
                                     X  |
                                        |
                                       T|
                                     X  |
    3                                X  +                                   3
                                        |T
                                     X  |
                                    XX  |  X
                                    XX  |
                                    XX S|  X
    2                              XXX  +                                   2
                                    XX  |  X
                                    XX  |  XX
                                    XX  |
                                    XX  |S
                                XXXXXX  |  X
    1                              XXX  +                                   1
                                   XXX M|
                                 XXXXX  |  XX
                                   XXX  |  XXX
                           XXXXXXXXXXX  |  X
                                  XXXX  |  XX
    0                              XXX  +M                                  0
                                 XXXXX  |
                                       S|  XX
                                     X  |  X
                                    XX  |  X
                                        |  X
   -1                               XX  +  X                               -1
                                        |
                                    XX  |S X
                                       T|  X
                                     X  |
                                        |
   -2                                   +  X                               -2
                                        |
                                        |
                                        |  X
                                        |
                                        |T
   -3                                   +                                  -3
                                        |  X
                                        |
                                        |
                                        |
                                        |
   -4                                   +                                  -4
                                        |
                                        |
                                        |
                                        |
                                        |
   -5                                   +                                  -5
  <less> --------------------- KIDS    -+- ACTS    ------------------<frequent>
```

Figure 6. Variable map.

```
EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT) (BY OBSERVED CATEGORY)
-6       -4        -2         0         2         4         6
|---------+---------+---------+---------+---------+---------|  NUM
0                               0   :     1     :   2       2    5
|                                                           |
0                              0   :     1     :   2        2   23
|                                                           |
0                            0   :     1     :    2         2   20
0                           0    :     1     :   2          2    4
0                           0   :     1     :    2          2    8
|                                                           |
|                                                           |
0                        0    :     1    :    2             2    7
|                                                           |
0                      0    :     1    :    2               2    9
0                      0   :     1     :   2                2   16
0                     0    :     1     :   2                2   25
0                     0   :     1     :   2                 2    3
0                     0   :     1     :   2                 2   14
0                    0    :     1    :    2                 2    6
0                   0    :     1     :   2                  2   17
0                   0   :     1     :    2                  2   22
|                                                           |
0                 0    :    1     :    2                    2   24
0                 0   :     1     :   2                     2    1
0                0    :     1     :   2                     2   15
|                                                           |
0               0    :    1     :    2                      2    2
0              0    :     1     :   2                       2   21
0              0   :     1     :    2                       2   11
|                                                           |
0            0    :     1    :    2                         2   13
0           0    :     1     :   2                          2   10
|                                                           |
|                                                           |
0        0    :     1     :   2                             2   12
|                                                           |
0      0    :     1     :   2                               2   19
|                                                           |
|                                                           |
0   0   :     1     :    2                                  2   18
|---------+---------+---------+---------+---------+---------|  NUM
-6       -4        -2         0         2         4         6

                               1
                     12 2 2117083563422342 111 1 1    1     1  PER
                     T      S      M      S     T
```

Figure 7. General keyform.

data and should be diagnosed and assessed for their potential threat to unidimensionality. PCA of standardized residuals provides an additional diagnostic tool to EFA and to Rasch item analysis discussed earlier in the diagnosis of multidimensionality.

In this section the theoretical and operational basis for a Rasch rating scale model item analysis have been summarized. The guidelines outlined here form the foundation for the analyses conducted in this study. Ultimately, item analysis allows the researcher to discern what the data have to say about the substantive theory being investigated and the theory tells us something about the persons and items under investigation in an ongoing dialectical process (Bond & Fox, 2001). This process is reflected in the next step of revision and re-administration.

Revise and re-administer. The item analysis informs the revision of items and scales in a variety of ways. Where persons or items misfit, the researcher may choose to simply delete them from the analysis. For example, persons who respond to the rating scale indiscriminately without reading the items will likely demonstrate considerable misfit. Their responses contribute little information regarding the latent variable and their responses should be eliminated from the analysis, usually resulting in improved item fit statistics.

Similarly, an item may misfit because it is capturing a variable very different from the other items in the scale. The item fit statistics allow the researcher to identify a problem item and then to theorize as to why it is working differently than the other items. Although great care may be taken in the variable definition and item writing process, it is not unusual that the way in which an item is written results in poor fit. The reasons are usually relatively easy to discern once the item has been identified as misfitting. Once misfitting items have been eliminated it is not unusual for the person and item reliability estimates to improve. However, if a scale consists of relatively few items and one or two are deleted because of misfit, the resulting reliability estimates may be unsatisfactorily low. This dictates the addition of items or the revision of previously misfitting items to better align with the latent variable under investigation.

An analysis of threshold structure may demonstrate that some of the response options are simply not being used by respondents. Understanding which categories are not being used may lead to the decision to collapse adjacent response categories because respondents are simply not differentiating between the theorized gradations.

Finally, a PCA of the standardized residuals may suggest additional dimensions in the data. This may result in the splintering of a group of items previously thought to capture one variable into multiple sub-scales, each representing new variables that may not have been identified in the initial variable definition process. These sub-scales would then be treated as independent scales and analyzed independently of one another.

Having summarized the theory and practice behind a methodology for rating scale development that will produce invariant, equal-interval measures built upon a firm foundation of precision (reliability) and meaning (validity), attention will now be turned to potential threats to the validity of conclusions drawn from rating scale data.

Halo Effect

In the context of rating environments, data are subject to a variety of sources of error. One source of error is *halo effect* which was broadly defined earlier in this study as the tendency of a rater to attend to a global impression of the ratee rather than to distinguish differing levels of performance on separate dimensions. If, for example, my feeling of rapport with a teacher influences my rating of the teacher's performance in areas other than rapport, then halo effect would be present in those ratings. This section of the literature review will summarize the research related to the conceptualization and estimation of halo effect.

Researchers generally consider halo effect to be one of the most common and persistent errors in performance ratings (Cooper, 1981; Feeley, 2002; King, Hunter, & Schmidt, 1980; K. Murphy, 1982; Pike, 1999; Frank E. Saal, Downey, & Layhey, 1980). Feeley (2002) explains that a variety of different labels have been given to raters' tendency to overestimate the covariance between traits including *the logical error* (Newcomb, 1931), *correlational bias* (Berman & Kenny, 1976; Kenny & Berman, 1980), *illusory halo* (Cooper, 1981) and most popularly *halo effect* (Thorndike, 1920). Cooper (1981) cites an even earlier source (Wells, 1907) who addressed this phenomenon and writes regarding ratings of literary merit

> There is noted introspectively a tendency to grade for general merit at the same time as for the qualities, and to allow an individual's general position to influence his position in the qualities . . . especially in the case of those qualities that are ill-defined in the minds of the subjects, and tended to be interpreted rather in terms of general merit. . . . This would make the correspondence of such qualities appear closer than they were. It probably does not play any serious part . . . but it is difficult to see how it could have been avoided. (p. 21)

Fisicaro and Lance (1990) group the majority of conceptual definitions of halo into three categories based on the stated or implied cause of halo effect: (a) inadequate discrimination of ratee behavior, (b) a general impression of the ratee, and (c) perception of the ratee in terms of a salient dimension or characteristic.

Cooper (1981) makes an important distinction by conceptualizing halo as a combination of *true halo* and *illusory halo*. He explains that the absence of true halo would mean that different dimensions of job performance, for example, are completely uncorrelated. This is highly unlikely, however, because job behaviors are frequently and necessarily correlated with one another. Cooper explains that “job performance ratings with substantial true halo should be particularly likely in organizations with stringent job retention policies, in mechanistic organizations, and in positions with a low selection ratio” (p. 221). True halo is defined, therefore, as real (true) correlations between different abilities or traits. Cooper continues “unless cross-category performance ratings are entirely composed of error scores, some of the halo observed will reflect true halo.” In other words, ratings from homogenous domains will always contain true halo. The definition of illusory halo is the amount of halo in observed ratings that exceeds true halo. True halo contributes meaningfully to the ratings, while illusory halo blurs ratings. In the relatively homogenous domain of teacher performance it would be expected that true halo would be present.

While there is some consensus on the conceptualization of halo, procedures for estimating halo effect have enjoyed significantly less agreement. A variety of approaches have been proposed. These include (a) intervariable correlations (Keaveny & McGann, 1975; Thorndike, 1920), (b) intraratee variance across variables (Bernardin & Pence, 1979; Bernardin & Walter, 1977; Borman, 1975, 1977; Kiker & Motowidlo, 1998; Moritsch & Suter, 1988), (c) general impression procedure (Fisicaro & Lance, 1990;

Lance, Fisicaro, & LaPointe, 1990), (d) intervariable factor structure (Blanz & Ghiselli, 1972; Kraut, 1975), (e) rater by ratee by variable analysis of variance (ANOVA) (Dickinson & Tice, 1977; Guilford, 1954; Kavanagh, MacKinney, & Wollins, 1971), (f) effect of manipulated variables on non-manipulated variable ratings (Bartlett, 1983; Wherry & Bartlett, 1982), and (g) many-facet Rasch model estimation (Engelhard, 1994; Linacre, 2004b; Myford & Wolfe, 2004a, 2004b).

Classical Approaches

Intervariable correlations. This approach involves the calculation of intercorrelations among ratings on a number of variables. Higher correlations indicate an inability to discriminate between variables. These correlations are then compared to some model of allegedly true correlation values to determine whether halo effect is exhibited.

Intraratee variance. This approach requires the calculation of intraratee variance or standard deviation across variables. Small variance estimates are an indication that halo is operating. Inferences are drawn regarding the observed variance compared to some model of true variance.

General impression. This procedure requires the calculation of correlations between each variable (E_i, E_j) and the raters' overall evaluation (GI) of the ratee(s). Halo effect is given by Equation 12:

$$HE_{GIij} = r_{e_i},_{GI}\; r_{e_i,GI} \;. \qquad (12)$$

Intervariable factor structure. This approach requires a factor analysis to determine the structure of the data. The assumption underlying this approach is that the presence of a dominant general factor accounting for a significant amount of variance does not reflect the expected multidimensionality based on the several variables represented, but the failure of raters to differentiate between variables, and therefore halo effect.

Analysis of variance. If an analysis of variance (ANOVA) conducted in a rater by ratee by variable design produces a rater by ratee effect, particularly if the interaction explains much of the variance in the ratings, then halo effect is considered present.

Manipulated variables. An example of this approach (Bartlett, 1983) is based on the theory that the rating process may be described in terms of various systematic and random error components as expressed in Equation 13:

$$z_R = W_T z_T + W_B z_B + W_I z_I + W_E z_E \tag{13}$$

where z_R is the rating response broken into the components due to true ability of the ratee ($W_T z_T$), bias of the rater ($W_B z_B$), environmental influences ($W_I z_I$), and various sources of random error ($W_E z_E$). In an effort to distinguish between valid (true) halo, and invalid (illusory) halo, Bartlett suggests the use of differential item validities. This approach grew out of a study that demonstrated high correlations between item validities and strong factor loadings on a general factor (i.e., halo) (Bartlett, 1966). In order to distinguish between valid (true) halo and invalid (illusory) halo, variables were manipulated producing two sets of items, one set of valid items (subscript $_v$ in Equation 14) and another set of less valid items (subscript $_i$ in Equation 15). Assuming that differential environmental influences and random error have been controlled, the scores on these two sets of items are represented by Equations 14 and 15:

$$z_{R_v} = W_{T_v} z_{T_v} + W_B z_B \tag{14}$$

$$z_{R_i} = W_{T_i} z_{T_i} + W_B z_B \ . \tag{15}$$

By subtracting or partialing these less valid ratings from the more valid ratings the bias component is reduced or eliminated.

Rasch Model Approaches

Individual-level statistical indicators. Engelhard (1994) argues that classical methods for halo diagnosis cannot distinguish between true halo and illusory halo. He proposes a multi-facet Rasch model approach to diagnosing illusory halo. The multi-facet Rasch model is expressed in Equation 16:

$$\ln[\frac{P_{nijk}}{P_{nij(k-1)}}] = \theta_n - b_i - s_j - c_k \tag{16}$$

where

P_{nijk} is the probability of ratee *n* being rated *k* on domain *i* by rater *j*,

$P_{nij(k-1)}$ is the probability of ratee *n* being rated *k - 1* on domain *i* by rater *j*,

θ_n is the ability of the ratee *n*,

b_i is the difficulty of domain *i*,

s_j is the severity of rater *j*,

c_k is the difficulty of rating threshold k relative to threshold $k - 1$.

Engelhard operationalizes halo error as unexpected uniformity represented by infit and outfit statistics for a rater that are significantly less than the expected value of 1 which he defines as an infit or outfit mean square value less than 0.5. If the rater’s fit statistics calculated over ratees are small, then the rater is not reliably distinguishing among the several variables intended by the developers of the scales.

Like Engelhard, Myford and Wolfe (2004b) recommend the use of the infit and outfit mean square indices for halo effect diagnosis. However, they also propose an additional individual-level statistical indicator in the form of a bias interaction analysis.

An expanded version of Equation 17 allows for estimation of a measure for each trait (T_i) and separate performance measures for each and every combination of an individual rater with a trait, (I_{ji}):

$$\ln[\frac{P_{nijk}}{P_{nij(k-1)}}] = \theta_n - b_i - s_j - T_i - c_k - I_{ji}\,. \tag{17}$$

In Equation 18, I_{ji} is the bias interaction term. This model allows the investigator to determine the degree to which the observed rating for every rater/trait pairing deviates from the rating predicted by the model that does not contain the bias interaction term. FACETS computes the interaction term and a z-statistic that tests the null hypothesis that $I_{jt} = 0$:

$$Z_{bias} = \frac{I_{ji}}{SE_{I_{ji}}} . \tag{18}$$

If the value of z_{bias} is significant (± 2) then the particular rater is exhibiting unexpectedly severe or lenient rating behavior. Once the rater is identified as misfitting, then the researcher must examine the observed and expected ratings for misfitting raters to determine if the misfit is the result of halo effect. Halo effect would be diagnosed if there were consistently unexpectedly high or low ratings relative to the difficulty of the specific items.

A final individual-level Rasch model approach to halo effect diagnosis has been proposed by Linacre (2004b). This approach will be referred to as like-difficulty item anchoring. This approach to detecting halo effect involves anchoring all items at the same difficulty, usually 0, and then generating rater fit statistics. Raters who best fit this scenario are most likely to be exhibiting halo effect.

Group-level statistical indicators. In addition to individual-level indicators, Myford and Wolfe (2004b) recommend investigation of halo effect based on a number of group-level indicators including (a) the fixed chi-square test, (b) trait separation ratio, (c) trait separation index, and (d) reliability of the trait separation index.

In conclusion, halo is correctly labeled as "ubiquitous" (Cooper, 1981). Given its ever present nature, numerous approaches have been developed to estimate and control halo. Engelhard (1994) points out, however, the problematic nature of applying a collection of ad hoc procedures that yield inconsistent findings and recommendations for the improvement of the quality of rating scale data. Applying the Rasch model, a special case of additive conjoint measurement, Engelhard proposes that a more consistent strategy lies within the context of objective measurement based on the arguments articulated earlier in this chapter. While optimistic about the potential of objective measurement models, Engelhard acknowledges the need for additional research to determine how best to assess halo effect within the context of model fit indices.

This chapter has examined the literature relevant to the development and assessment of rating scales as well as the definition and estimation of halo effect. The next chapter will outline the method employed to answer the research questions identified in Chapter 1.

Chapter 3: Method

Scale Development

Identification of Areas of Valued Teacher Performance

A thorough content analysis of CES objectives, current training materials, and policy manuals related to teaching performance was conducted in order to identify areas of valued teacher performance (Church Educational System, 1994, 2000, 2003a, 2003b). A committee of five staff members representing CES Training Services and Research Information Services was formed to facilitate this process. Committee members were instructed to independently review the institutional training documents and to identify the competencies identified with teaching performance. The committee created a matrix of with competencies as rows and institutional documents as columns allowing the committee to identify the relative emphasis given to each proposed competency. The wording of each competency was negotiated until a consensus was reached.

Conceptualization of Areas of Valued Teacher Performance

Operational definitions of each area of valued teacher performance were constructed in terms of the language of the institutional training documents. The purpose of this exercise was to more closely align the measurement instruments with the explicit values and objectives of CES. The conceptualization of the core competencies were tested in focus group interviews with CES teachers and students. The purpose of the focus group interviews was to determine to what degree the verbal conceptualizations of core competencies align with the way in which key stakeholders conceptualize those competencies. Given the possibility that the committee’s conceptualizations of core competencies were lacking in some respect, stakeholders were solicited for additional competencies and indicators that may not have been identified by the committee. Focus groups also allowed the author to obtain stakeholder language related to the core competencies so that scale items could be written in a way that reflected the language used by stakeholders in communicating about the core competencies. Appendix A contains the protocol for conducting the focus groups interviews that was developed and implemented. Eight groups of teachers were interviewed including a total of thirty-two individual teachers.

Three groups of students were interviewed including a total of twenty four students. The interviews conducted with teachers ranged in length from 60 to 90 minutes, while student interviews lasted from 40 to 60 minutes. Informed consent forms were obtained from all participants (Appendix B). Questions addressed in the focus group interviews included

1. What are the key abilities or competencies that make for effective teaching in a seminary or institute class?
2. What would you accept as evidence that a teacher has these abilities?
3. After reviewing the competencies the committee identified what competencies do you think are missing, if any?
4. What would you accept as evidence that a teacher has each of these abilities?
5. How would you word a question to get at the indicators you have identified? Would that wording make sense to both seminary and institute students, teachers, and administrators? How might they be worded so as to make sense to everybody?

Identifying Indicators and Item Writing

Lists of indicators for each of the core competencies were constructed using the content analysis of the institutional documents and the focus group interviews. Indicators are those observable behaviors or expressed attitudes that stakeholders agree demonstrate the presence or absence of the ability. Items and scales were constructed around each of the core abilities and their respective indicators. The resulting items were reviewed by content experts and members of the CES Training Services division to determine appropriateness and to reinforce a content validity argument for the scales.

Study Sample

The scales were tested using a convenience sample at the Lehi Senior Seminary in Lehi, Utah. This seminary is of average size for programs along the Wasatch Front with seven full-time faculty and 1,090 students enrolled during term 1 of the 2004-2005 school year. The seminary serves grades 10, 11, and 12. This program has an alternating day schedule (A/B schedule) with ninety minute class periods. One class period prior to administration all teachers and students were given a consent form and invited to participate (student and teacher consent forms are provided in appendices C and D). The scales were then administered to all students taught by seven different teachers. Students were asked to respond to all items on all scales (i.e., 92 items constituting 12 scales). Four-hundred-six usable surveys were obtained. Gender was identified on 388 of the surveys (206 males, 182 females). Respondents identified grade level on 389 of the surveys (164 in grade 10, 153 in grade 11, and 72 in grade 12).

Item and Scale Analysis

Exploratory factor analyses were conducted using the SPSS software package to assess the dimensionality of the response data and to reconfigure items on scales with defensible unidimensionality. The Rasch rating-scale model was then applied to the resulting constellation of items to diagnose scale and item functioning. Additionally, confirmatory factor analyses were conducted using the AMOS software package. Based on this array of analyses, recommendations for revisions to scales, items, and response categories were outlined.

Halo Effect Diagnosis and Gender Comparisons

Using the reconstituted scales from the above analyses, halo effect was diagnosed using those methods suited to the particular rating situation addressed in this study. Several of the diagnostic approaches outlined in Chapter 2 were used to diagnose halo effect including (a) intervariable correlations, (b) intraratee variance, (c) intervariable factor structure, (d) many-facet Rasch model infit/outfit mean squares, (e) many-facet Rasch model like-difficulty item anchoring, (f) many-facet Rasch model group-level statistical indicators, and (g) many-facet Rasch model individual-level bias-interaction estimation. Estimates were calculated by gender group for comparison of variable halo effect by gender.

Chapter 4: Results

Research Question 1: What are the key areas of teacher performance valued by CES administrators, teachers, and students?

The committee reached a consensus on the following areas of desired performance by CES teachers:

1. teaches students the Gospel of Jesus Christ,
2. teaches by the Spirit,
3. teaches by example,
4. establishes and maintains an appropriate setting,
5. helps students accept responsibility for gospel learning,
6. effectively decides what to teach,
7. effectively decides how to teach,
8. effectively uses scripture study skills,
9. effectively uses teaching skills,
10. relates well with students,
11. prepares young people for effective church service, and
12. has high expectations for students.

All of the competencies were identified in the institutional documents with the exception of "relates well with students" and "has high expectations of students." The committee determined that while not explicit in the institutional documents the domain of student rapport is essential to effective teaching and implicit in the institutional documents. Additionally, recent statements by members of the CES Board of Education have focused attention on the need to communicate high expectations for student performance. Therefore, the committee decided to add an additional scale dealing specifically with teacher expectations regarding scripture mastery. Focus group interviews revealed a consensus among administrators, teachers, and students with regards to the areas of performance listed with the exception of "the teacher has high expectations of students for scripture mastery." This last competency was not an area of valued teacher performance reflected in the focus group interviews. It was included in the pilot study, however, at the request of the committee.

Research Question 2: In what ways do students conceptualize these areas of valued teacher performance?

The results of the focus group interviews were transcribed and analyzed. Core indicators were distilled from this analysis in the language of the focus group participants. Some representative expressions of the focus group participants are listed below.

1. Teaches students the gospel of Jesus Christ as found in the standard works and the words of the prophets.
    a. They teach from the scriptures.
    b. They teach less opinion and more doctrine.
    c. They avoid expressing personal opinions.
    d. They recognize their own opinions.
    e. They teach what the prophets teach.
2. Teaches by the Spirit

   a. Students feel the Spirit (burning in the bosom, good feeling throughout body, peace, happy, want to come back, comforted).

   b. Teachers are passionate about what they teach.

   c. Teachers bear their testimonies frequently.

   d. The lesson flows easier.

3. Teaches by example

   a. They are dependable; they don’t miss classes often.

   b. They live what they teach.

   c. They are committed to seminary.

   d. They were living righteous lives by the power of their testimony.

   e. They are happy to be teaching seminary; to be involved in the Church.

4. Establishes and maintains an appropriate setting.

   a. There is a relaxed environment.

   b. We are expected to plan and present good devotionals.

   c. The teacher keeps things clean and neat.

   d. The teacher expects students respect each other.

5. Helps students accept their responsibility for gospel learning.

   a. Teacher provides incentives for scripture reading.

   b. Teacher shows by example.

6. Effectively decides what to teach.

   a. They got me into the scriptures.

   b. You can be over-prepared and therefore inflexible and unsusceptible to the Spirit.

   c. They stayed in the scriptures and applied it to our lives.

   d. They showed how to apply the scriptures in our lives.

7. Effectively decides how to teach.

   a. They make you personally think about the questions and draw your own conclusions.

   b. They asked questions that caused me thinking.

   c. They shared the background so that I could make sense of what I was reading.

   d. They teach in a lot of different ways.

8. Effectively uses scripture study skills.

   a. The teacher helps us see the big picture in the scriptures.

   b. They teach us ways to mark our scriptures that helps us down the road.

   c. They encourage us to memorize the scripture master scriptures.

   d. We connect scriptures together that help us understand more.

9. Effectively uses teaching skills.

   a. Students are paying attention; they aren't staring at the wall.

   b. Teachers use fun and games to teach the gospel.

   c. Teachers use variety; they change it up.

   d. They use visual aids; they use object lessons and props.

   e. It's bad when the teacher goes on a tangent and doesn't bring the lesson back to the main point.

   f. A good teacher answers questions when they are raised because the students are ready to learn.

   g. Teachers are ineffective when they just read the scriptures without much thought or preparation.

   h. Poor teachers ignore or fail to acknowledge the comments or questions of the students; comments like “I don’t know where you got that idea;” We’ve got to move on or we won’t cover the whole chapter;” “I’d love to get your comments, but we’ve got to move on.”

10. Relates well with students.

   a. Teachers listen to their students.

   b. Teachers are understanding of their students.

   c. They don’t have favorites.

   d. They compliment students.

   e. They aren’t mean.

   f. They don’t look down on people.

   g. They see the positive side.

   h. I knew the teacher cared about me and because of it I listened; they listened to me and I knew they cared and considered by question or comment.

   i. Teachers know the students personal lives and acknowledge their accomplishments and special days.

   j. Shake their hand and look them in the eye every day; make some kind of personal connection every day.

11. Prepares students for effective Church service.

a. Students would be willing to accept and fulfill callings and honor the priesthood.

b. Students would come away from seminary knowing how the spirit communicates; how to recognize the spirit and follow it.

c. A sign that students are well prepared for church service is their willingness to teach others what they know and to encourage others to be obedient.

d. Teachers accomplish this by being an example in their own church service.

e. Class structure like presidency or devotional assignments can help them prepare.

f. Build faith in God and in the Church by sharing bits of testimony throughout the year that will create a desire to serve in the Church.

g. Affirm that the prophets and Church leaders have made sacrifices and apply that to the students.

Based on these conceptualizations the committee reached consensus on questions intended to sample from the universe of possible questions for each area of performance. Those items are listed below. The actual forms, as administered to students, are found in Appendix E.

1. Teaches students the Gospel of Jesus Christ as found in the standard works and the words of the prophets.

My teacher

a. Avoids sharing a lot of personal opinions not found in the scriptures.

b. Helps us to understand how specific scriptures relate to the plan of salvation.

c. Teaches us from the teachings of the modern prophets.

d. Avoids speculating or guessing about things that are not clear in the scriptures.

e. Teaches us what the scriptures and the prophets say not what others say about them.

2. Teaches by the Spirit.

My teacher

a. Helps us to recognize the influence of the Spirit in the classroom.

b. Focuses on us and what we are feeling.

c. Invites the Spirit by the way he/she acts in class.

d. Gives us a chance to share spiritual thoughts and feelings.

e. Helps us to understand how to invite the Spirit in class.

f. Testifies of the truthfulness of gospel truths.

g. Avoids doing things that drive the Spirit away.

3. Teaches by example.

My teacher

a. Practices what he/she preaches.

b. Shows us how the gospel has changed him/her by the way he/she acts.

c. Shares examples of gospel principles at work in his/her own life.

d. Sets an example for us by always being well prepared to teach.

e. Deals with discipline problems in a Christ-like way.

f. Sets a righteous example for us in everything he/she does.

g. Shows love for God by the way he/she treats others.

4. Establishes an appropriate setting.

My teacher

a. Keeps the classroom neat and well-organized.

b. Arranges the seating in a way that makes it easy for all of us to pay attention and learn.

c. Expects us to do meaningful devotional every class period.

d. Always starts and ends class on time.

e. Does not allow inappropriate behavior in class (e.g., bad language, disrespect).

f. Avoids being a friend rather than a teacher to us.

g. expects us to respect the house or building where class is held.

5. Helps students accept responsibility for gospel learning.

My teacher

a. Asks questions that help us figure things out ourselves.

b. Expects us to ask questions when we do not understand.

c. Expects us to study the scriptures on our own without being reminded.

d. Does not speak down to us but treats us as fellow students of the scriptures.

e. Asks us what we have discovered in the scriptures in our own study.

f. Makes us work at understanding the scriptures ourselves instead of just explaining it to us.

g. Teaches us how to study the gospel on our own.

6. Makes good decisions about what to teach.

My teacher

a. teaches us principles that really apply in our lives.

b. focuses on parts of the scriptures that are most important.

c. teaches us in a way that is not too simple, but also not over our head.

d. organizes lessons so that the ideas are presented in an order that makes sense.

e. avoids spending a lot of time on things that are not very important.

f. takes time to answer our specific concerns or questions.

7. Makes good decisions about how to teach.

My teacher

a. teaches in a way that is very uplifting.

b. encourages us to get into the lesson by participating in class activities (group work, discussions, writing activities).

c. avoids taking too much time on one part of the lesson and rushing through the rest.

d. uses a lot of variety when teaching.

e. avoids teaching in a way that offends students.

8. Effectively uses scripture study skills

My teacher

a. teaches us to use the scripture study aids (footnotes, chapter headings, topical guide, Bible dictionary) by using them frequently in class.

b. helps us to learn the references for the scripture mastery scriptures (book, chapter, and verses).

c. asks us to look for particular things in the scriptures we are studying (principles, definitions, symbols, if/then relationships, patterns).

d. helps us to understand the doctrines found in the scripture mastery scriptures.

e. encourages us to mark our scriptures as we study.

f. helps us to see patterns in the scriptures by "chaining" or connecting related scriptures together.

g. reviews scripture mastery scriptures with us regularly.

h. helps us to liken the scriptures to our own situation in life.

i. summarizes the scriptures in a way that really helps us understand what is being taught.

j. gives us time to ponder and reflect on how the scriptures relate to us.

k. helps us to understand the language of the scriptures (e.g., when the use of "man" or "men" in the scriptures refers to both men and women).

l. helps us to memorize the scripture mastery scriptures.

9. Effectively uses teaching skills

My teacher

a. gives us writing assignments in class to help us learn (study exercises, tests, quizzes, instructional games, and essays).

b. avoids lecturing too much.

c. uses media (e.g., music, video) to effectively teach gospel principles. asks questions that really make us think.

d. listens carefully to the answers that we give to questions.

e. helps us learn by giving us productive and meaningful group assignments.

f. frequently calls on us by name.

g. gives respectful answers to our questions.

h. uses the chalk/white board really well to teach concepts.

i. avoids using videos all the time.

j. shares true stories to help us better understand gospel principles.

k. uses music to effectively teach gospel principles.

l. avoids exaggerating true stories to get an emotional reaction.

10. Relates well with students

My teacher

a. shows sincere interest in what we are doing in our lives.

b. looks me in the eye when we are talking.

c. knows my name.

d. does not embarrass us.

e. gives us sincere compliments when appropriate.

f. shows love and respect to all of us.

g. makes us feel comfortable talking to him/her.

11. Prepares young people for effective church service

My teacher

a. shares personal experiences of Church service that helps us prepare to serve.

b. gives us a chance to learn service in class (class leadership, devotionals, teaching, or other assignments).

c. frequently explains how the lesson relates to serving in the Church.

d. teaches us how the Spirit works when we are serving others.

e. explains how the lesson relates to being an effective parent.

f. gives us responsibilities that help us develop leadership skills (e.g., delegation, follow-up, accountability).

12. Teacher has high expectations of students

My teacher expects us to

a. read the scriptures outside of class every day.

b. understand key doctrines like the plan of salvation, the Atonement of Jesus Christ, the Apostasy, and the Restoration.

c. find specific principles in the scriptures that we can apply in our lives.

d. know and be able to explain the background of each of the scripture mastery scriptures (time, people, situation).

e. be able to explain the principles we find in the scriptures.

f. memorize all of the scripture mastery scriptures.

g. find relationships or connections between different scriptures.

h. know all of the references to the scripture mastery scriptures (book, chapter, verses).

i. be able to clearly explain key doctrines like the plan of salvation, the Atonement of Jesus Christ, the Apostasy, and the Restoration.

Research Question 3: To what degree do the items derived from student conceptualizations function to produce reliable data from which valid conclusions may be drawn about teacher performance?

Classical Analyses

Exploratory factor analysis. Exploratory factor analyses were conducted in order to assess the dimensionality of the scales. Because the scales were expected to be correlated (i.e., teachers who have good rapport with students are likely to also be good at creating a spiritual learning environment) and the obtained bivariate correlation estimates support this assumption (see Table 6), an oblique rotation (Promax) was applied in the EFA. Netemeyer, Bearden, and Sharma (2003) explain that oblique rotation tends to reveal the more meaningful theoretical factors when the factors are correlated.

Table 6

Bivariate Correlation Matrix for Three Scales

Scale	Student-Teacher Rapport Scale	Scripture Mastery Expectation Scale	Spiritual Learning Environment Scale
Student-Teacher Rapport Scale	1.00		
Scripture Mastery Expectation Scale	.57	1.00	
Spiritual Learning Environment Scale	.66	.50	1.00

Initial analyses revealed one dominant factor and several additional factors with eigenvalues greater than 1.0. Based on the factor loadings from these analyses, the items were regrouped and reanalyzed. This process was repeated in an iterative fashion until only three 6-item scales remained that the author considered defensible: (a) the Student-Teacher Rapport Scale (STRS), (b) the Scripture Mastery Expectation Scale (SMES), and (c) the Spiritual Learning Environment Scale (SLES). The items constituting each of the revised scales are displayed in Table 7. All three scales produced responses that meet generally accepted criteria. These criteria include (a) eigenvalues greater than 1.0, (b) satisfactory results of the scree test, (c) item factor loadings greater than .40, (d) satisfactory variance explained by the set of extracted factors (50-60 %), and satisfactory variance explained by each individual factor extracted (greater than 5%). Table 8 displays the eigenvalues for each of the three factors in column 2. Each of the factors generated eigenvalues greater than 1.0. The percent of total variance explained by each factor is displayed in column 3. The variance explained by each factor exceeds 5%. The cumulative percent variance explained by the factors is displayed in column 4. The total variance explained by the three factors was 55%. The scree plot in Figure 8 illustrates the dominance of the first factor and the relative magnitude of factors 2 and 3. The grouping of items is graphically depicted in rotated factor space in Figure 9. The criterion of simple structure was achieved in that the items constituting each scale all have a high loading on one factor and quite small loadings on each of the other factors. The primary factor loadings (shaded in gray) displayed in Table 9 exceed .40 and most exceed .50.

In addition to conducting exploratory factor analyses that included all items from all three scales, separate factor analyses were conducted for each individual scale. Table

Table 7

Items Constituting Three Revised Scales

Scale	Item Stem
Student-Teacher Rapport Scale	
1a	shows sincere interest in what we are doing in our lives.
1b	teaches in a way that is very uplifting.
1c	gives respectful answers to our questions.
1d	gives us sincere compliments when appropriate.
1e	shows love and respect to all of us.
1f	makes us feel comfortable talking to him/her.
Scripture Mastery Expectation Scale	
2a	reviews scripture mastery scriptures with us regularly.
2b	helps us to memorize the scripture mastery scriptures.
2c	expects us to know and be able to explain the background of each of the scripture mastery scriptures (time, people, situation).
2d	expects us to be able to explain the principles we find in the scriptures.
2e	expects us to memorize all of the scripture mastery scriptures.
2f	expects us to know all of the references to the scripture mastery scriptures (book, chapter, verses).
Spiritual Learning Environment Scale	
3a	teaches us the gospel according to the scriptures not "the gospel according to the teacher."
3b	teaches us what the scriptures and the prophets say not what others say about them.
3c	helps us to recognize the influence of the Spirit in the classroom.
3d	invites the Spirit by the way he/she acts in class.
3e	testifies of the truthfulness of the gospel principles he/she teaches.
3f	avoids doing things that drive the Spirit away.

Table 8

Eigenvalues and Percent of Variance Explained by Three-Factor Solution

Factor	Eigenvalue	Percent of Variance Explained	Cumulative Percent of Variance Explained
1	7.41	41.17	41.17
2	1.33	7.40	48.57
3	1.08	5.98	54.55

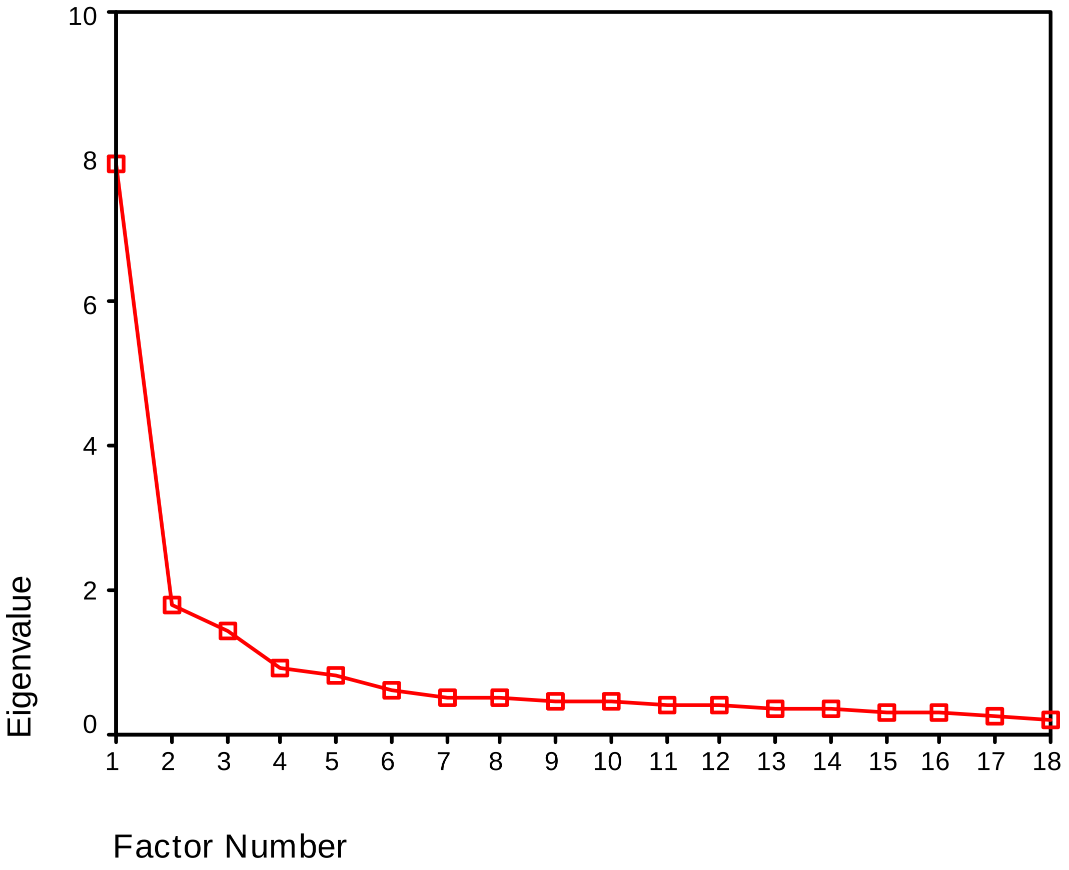


Figure 8. Scree plot for three factor solution.

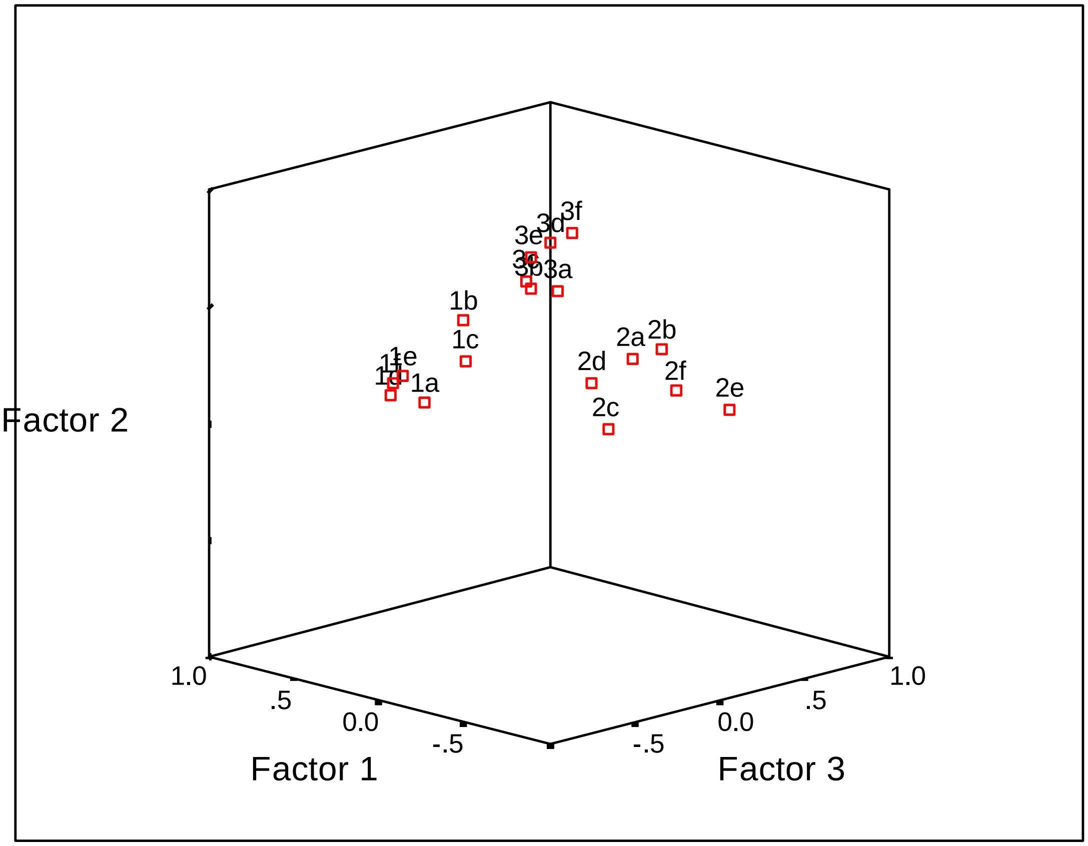


Figure 9. Three factor plot in rotated space.

Table 9

Pattern Matrix for Three-Factor Solution

	Factor Loadings		
Item	1	2	3
1a	.79	-.07	.06
1b	.50	.35	-.01
1c	.58	.14	.09
1d	.90	-.05	-.04
1e	.83	.05	-.03
1f	.89	.01	-.03
2a	.07	.16	.56
2b	-.04	.20	.62
2c	.26	-.19	.60
2d	.19	.06	.45
2e	-.13	-.09	.93
2f	-.05	.01	.70
3a	.00	.56	.05
3b	.08	.57	-.03
3c	.16	.57	.03
3d	.01	.77	.02
3e	.02	.73	-.09
3f	-.13	.84	.01

10 displays the eigenvalues and percent of variance explained from these follow-up analyses. The response data from each of these scales produced only one factor with an eigenvalue greater than 1. The first factor on the STRS produced an eigenvalue of 4.17

Table 10

Eigenvalues and Percent of Variance Explained for the STRS, SMES, and SLES

Factor	Eigenvalue	Percent of Variance Explained	Cumulative Percent of Variance Explained
Student-Teacher Rapport Scale			
1	4.17	69.56	69.56
2	.50	8.25	77.81
3	.44	7.28	85.09
4	.39	6.50	91.59
5	.27	4.54	96.13
6	.23	3.87	100.00
Scripture Mastery Expectation Scale			
1	3.36	55.95	55.95
2	.90	15.01	70.96
3	.58	9.69	80.65
4	.47	7.82	88.46
5	.36	6.07	94.53
6	.33	5.47	100.00
Spiritual Learning Environment Scale			
1	3.39	56.51	56.51
2	.78	12.93	69.44
3	.58	9.68	79.12
4	.48	8.05	87.17
5	.42	6.95	94.11
6	.35	5.89	100.00

and explained 70% of the total variance in the students' responses. The first factor on the SMES produced an eigenvalue of 3.36 and explained 56% of the total variance in the data. The first factor on the SLES produced an eigenvalue of 3.39 and explained 57% of the total variance in the students' responses.

Classical item statistics. A minimum coefficient alpha of .80 is generally sought in new scales (Bearden & Netemeyer, 1998; Clark & Watson, 1995). Additionally, inter-item correlations are expected to exceed .30 (Netemeyer et al., 2003; Robinson et al., 1991) and corrected item-total correlations should exceed .50 (Netemeyer et al., 1996; Obermiller & Spangenberg, 1998; Tian et al., 2001).

Coefficient alpha, inter-item correlations, and item-total correlations were estimated for all three scales. Table 11 displays the coefficient alpha estimates in column 2, the minimum and maximum inter-item correlations in columns 3 and 4, and the

Table 11

Classical Reliability Statistics for Three Scales

		Inter-Item Correlations		Corrected Item-Total Correlations	
Scale	Coefficient Alpha	Minimum	Maximum	Minimum	Maximum
STRS	.91	.54	.75	.69	.81
SMES	.84	.34	.64	.56	.69
SLES	.84	.37	.62	.55	.69

corrected item-total correlations in columns 5 and 6. All of the estimates meet or exceed established criteria.

Rasch Model Analyses

In the Rasch model analyses maximum extreme scores, minimum extreme scores, and scales on which a respondent provided no answers were excluded from the calculation of person reliability and fit statistics. Additionally, the most misfitting persons were deleted from the analysis for reasons outlined in Chapter 2. Because of the differences between classical reliability estimation and Rasch person reliability estimation, it is not expected that the estimates will be the same. Rasch estimates are almost always more conservative than estimates of coefficient alpha. Recall that the reliability criteria identified in Chapter 2 for satisfactory Rasch estimates was .70, whereas the criteria for classical analysis was .80. Fit or misfit is assessed by the nearness of the observed fit estimates to the expected values of a mean square of 1.00 and a standardized *z* of 0.00. Person reliability and fit statistics are summarized in Table 12.

Table 12

Rasch Summary Person Statistics for Three Scales

		Person Infit		Person Outfit	
Scale	Person Reliability	Mean Square	Standardized *Z*	Mean Square	Standardized *Z*
STRS	.85	0.92	-0.10	0.94	-0.10
SMES	.80	0.98	-0.20	0.98	-0.20
SLES	.81	0.97	-0.10	0.98	-0.10

Column 2 contains the person reliability estimates, columns 3 and 4 contain person infit estimates, and columns 5 and 6 contain person outfit estimates. The person infit mean square and standardized *z* estimates range from 0.92 to 0.97 and -0.20 to -0.10 respectively. The person outfit mean square and standardized *z* estimates range from .94 to .98 and -0.10 to -0.20 respectively.

Item reliability, fit statistics, and point-measure correlations are summarized in Table 13. The Rasch model permits calculation of the consistency of the estimates for both persons and items. While Rasch model person reliability is analogous to CTT reliability estimates such as Cronbach's coefficient alpha, there is not a CTT estimate for item precision that is analogous to item reliability in the Rasch model. The Rasch item reliability is an indicator of the "replicability of item placements along the pathway [continuum] if these same items were given to another sample with comparable ability levels. . . . Therefore, from high item reliability, we can infer that we have developed a

Table 13

Rasch Summary Item Statistics for Three Scales

		Item Infit		Item Outfit		Point-Measure Correlation	
Scale	Item Reliability	Mean Square	Standardized *Z*	Mean Square	Standardized *Z*	Min.	Max.
STRS	.95	0.98	-0.30	0.94	-0.60	.80	.87
SMES	.96	1.00	0.00	0.98	-0.20	.74	.82
SLES	.94	0.98	-0.20	0.98	-0.20	.76	.81

line of inquiry in which some items are more difficult and some items are easier, and that we can place confidence in the consistency of these inferences" (Bond & Fox, 2001). Item reliability estimates are almost always higher than person reliability estimates because the number of respondents generally far exceeds the number items. Larger numbers of responses result in higher item reliability just as larger numbers of items result in higher person reliability.

Column 2 contains item reliability estimates ranging from .94 to .96 indicating that the author may have confidence in the consistency of inferences drawn from the items representing the three scales; columns 3 and 4 contain item infit statistics with mean square estimates ranging from 0.92 to 1.00 and standardized *z* estimates ranging from -0.30 to 0.00; columns 5 and 6 contain item outfit statistics with mean square estimates ranging from 0.94 to 0.98 and standardized *z* estimates ranging from -0.60 to -0.20; and columns 8 and 9 contain the minimum and maximum item point-measure correlations ranging from .74 to .87. All of these estimates reveal acceptable fit with the theoretical model and, therefore, provide a basis for confidence in drawing inferences about teacher performance from the results.

Having provided a description of summary scale statistics, attention will now be turned to a description of the detailed individual scale statistics. The presentation of these results will follow the pattern outlined in Chapter 2. A paragraph will describe how each scale functions in terms of each of the following issues: (a) person fit and reliability, (b) item fit and reliability, (c) differential item functioning, (d) response category structure, (e) variable map, (f) general keyform, and (g) PCA of standardized residuals. Key statistics are reported in the body of the paper, while the complete WINSTEPS tables from which these estimates were derived are included in Appendix F for the STRS, in Appendix G for the SMES, and in Appendix H for the SLES.

Student-teacher rapport scale. The Rasch person and item reliability estimates are .85 and .95 respectively well exceeding the criteria identified earlier for reliability. Based on these estimates we may have confidence in the replicability of person and item ordering along the trait continuum if persons were given another set of items measuring the same construct or if another group of persons' responses were used to estimate item measures. Individual item fit statistics and point-measure correlations are displayed in Table 14 and approximate expected values suggesting that the items on each scale are working together to reflect a single underlying construct. These fit statistics support the assumption of unidimensionality.

Table 14

Individual Item Statistics for the STRS

		Item Infit		Item Outfit	
Item	Point-Measure Correlation	Mean Square	Standardized *Z*	Mean Square	Standardized *Z*
1a	.87	1.05	0.60	1.07	0.70
1b	.80	1.18	1.90	1.18	1.50
1c	.81	1.13	1.40	1.15	1.20
1d	.85	0.87	-1.50	0.78	-2.10
1e	.87	0.69	-3.70	0.60	-3.90
1f	.87	0.93	-0.70	0.86	-1.30

As outlined in Chapter 2 an additional important consideration in scale analysis is the diagnosis of unusual differences in item performance for subpopulations. Gender groups were the only subpopulation of interest identified for this study. Differential item function by gender suggests the possibility of additional dimensions in the data that threaten the assumption of unidimensionality. It is important, therefore, to diagnose and address such threats. Using WINSTEPS differential item function by gender was examined. The full results of these analyses are displayed in the appendices in Figure F3 for the STRS, Figure G3 for the SMES, and Figure H3 for the SLES. The analysis of differential item function by gender in the STRS—defined by a *t*-statistic exceeding ± 2.0—is exhibited in items 1c ($t = 2.46$) and 1f ($t = -2.13$) (see Figure F3 in Appendix F). However, these statistics fall within acceptable parameters when response categories are collapsed suggesting that the revision of response categories may eliminate differential item function by gender in subsequent administrations and the assumption of unidimensionality can be defended. Further details regarding this issue will be addressed later in this chapter.

The next step in the process of scale analysis is an examination of response category structure. At least 10 observations in each response category were obtained using the STRS. However, a uniform distribution of observations across response categories was not achieved; the upper three categories (*undecided*, *agree*, *strongly agree*) garnered 95% of the responses. Average category measures advance monotonically and category fit statistics fall within acceptable parameters. Threshold calibrations advance by at least 1.4 logits, however, the threshold calibration between categories 4 and 5 advance by more than the targeted maximum of 5.0 logits. The category probability curves are displayed in Figure 10 revealing the large space between categories 4 and 5 suggesting the need to create gradations between these levels that will be meaningful to students and allow for more precise location of teachers at the upper end of the scale.

The relative difficulties of the various items along the theta continuum approximate the committee's understanding of the targeted construct. However, mean item and person measures are separated by approximately 4 logits illustrating the misalignment of items and persons that results in reduced precision because the items are too *easy*, in other words, the items do not tap the upper end of the continuum as well as desired.

The general keyform reveals relatively homogenous item difficulties and the predominance of person measures at the upper end of the theta continuum. This means that the item difficulties are easier than the person measures and suggests the need to revise items and response categories in order to better separate persons along the continuum. This is roughly analogous to measuring children's ability to do math by administering addition and subtraction questions when the children are already capable of multiplication. Proposed changes to achieve items that are more heterogeneous in difficulty and therefore better able to separate persons along the continuum will be addressed in the next chapter.

Finally, the measures account for 65% of the variance. Of the remaining unexplained variance, 9% is explained by an additional factor. Items 1b and 1c load on the secondary factor with loadings of .71 and .69 respectively. Remember that the PCA conducted in the Rasch model analyses examines the factor loadings of the standardized

```
P      ++-------+-------+-------+-------+-------+-------+-------++
R  1.0 +                                                         +
O      |                                                         |
B      |                                  4444444                |
A      |11                              44       44             5|
B   .8 +  1                            4           4          55 +
I      |   1                          4             4        5   |
L      |    1                        4               4      5    |
I      |     1            33333     4                 4    5     |
T   .6 +      1          3     3   4                   4  5      +
Y      |      1    2    3       3 4                     4 5      |
    .5 +       1 22 22 3         *                       *       +
O      |        *     *          43                     5 4      |
F   .4 +      221    3 2        4  3                   5  4      +
       |     2   1   3  2      4    3                 5    4     |
R      |    2     1 3    2    4      3               5      4    |
E      |   2       *      2  4        3             55       4   |
S   .2 +  2       3 1      24          3           5          44 +
P      |22       3   1     42           3         5             4|
O      |        3     1  44  22          333   555               |
N      |     333      4**1     222         5***3                 |
S   .0 +**************5555***************************************+
E      ++-------+-------+-------+-------+-------+-------+-------++
       -6      -4      -2       0       2       4       6       8
```

Figure 10. Category probability curves for the STRS.

residuals, whereas the factor analysis outlined earlier as part of the EFA examines the factor loadings of the raw scores. Both approaches allow the researcher to assess the unidimensionality of the scales, but from different perspectives. The unidimensionality of the STRS is strongly supported by these figures.

Scripture mastery expectation scale. The person and item reliability estimates were satisfactory (.80 and .96 respectively). Individual item fit statistics and point-measure correlations fall within established parameter and are displayed in Table 15.

Table 15

Individual Item Statistics for the SMES

		Item Infit		Item Outfit	
Item	Point-Measure Correlation	Mean Square	*Standardized Z*	Mean Square	*Standardized Z*
2a	.75	1.12	1.40	1.06	0.70
2b	.77	1.10	1.20	1.15	1.70
2c	.80	0.96	-0.50	0.94	-0.70
2d	.74	1.00	0.00	0.96	-0.40
2e	.82	0.88	-1.50	0.88	-1.50
2f	.77	0.96	-0.50	0.91	-1.10

With regards to differential item functioning, none of the *t*-statistics exceed ± 2.0 supporting the argument that the items do not function differentially by gender (see Figure G3 in Appendix G).

Category 1 (*strongly disagree*) was observed in only 9 instances in the SMES, however, the distribution of responses across the other categories was more even than in the previous scale. The skewness in responses in the STRS is also reflected in the SMES with categories 3 (*undecided*), 4 (*agree*), and 5 (*strongly agree*) accounting for 92% of the responses. Average category measures advance monotonically and category fit statistics fall within acceptable parameters. Threshold calibrations advance by at least 1.4 logits except between thresholds 3 and 4. None of the threshold calibrations advance by more than 5.0 logits. The category probability curves are displayed in Figure 11.

The relative difficulty of the various items approximates the committee's understanding of the targeted construct. In other words, the difficulty of each item on the scale (its location on the trait continuum), reflects the committees expectations about relative differences in difficulty between items. Mean item and person measures are separated by approximately 2.5 logits.

The general keyform for the SMES reveals the relative homogenous difficulty of the items and the predominance of person measures at the upper end of the theta continuum as found in the analysis of the STRS.

Finally, measures account for 61% of the variance. Of the remaining unexplained variance, 12% is explained by an additional factor. Items 2a and 2b load on this factor (.65 and .71 respectively). This secondary loading of items 2a and 2b suggests an additional dimension and a threat to the assumption of unidimensionality. Possible explanations for this phenomena and approaches to improving item performance are provided in the next chapter.

```
P      ++---------+---------+---------+---------+---------+---------++
R  1.0 +                                                             +
O      |                                                             |
B      |1                                                            |
A      | 11                                                         5|
B   .8 +   1                                                      55 +
I      |    1                                    44444           5   |
L      |     11         2222                   44     44        5    |
I      |       1      22    22               44         4      5     |
T   .6 +        1   22        2             4            4    5      +
Y      |         1 2           2     3     4              4  5       |
    .5 +          *             2  33 333 4                45        +
O      |         2 1             *3      *3                54        |
F   .4 +        2   1           3 2     4  3              5  44      +
       |       2     1         3   2    4   3            5     4     |
R      |      2       1       3     2  4     3          5       4    |
E      |    22         1    33       24       33      55         4   |
S   .2 +   2            1  3        442         3    5            44 +
P      | 22              **        4   22        3355               4|
O      |2              33  11    44      2       5533                |
N      |           3333      1***         22**555    33333           |
S   .0 +**********************555***********11***********************+
E      ++---------+---------+---------+---------+---------+---------++
       -6        -4        -2         0         2         4         6
```

Figure 11. Category probability curves for the SMES.

Spiritual learning environment scale. The person and item reliability estimates are .81 and .94 respectively. Summarized person fit statistics approximated expected values. Individual item fit statistics and point-measure correlations are displayed in Table 16. Reliability and fit statistics fall within the parameters outlined and support the assumptions of the measurement model.

A comparison of item calibration differences based on gender did not produce *t*-statistics greater than ± 2.0 (see Figure H3 in Appendix H).

Table 16

Individual Item Statistics for the SLES

		Item Infit		Item Outfit	
Item	Point-Measure Correlation	Mean Square	Standardized *Z*	Mean Square	Standardized *Z*
3a	.76	1.08	0.90	1.03	0.40
3f	.78	1.03	0.40	1.05	0.60
3c	.81	0.97	-0.40	1.02	-0.30
3b	.76	0.99	-0.10	0.98	-0.20
3e	.76	0.94	-0.70	0.97	-0.30
3d	.81	0.89	-1.40	0.85	-1.90

The threshold structure statistics revealed that all categories were used at least 10 times. However, the skewness observed in the previous scales was exhibited in this scale as well with categories 3, 4, and 5 accounting for 97% of the responses and categories 4 and 5 accounting for 88% of responses. The average category measures advance monotonically and category fit statistics approximate expected values. Threshold calibrations advance by at least 1.4 logits, but not by more than 5.0 logits. The category probability curves are displayed in Figure 12. The relative difficulties of the various items approximated the committee's understanding of the targeted construct. Items and persons were misaligned with the mean person measure and the mean item difficulty separated by four logits.

The general keyform reveal homogenous item difficulties and the predominance of person measures at the upper end of the theta continuum.

Finally, measures account for 62% of the variance, with an additional factor explaining 10% of the remaining variance. Items 3a and 3b load on this additional factor with loadings of .65 and .66 respectively. The secondary loadings of items 3a and 3b suggest a possible threat to unidimensionality and must be explored and addressed.

Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) was conducted using the AMOS software package. The objective of this analysis was to determine how well the factor structure fulfills the assumptions of unidimensionality and local item independence using structural equation modeling. CFA has become a generally accepted statistical approach to refining and finalizing scales (Bryant, 2000; Bryant & Yarnold, 1995; Netemeyer et al., 2003; Wothke, 1996).

Although AMOS produces a number of fit indices, those targeted in this study were the root-mean-square-error-of-approximation (RMSEA) and the comparative fit

```
P      ++---------+---------+---------+---------+---------+---------++

R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |                                    444444                 55|
B   .8 +11                                44      44             55  +
I      |  1                             44          44          5    |
L      |   1                           4              4        5     |
I      |    1                         4                4     55      |
T   .6 +     1               33      4                  4   5        +
Y      |      1    2       33  33   4                    4 5         |
    .5 +       1 22 222  33      334                      *          +
O      |       2*      23         43                     5 4         |
F   .4 +      2 1      322       4  3                   5   4        +
       |     2   1    3   2     4    3                 5     44      |
R      |   22     1  3     2   4      3               5        4     |
E      |  2        13       2 4        33            5          4    |
S   .2 +22         31        *          3         55            44   +
P      |         33  11    44 22          33     55                44|
O      |       33      1144     22          33355                    |
N      |    333       444111      2222    5555533333                 |
S   .0 +**************555555*****************************************+
E      ++---------+---------+---------+---------+---------+---------++
       -5        -3        -1         1         3         5         7
```

Figure 12. Category probability curves for the SLES.

index (CFI). Goodness of fit has been operationally defined as a RMSEA of .08 or less (Browne & Cudeck, 1993) and as a CFI greater than .90 (Bentler, 1990). More recent research by Hu and Bentler (1999), however, has proposed a RMSEA value of .05 or lower and a CFI value of .95 or higher.

The initial CFA produced a RMSEA of .08 and a CFI of .92. This model assumed no correlations between variables. When the variables were allowed to correlate, however, the goodness of fit indices improve to meet the more rigorous criteria proposed by Hu and Benter (1999). Figure 13 displays the variables in the CFA with the correlations represented by curved double-arrowed lines and the factor loadings with straight single-arrowed lines. The values of the estimates are also displayed. For example, the correlation between factor 1 and factor 3 is .70. Item 1b loads on factor 1 (.48), but has a secondary loading on factor 3 (.35). The error estimates for items 2a and 2b are correlated (.30) as are the error estimates of items 2b and 2e (.22). The revised model that permitted correlated variables and the secondary factor loading produced a RMSEA of .05 and a CFI of .96.

Research Question 4: In what ways might items and scales be revised to improve reliability and validity?

Student-Teacher Rapport Scale

An item-specific weakness exists in item 1b. This item has the second smallest factor loading (.73) of all the items on its scale when factor analyzed with only STRS items. It also has the greatest misfit of all the items, though the fit statistics are within acceptable parameters (outfit standardized *z* = 1.5). Item 1b also has the lowest point-measure

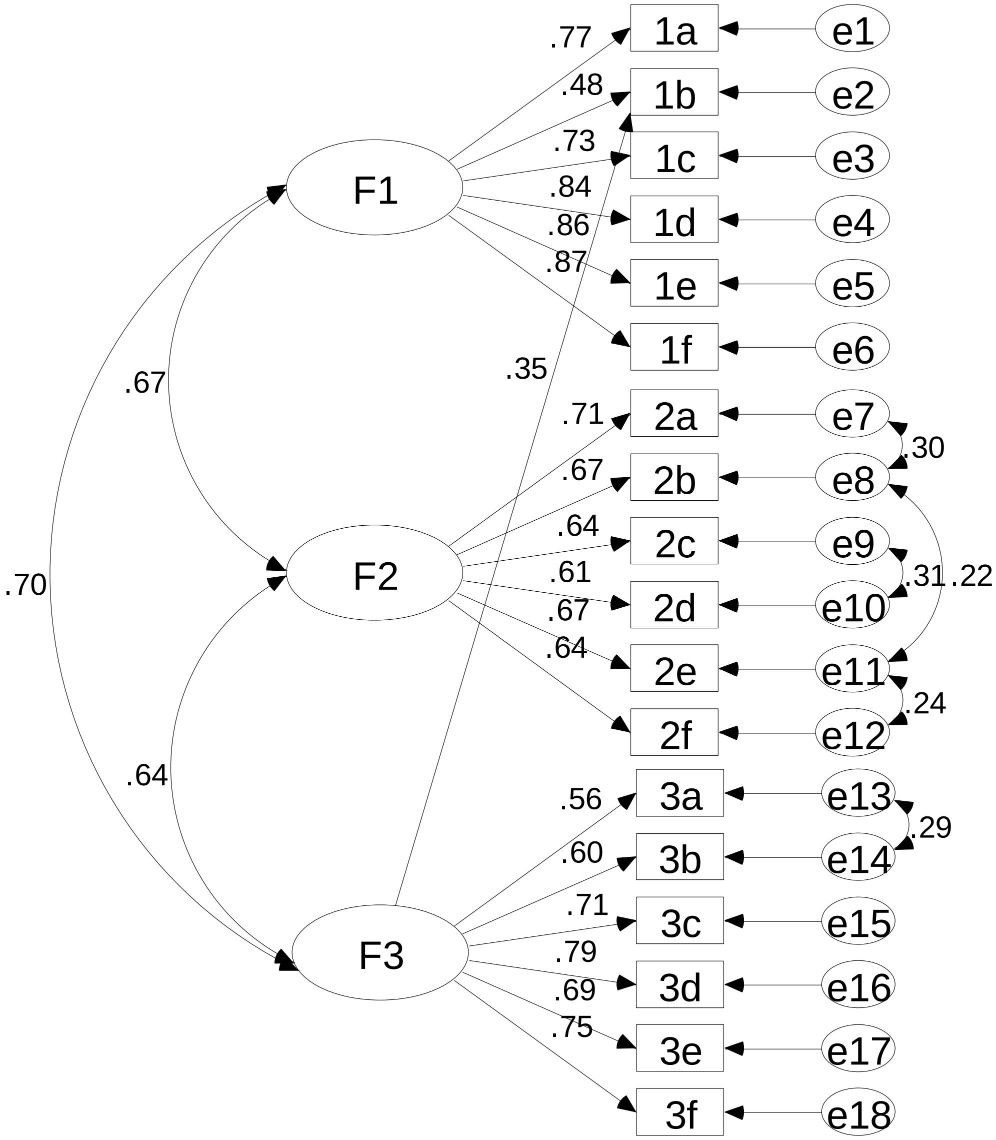


Figure 13. Variable map for three factor CFA.

correlation (.80), the smallest factor loading from the CFA (.48), and the largest additional factor loading in the PCA of standardized residuals. The CFA also reveals that 1b loads secondarily on factor 3 (.35). Item 1b is semantically different from all other items on the STRS scale in that it does not relate teacher performance to the students as directly as the other items on the scale. For example, item 1a reads: “shows sincere interest in what *we* are doing in our lives;” item 1c reads: “gives respectful answers to *our* questions;” and item 1d reads: “gives *us* sincere compliments when appropriate” (emphasis added). While teaching students is implicit in item 1b: “teaches in a way that is very uplifting,” it is likely to function better if it were to parallel the structure and focus in the other items. I would hypothesize, therefore, that the following wording would improve scale and item performance: “teaches in a way that uplifts *us*” by focusing attention on the students as do the other items in the scale. I anticipate that this change would result in improved fit and factor loadings.

Additional areas for scale improvement include better alignment of item difficulties with person measures and improved distribution of persons along the ability continuum. The variable map reveals that the mean item difficulty and the mean person measure are approximately 4 logits apart. It is difficult to conceive of items that would be sufficiently more difficult to endorse in order to better align items and persons. An alternative to adding new, more difficult items is to revise response categories in a way that would better tap the upper end of the scale. This approach is supported by the small number of respondents using the lower two categories (i.e., *strongly disagree*, *disagree*) and by the large distance (nearly 6 logits) between thresholds 3 and 4. This might be accomplished by collapsing categories 1-3 into a new category such as “somewhat agree” and retaining categories 4 and 5 (i.e., *agree*, *strongly agree*). Although the distance between the resulting thresholds exceeds the targeted maximum of 5 logits, the revised structure draws the item and person means closer together (approximately 1 logit separation) and suggests a potentially useful revision for subsequent administrations. The collapsing of these categories also eliminates the problematic DIF statistics that appear in items 1c and 1f by bringing the *t*-statistics within acceptable parameters (1.54 and -1.73 respectively). Additional options include the addition of a response category such as *very strongly agree* to better tap the upper limits of the scale or an entirely different response continuum such as *frequency*. However, the effects of these options can only be assessed through follow-up administrations.

Scripture Mastery Expectations Scale

Although close to acceptable parameters, the most misfitting items in this scale are items 2b (outfit standardized z = 1.7) and 2a (infit standardized z = 1.4). The CFA reveals that the error estimates for these items are correlated and thus threaten the assumption of local independence. The SMES is fraught with correlated error estimates. The threat to the assumption of unidimensionality in these items is corroborated by the standardized residuals PCA where items 2b and 2a load on an extraneous factor (.71 and .65 respectively). This may be explained by the fact that all the items constituting this scale begin with the word “expects” except 2a and 2b (error correlation = .30). On the other hand, the correlated errors between 2b and 2e (.22), 2c and 2d (.31), and 2e and 2f (.24) may similarly be explained by the fact that they share the same word. Items 2b and 2e share the word “memorize;” 2c and 2d share the word “explain;” and 2e and 2f share the word “all.” If the semantic similarities in items account for the correlated error estimates, a logical solution would be the rewording of individual items to make them more distinctive from one another. For example, change the items to read: My teacher expects us to

2a review the scripture mastery scriptures regularly.

2b learn the scripture mastery scriptures.

2c be able to explain the background of each of the scripture mastery scriptures (time, people, situation).

2d show that we can teach the principles we find in the scriptures.

2e memorize the text of the scripture mastery scriptures.

2f know all of the references to the scripture mastery scriptures (book, chapter, verses).

Although the category structure performs better for the SMES than for the STRS, the problem of a skewed person distribution is a concern. The collapsing of categories as described in the previous section produces improved scale function, particularly a better alignment of persons and items.

Spiritual Learning Environment Scale

In spite of excellent fit and good reliability estimates, the primary concern with these items is the correlated errors between items 3a and 3b (.29) from the CFA and the corresponding standardized residual PCA loadings of 3a and 3b on an additional factor (.66 and .65 respectively) from the Rasch analysis. The correlated errors and common loadings are likely due to the fact that both items address the teaching of opinions versus teaching directly from the scriptures. Item 3a focuses on the teacher's opinion and item 3b focuses on others' opinions. The problematic nature of these items may also be due to their beginning with the same expression, “teaches us.” I propose the following revisions to address this concern:

3a creates a spiritual feeling by teaching from the scriptures.

3b shares inspired teachings from the prophets to bring the Spirit into our lessons.

I also propose the revision of response categories recommended for the previous two scales using the same logic.

Research Question 5: To What Extent Do Male and Female Seminary Students Exhibit Differing Degrees of Halo Effect in their Ratings of Teachers?

Classical Approaches

Intervariable correlations. The bivariate correlation estimates and CFA revealed correlations between the scales that are within the range anticipated. In order to examine gender group differences a Pearson correlation coefficient was calculated for each possible pairs of scales. Table 17 displays the correlation coefficients for each scale pair in columns 2-4.

Table 17

Pearson Correlation Coefficients

Group	Rapport/Mastery	Rapport/Spirit	Mastery/Spirit
Males	.67	.75	.56
Females	.50	.57	.46

Intraratee variance. Recall that this approach defines halo in terms of small variability in an individual's responses to a number of items intended to measure different traits. In order to examine this approach to halo diagnosis, variances were calculated for each rater across all 18 items. Table 18 displays the summary statistics for these variances by gender group. Column 2 gives the number of respondents used in the calculation of the variance estimates. Slightly fewer female respondents were included in the analysis than males. The minimum and maximum variance estimates are displayed in columns 3 and 4. Seven percent of males and 3.8% of females had variances of 0. In other words, 7% of male and 3.8% of female respondents gave the same response to all 18 items. The mean variances and their standard deviations are given in columns 5 and 6. Both male and female respondents display small variances, but the means and standard deviations also reveal that female responses show greater variability than do males. Figure 14 contains histograms for males and females depicting the frequencies of the variance values.

Table 18

Intraratee Variance Statistics Across All 18 Items

Gender	Number	Minimum	Maximum	Mean	Standard Deviation
Males	185	.00	2.17	.38	.34
Females	172	.00	2.74	.51	.49

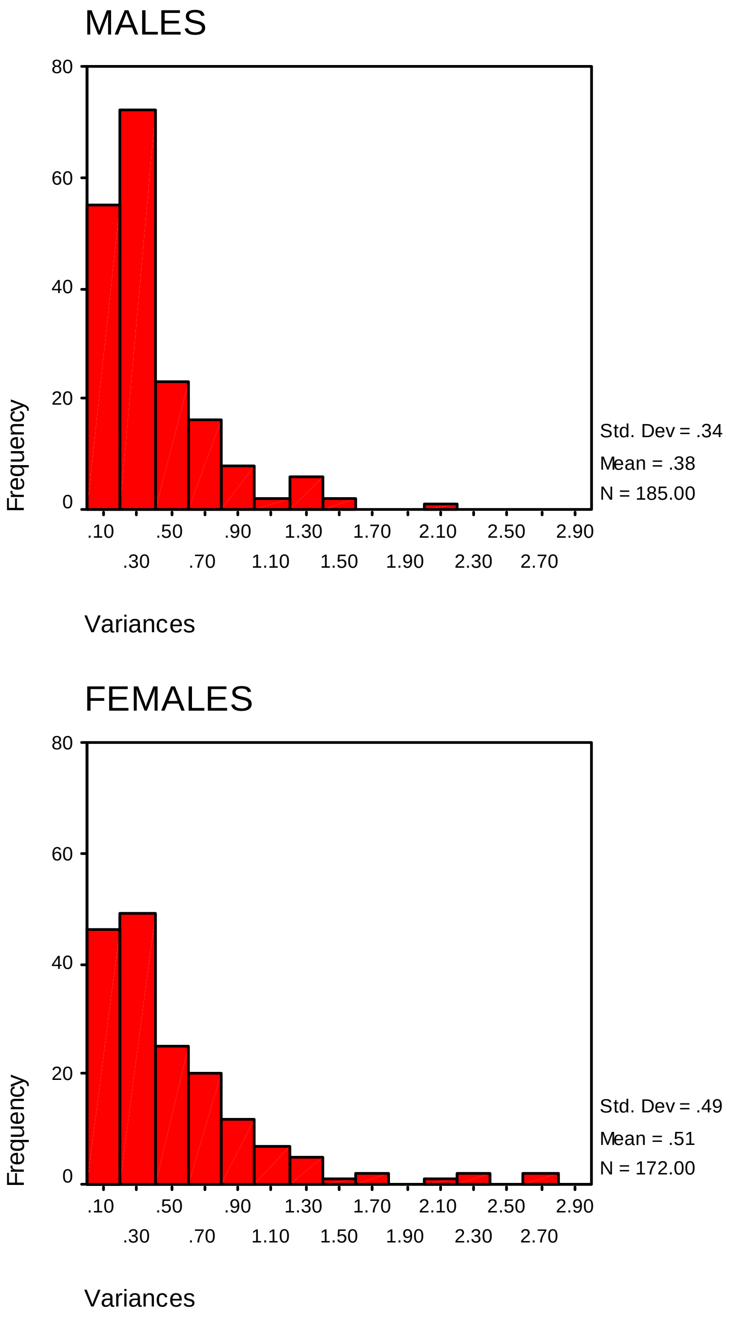


Figure 14. Frequency distributions of variance estimates by gender.

Intervariable factor structure. The presence of a dominant factor in the factor structure of the scales has already been documented. Table 19 displays the percentage of variance explained for each of the three factors when calculated using only male responses (column 2) and only female responses (column 3). However, the magnitude of the first factor when calculated separately for each gender group reveals that the dominant factor accounts for less variability when calculated using only female respondents (40%) compared to male respondents (45%). More variance is explained by factors 2 and 3 than factor 1 in the females' responses than in the males' responses.

Table 19

Percent of Variance Explained in Male and Female Response Data

Factor	Percent of Variance Explained in Male Responses	Percent of Variance Explained in Female Responses
1	45%	40%
2	7%	9%
3	4%	7%
Total	56%	56%

Rasch Model Approaches

Group-level statistical indicators. Four group-level statistical indicators were calculated: (a) a fixed chi-square test of the hypothesis that all items are of the same calibrated level of difficulty, (b) an item separation ratio, (c) an item separation index, and (d) a reliability of item separation index.

In the fixed chi-square test a nonsignificant chi-square value suggests halo in ratings of all raters. The chi-square statistics, their degrees of freedom, and their respective *p*-values for each scale are displayed in columns 2, 3 and 4 of Table 20. The chi-square values for all three scales are statistically significant (*p* < .005), indicating that at least two items are significantly different in terms of their difficulty. These results suggest that there is not a group-level halo effect present in this data set.

Table 20

Group-Level Multi-Facet Rasch Model Indicators of Halo for the STRS, SMES, and SLES

Gender	Chi-Square	Degrees of Freedom	Probability	Item Separation Ratio	Item Separation Index	Reliability of Item Separation Index
Student-Teacher Rapport Scale						
Males	71.8	5	.00	3.20	4.60	.91
Females	86.2	5	.00	3.51	5.01	.92
Scripture Mastery Expectation Scale						
Males	56.0	5	.00	2.98	4.31	.90
Females	85.9	5	.00	3.70	5.27	.93
Spiritual Learning Environment Scale						
Males	50.4	5	.00	2.69	3.92	.88
Females	43.2	5	.00	2.61	3.81	.87

The *item separation ratio* is a measure of the spread of item difficulties in relation to the precision of their estimates. Item separation ratios that reveal minimal separation suggest group-level halo effect. These ratios are displayed in column 5 of Table 20. The trait separation ratios for the three scales analyzed range from 2.61 (females responses to the SLES) to 3.70 (females responses to the SMES). These results indicate that the spread of item difficulty measures is about 3 times larger than the precision of those measures on the scales analyzed. This ratio is considerably smaller than those reported in other measurement studies (Myford & Wolfe, 2004b) suggesting the possibility of group-level halo. Females' responses result in larger separation ratios than males on the STRS and SMES, but a slightly smaller ratio on the SLES.

The *item separation index* is an indicator of how many statistically distinct levels of item difficulty are present in the data. Indices that reveal few distinct levels reflect group-level halo. These indices are given in column 6 of Table 20. The separation indices range from 3.81 (females responses to the SLES) to 5.27 (females responses to the SMES). These indices suggest that there are approximately 4 to 5 distinct strata of item difficulty. Like the previous indicator of group-level halo indicators females' responses result in a larger value than do males' responses on the STRS and SMES, but a slightly smaller value on the SLES. The relatively small number of strata indicate group-level halo.

The *reliability of item separation index* reveals how well item difficulties are separated. The estimates reveal how well raters distinguish items. When item difficulties are homogenous and, therefore, not well separated, halo effect is indicated. These estimates are displayed in column 7 of Table 20. The relatively high item separation reliabilities, ranging from .87 (females responses to the SLES) to .93 (females responses to the SMES) do not suggest a group-level halo effect in the data. The reliability estimates resulting from female responses exceed the estimates from males responses on the STRS and SMES, but are smaller on the SLES.

Individual-level statistical indicators. Using the multi-facet Rasch model like-difficulty item anchoring approach to halo diagnosis, all item difficulties were anchored at 0. In a sense, a simulated halo effect was created and the degree to which persons fit the model was examined. Fit was defined by an outfit mean square estimate less than 1.50 and an outfit standardized *z* estimate less than 1.90. Where persons fit the model we conclude that halo effect is at work. Where persons do not fit the model we conclude that they are not exhibiting halo. The results of these analyses are expressed in terms of the percentage of persons fitting the model. Table 21 displays the percentage of raters that fit the anchored model with the two gender categories represented in the rows and the three scales in the columns. The results of this analysis reflect the pattern noted in the group-level analyses, females' responses result in smaller values on the STRS and SMES, but larger on the SLES.

Table 21

Percent of Raters with Infit and Outfit Mean Square Statistics Less Than .50

Gender	STRS	SMES	SLES
Males	23%	29%	18%
Females	20%	27%	20%

The multi-facet Rasch model infit/outfit mean square approach to diagnosing halo effect sought to identify raters with infit and outfit statistics that were significantly less than the expected value of 1.00. Engelhard (2002) defines this value as less than .5. A value less than .5 suggests that the rater is not distinguishing among distinct variables and is therefore exhibiting halo. In order to assess these values a FACETS analysis was conducted using response data from the separate gender groups. Those raters with both infit and outfit mean square values less than .5 were counted and percentages of total raters were calculated for males and females. Table 22 displays the percentage of respondents whose fit statistics are less than .5. Males and females are represented on separate rows and the columns represent the three scales under investigation.

Table 22

Percent of Raters with Outfit Mean Square < 1.50 and Standardized Z <1.90 with Item Difficulties Anchored at 0

	STRS		SMES		SLES	
Gender	Mean Square	Standardized *Z*	Mean Square	Standardized *Z*	Mean Square	Standardized *Z*
Males	85%	93%	85%	93%	91%	95%
Females	89%	93%	80%	96%	86%	93%

The multi-facet Rasch model bias-interaction analysis required the calculation of interactions between facets. In this study the rater (student) by trait (item) interaction was investigated. A rater by ratee interaction was not possible because no student rated more than one teacher. The magnitude of the interaction is expressed as a *z*-score. *Z*-scores greater than ± 2 signify statistically significant interactions. A significant interaction occurs when the observed score a rater gives to an item is substantially different than the expected score based on the model. The difference may be positively or negatively oriented, but it suggests that something other than the underlying latent trait is influencing the rating. Table 23 displays the percentage of rater by item interactions that result in a *z*-score of ± 2 in columns 2, 4, and 6, as well as the percentage of individual raters whose responses resulted in two or more significant bias terms in columns 3, 5, and 7.

Table 23

Percent of Significant Bias Terms and Percent of Raters with 2 or More Significant Bias Terms by Gender for Three Scales

	STRS		SMES		SLES	
Gender	Percent of significant bias terms	Percent of raters with 2 or more	Percent of significant bias terms	Percent of raters with 2 or more	Percent of significant bias terms	Percent of raters with 2 or more
Males	2.69%	1.46%	2.85%	1.46%	3.35%	2.43%
Females	3.14%	1.66%	2.30%	0.00%	2.42%	0.55%

Chapter 5: Discussion

The research questions addressed in this study were

1. What are the key areas of teacher performance valued by CES administrators, teachers, and students?
2. In what ways do students conceptualize these areas of valued teacher performance?
3. To what degree do the items derived from student conceptualizations function to produce reliable ratings from which valid conclusions may be drawn about teacher performance?
4. In what ways should the items and scales be revised to improve reliability and validity?
5. To what extent do male and female seminary students exhibit differing degrees of halo effect in their ratings of teachers?

The answers to each of these questions will be addressed in the light of the results outlined in Chapter 4.

Research Question 1

Based on the content analysis of the CES core documents twelve distinct areas of desired teacher performance were identified. While adult professional and volunteer religious educators differentiated between these areas of performance, students were less likely to differentiate between them. As students were asked to identify elements of effective teaching, they were most likely to define it in terms of the teacher caring for and being interested in them as individuals and with the teacher’s ability to help them “feel the Spirit." When asked in the focus group interviews to identify indicators of other areas of teacher performance they struggled to do so. When prompted with a specific area of performance such as "establishing and maintaining an appropriate setting" or "effectively using scripture study skills," they agreed that these were important and were able to identify a few indicators, but they did not attach the same degree of importance to these indicators as did teachers and administrators.

Based on students' response patterns the number of scales was reduced from twelve to three (STRS, SMES, SLES). The results of statistical analyses reflect the likelihood of students to conceptualize effective teaching in terms of far fewer dimensions than were hypothesized based on the content analysis of core teacher training documents. As a matter of fact, the most important dimension of teaching effectiveness from the students' perspective is absent in any explicit form from published CES training materials. It could be argued that student-teacher rapport is implicit in teaching by the Spirit. I would argue, however, that the single most important dimension of teaching effectiveness—according to students—should play a more explicit and prominent role in CES training materials and efforts and in the process of teacher selection. If CES administrators choose to define teaching effectiveness exclusively in terms of the attributes of the teacher, then more attention should be given to the development of student-teacher rapport in its training philosophy and efforts. If, however, CES administrators choose to define teaching effectiveness in terms other than teacher attributes, such as students' understanding, attitudes, and behaviors, then student-teacher rapport may play a smaller role and be of lesser concern as long as student outcomes are achieved.

Research Question 2

Students' conceptualizations of the areas of teacher performance identified in response to research question 1 were outlined in Chapter 4. The relative importance of the various areas of teacher performance and students' abilities to identify indicators of those areas were described in the previous section. While it may have been desirable to select a larger sample for the focus group interviews to promote greater representativeness, the later interviews conducted with students failed to yield new indicators. This suggests that additional interviews were unlikely to have added substantively to the indicators already identified.

Research Question 3

Analyses of student responses to the items generated from student conceptualizations revealed only three 6-item scales that met generally accepted criteria of reliability and validity:

1. the Student-Teacher Rapport Scale,
2. the Scripture Mastery Expectation Scale, and
3. the Spiritual Learning Environment Scale.

For these three scales the classical test theory indictors of item quality (i.e., coefficient alpha, inter-item correlations, and item-total correlations) well exceeded established standards. Additionally, these standards were achieved with only six items in each scale. Many scales used in the social sciences require significantly more items in order to achieve a coefficient alpha of .80. Rasch model reliability estimates also met generally accepted standards with values of .80 or greater.

The construct validity of the three scales is supported by both exploratory and confirmatory factor analyses, as well as the Rasch model analyses, although the relative importance of student-teacher rapport revealed in the focus group interviews with students was reflected in these analyses as well. A three-factor solution resulted in 41% of the variance explained by factor 1. The largest factor loadings on factor 1 are associated with those items constituting the student-teacher rapport scale. Both the qualitative data from the focus group interviews and the quantitative data obtained by the administration and analysis of the scale items support the preeminence of student-teacher rapport in students' perceptions of teacher effectiveness. This finding corroborates the research of Sudweeks (1979) and Elzey (1998).

The present study, however, provides evidence of the construct validity of two additional dimensions of teacher performance not previously found in the research. Both the Scripture Mastery Expectation Scale and the Spiritual Learning Environment Scale produced responses that explain more than the minimum amount of total variance in a factor analysis to provide meaningful information about the variables. The factor loadings produced in the three-factor solution also support the argument for three defensible dimensions in the data with acceptable loadings for each of the items on their respective scales. Additionally, when the scales were analyzed independently of one another, responses to items on the three scales produced acceptable eigenvalues, factor loadings, and percent of variance explained.

The Rasch model analyses also support the construct validity of the STRS, SMES, and SLES. Messick (1989; 1995) identifies six facets of construct validity (content, substantive, structural, generalizability, external, and consequential). E. V. Smith (2004) argues that each of the facets identified by Messick are addressed

> by three general aspects in Rasch measurement: the model requirements and measurement properties if the data fit the model, the order of items and persons on a common linear scale along with the associated individual SE, and the fit of the items and persons to the model requirements. (p. 100)

However, a few areas of concern were diagnosed. The most serious concern lies in the response categories, where a uniform distribution of observations across response categories was not achieved. The vast majority of responses utilized the upper three categories (*undecided*, *agree*, *strongly agree*). Students were highly unlikely to select *strongly disagree* or *disagree*. The consequence of this pattern was to restrict the degree to which items and persons were separated along the difficulty/performance continuum, thus reducing item and person reliability and item fit statistics. Specific recommendations for addressing this area of concern were presented in Chapter 4. In summary, the proposed revisions seek to stretch the upper end of the continuum to better align item difficulties and person measures, and to better separate persons producing better reliability and fit statistics. Restricted item separation also relates to the diagnosis of halo effect and will be elaborated in response to research question 5.

The confirmatory factor analysis produced only marginal fit statistics, although it should be noted that the assumptions of the model were very demanding (i.e., no correlations between variables). The proposed revisions to eliminate the secondary factor loading of item 1b on the SLES and the correlated errors between items particularly on the SMES were described in detail in Chapter 4.

Although a sufficient number of responses was obtained to meet generally accepted criteria for the exploratory and confirmatory factor analyses and the Rasch model analyses, the responses may not be representative because a convenience sample was used. Therefore, the findings of this study should be corroborated by additional research employing randomized sampling techniques.

Research Question 4

While demonstrating reasonable adherence to accepted standards of scale performance, the study revealed areas in which scale function can be improved. Specific recommendations have already been detailed in Chapter 4. In summary, it is proposed that revisions to individual items be made in an attempt to eliminate or reduce significant secondary factor loadings (e.g., item 1b) and to eliminate or reduce correlated error between items (e.g., items constituting the SMES). It is also proposed that the response categories be revised to stretch the upper end of the continuum and to better align persons and items in order to obtain better item and person separation and therefore improved reliability, fit statistics, and ultimately better construct validity.

Research Question 5

This study employed several classical (intervariable correlation, intraratee variance, and intervariable factor structure) and Rasch model (group-level and individual-level indicators) approaches to diagnose halo effect. The intervariable correlation approach examined the correlation between scores on each pair of scales (i.e., STRS/SMES, STRS/SLES, SMES/SLES). Large correlations are considered evidence of halo effect. The correlation estimates for each pair of scales was greater for males than for females (see Table 19). The second approach examined the intraratee variance across all scales. Small variance estimates are indicative of halo. Although both variance estimates were small (see Table 20), the mean variance for males was less than the mean variance for females. The third approach examined the factor structure of the data using factor analysis. Where a single dominant factor accounts for a large percentage of variance explained, halo effect is considered to be at work. Table 21 shows that the percent of variance explained by the first factor was substantially greater for males than it was for females. Each of these approaches suggests that males exhibited halo to a greater degree than females.

The group- and item-level indicators used to diagnose halo effect with the Rasch model were less definitive with regards to gender differences than the classical approaches. The group-level indicators for both males and females included chi-square statistics with their respective *p*-values, trait separation ratio, trait separation index, and reliability of the trait separation index. All chi-square statistics produced significant *p*-values suggesting no group-level halo. The other group-level indicators, however, provided evidence of halo effect. In those instances, a comparison of the statistical indicators by gender reveals that males exhibit halo to a greater degree than do females on the STRS and SMES. The SLES, on the other hand, produced statistical indicators that suggest females exhibited halo to a slightly greater degree than males. However, the gender differences were much less pronounced on the SLES than on the other scales.

The individual-level indicators included infit and outfit mean square, like-difficulty anchoring, and bias-interaction. The infit and outfit mean square analysis revealed more males exhibiting halo on the STRS and SMES, but somewhat fewer on the SLES (see Table 25). The results of the like-difficulty item anchoring approach was mixed (see Table 26). On the STRS males were less likely than females to exhibit halo based upon the mean square value and equally as likely to exhibit halo based on the standardized *z*. On the SMES males were more likely to exhibit halo based upon the mean square and less likely when based upon the standardized *z*. On the SLES males were more likely to exhibit halo based on both the mean square and standardized *z*. Using the bias-interaction approach, males were less slightly likely to exhibit halo on the STRS and somewhat more likely to exhibit halo on the SMES and SLES (see Table 27).

In summary, the classical approaches to halo effect diagnosis suggested, in every instance, that male seminary students exhibit halo to a greater degree than do their female counterparts. The Rasch model approaches were less consistent in diagnosing greater halo effect among males than females, but the evidence does point toward differential halo effect by gender, with males exhibiting halo more than females.

If the source of the halo is rapport, as suggested by the factor analyses, how do we interpret the diagnosis of halo in the responses to the STRS? In other words, how can we argue that rapport is influencing responses to items designed to measure rapport causing the ratings to be artificially increased or decreased? When items and persons are poorly aligned and item difficulty is homogenous, what may appear to be halo may simply be a lack of variability in the responses. To reason that the indicators necessarily reflect halo would not be valid.

What may be reasonably concluded is that the three scales analyzed in this study produced responses that meet generally accepted criteria for reliability and construct validity. The analyses provided several insights into areas of potential improvement in scale function, although the proposed solutions to these areas of weakness must be evaluated through additional administrations of the scales. In general, males exhibited less variability in their responses than did females. The evidence supports the conclusion that female students in LDS seminaries discriminate more meaningfully between areas of teacher performance than do their male counterparts. As a result, decisions based upon female raters' responses would be considered more valid than decisions based upon male raters' responses. Consideration must, therefore, be given to the gender of raters when making judgments about teacher performance. Additional research is needed to determine if greater variability and, therefore, precision can be achieved by making the changes to items and response categories proposed in this study and the effect of that increased variability on the diagnosis of halo effect and the assessment of differences between males and females.

Having answered the five specific research questions posed in this study, attention will now be turned to broader issues such as the contributions of this study, the implications of this study for the practice of teacher evaluation and instructional design, as well as areas requiring additional research.

Study Contributions

This study provides a developmental framework for scale construction that integrates Classical Test Theory, Item Response Theory, and factor analytic techniques in a way that leads to defensibly reliable data from which valid conclusions may be drawn. It also establishes a firm basis for three scales that measure traits of importance to CES that meet widely acceptable psychometric standards, whereas past efforts were limited to only one trait—rapport. Finally, this study provides evidence that males exhibit halo to a greater degree than do females among secondary students on the traits examined.

Although not generalizable to other traits or other instructional settings, it raises a caution about drawing conclusions about teachers from ratings produced by differing gender distributions.

Implications for Teacher Evaluation

CES religious educators are expected to "live the gospel, teach effectively, and administer appropriately" (Church Educational System, 1994, p. 4). Judgments about the degree to which religious educators live the gospel are left primarily to the individual and their ecclesiastical leaders. Evaluating the ability of religious educators to teach effectively and to administer appropriately, however, is essential to the progress of CES. This study establishes the STRS, SMES, and SLES as defensible measures of three traits of importance to CES related to teaching effectively. However, it also highlights the relatively limited scope of traits about which secondary students can provide meaningful information. This study suggests that other traits related to the current definition of teaching effectiveness in CES cannot be validly judged based on student ratings of teachers.

Because student judgments about traits related to teaching effectiveness are limited, it is critical that other sources of information about teacher performance be considered when making judgments about areas of needed teacher improvement, hiring, or termination. These sources include training personnel, peer teachers, supervisors, and other stakeholders.

Another approach to evaluating teaching effectiveness is to define effectiveness in terms of the achievement of student outcomes. A student outcomes approach would define teaching effectiveness in terms of a variety of student cognitive, affective, and behavioral outcomes that reflect the objectives and philosophy of the organization. This approach is based on the assumption that teaching implies learning. The testing of that assumption would require, however, the assessment of student learning not currently practiced by CES in any formal or systematic fashion.

Implications for Instructional Design

Conclusions about the effectiveness of instruction, whatever the setting or the instructional design model, are based upon evidence that objectives have been achieved. Despite the criticism of some scholars, when carefully designed, developed, and implemented, rating scales provide a basis on which to make valid judgments about instruction and the design models upon which the instruction is based. Nevertheless, threats to the validity of conclusions about instructional interventions abound. Instructional designers should be aware of these threats and take appropriate steps to diagnose and mitigate them as they assess. In particular, this study notes the requirements of fundamental measurement when applying statistical analyses to assessment data and highlights the potential threat of halo error in ratings.

Future Research

A number of practical and theoretical questions evolve from this study. For example, do the twelve scales developed in this study function as desired when more mature raters (e.g., adult students, peer teachers, supervisors, trainers) are employed in the rating process? Significant effort went into the construction of the original twelve scales. Although not appropriate for use with secondary student raters, the scales may function well when used with adults. The approach to scale diagnostics used in this study could be used to assess issues of reliability and validity with data generated by other rater groups.

Additionally, this study focused on the reliability and construct validity of measures of teacher performance while paying relatively little attention to issues of administration and reporting. Because of this limitation, other facets of a validity argument have not been addressed such as Messick’s (1989, 1995) notion of consequential validity. This tact was based on a conscious decision to limit the scope of the study to the construction of instruments in which we can have confidence; and then, secondarily (outside the scope of the present study) to focus on how to administer the instrument and report the findings. Additional research is needed, therefore, to identify the best administration practices and to explore meaningful ways of communicating the ratings to various audiences. Of particular concern is the reporting of Rasch model data that is expressed in logits. Most audiences will not be familiar with log odds ratios or their interpretation. Substantial experimentation will be required to determine how best to standardize administration and reporting efforts in a manner that will lead to improved teacher performance and decision making by teacher and administrators.

Other possible research questions have to do with the degree to which potentially improved item and scale functioning, obtained by implementation of the recommendations made in this study, impact the diagnosis of halo effect and gender-based differences in halo. In other words, if the proposed changes to items and response categories result in improved scale function, will halo still be evident? Will males still be more likely to exhibit halo than females? Can illusory halo be mitigated to a substantial degree by more rigorous standards for scale performance? Significant questions remain as to how researchers can meaningfully differentiate between restricted variability in ratings and halo effect.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1968). *Psychological testing* (3 ed.). New York: Macmillan.

Anderson, E. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*(32), 283-301.

Anderson, E. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*(34), 42-54.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scores with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Barnette, W. (1976). *Readings in psychological tests and measurements* (3rd ed.). Baltimore: Williams & Wilkins.

Bartlett, C. (1966). The use of an internal discrimination index in forced-choice scale construction. *Personnel Psychology, 19*, 209-213.

Bartlett, C. (1983). What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *Journal of Applied Psychology, 68*, 218-226.

Bearden, W. O., & Netemeyer, R. G. (1998). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research.* Thousand Oaks, CA: Sage.

Beerens, D. R. (2000). *Evaluating teachers for professional growth: Creating a culture of motivation and learning.* Thousand Oaks, CA: Corwin Press.

Bentler, P. M. (1990). Comparative fit indices in structural equation models. *Psychological Bulletin, 107*, 238-246.

Berman, J., & Kenny, D. (1976). Correlational bias in observer ratings. *Journal of Personality and Social Psychology, 34*, 263-273.

Bernardin, H. J., & Pence, E. C. (1979). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60-66.

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*, 64-69.

Binet, A. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Annee Psychologique, 11*, 245-336.

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems.* Randolph Air Force Base, Texas: USAF School of Aviation Medicine.

Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability.* Randolph Air Force Base, Texas: USAF School of Aviation Medicine.

Birnbaum, A. (1958b). *On the estimation of mental ability.* Randolph Air Force Base, Texas: USAF School of Aviation Medicine.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology, 25*, 185-199.

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

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 64*, 64-69.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*, 233-252.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bryant, F. B. (2000). Assessing the validity of measurement. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 99-146). Washington, DC: American Psychological Association.

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). Washington, DC: American Psychological Association.

Cattel, R. B. (1966). The meaning and strategic use of factor analysis. In R. B. Cattel (Ed.), *Handbook of multivariate experimental psychology* (pp. 174-243). Chicago: Rand McNally.

Church Educational System (1994). *Teaching the gospel: A handbook for CES teachers and leaders*. Salt Lake City, UT: The Church of Jesus Christ of Latter-day Saints.

Church Educational System (2000). *Teaching the gospel: A CES training resource for teaching improvement*. Salt Lake City, UT: Intellectual Reserve, Inc.

Church Educational System (2003a). *Administering appropriately: A handbook for CES teachers and leaders*. Salt Lake City, UT: Intellectual Reserve, Inc.

Church Educational System (2003b). *A current teaching emphasis for the Church Educational System*. Salt Lake City, UT: Intellectual Reserve, Inc.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in scale development. *Psychological Assessment, 7*(3), 309-319.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*(2), 161-173.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218-244.

Cudeck, R. (2000). Exploratory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 265-296). San Diego, CA: Academic Press.

Currey, S., Callahan, L., & DeVellis, R. F. (2002). *Five-item Rheumatology Attitudes Index (RAI): Disadvantages of a single positively worded item.* Chapel Hill, NC: Thurston Arthritis Research Center.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* Princeton, NJ: Educational Testing Service.

DeVellis, R. F. (2003). *Scale development: Theory and applications.* (2 ed.). Thousand Oaks, CA: Sage.

DeVellis, R. F., & Callahan, L. (1993). A brief measure of helplessness: The helplessness subscale of the Rheumatology Attitudes Index. *Journal of Rheumatology, 20*, 866-869.

Dickinson, T. L., & Tice, T. E. (1977). The discriminant validity of scales developed by retranslation. *Personnel Psychology, 30*, 255-268.

Dillman, D. (2000). *Mail and internet surveys.* New York: John Wiley and Sons.

Drucker, P. F. (1998). Management's new paradigms. *Forbes, 162,* 152-176.

Duncan, O. (1984). *Notes on social measurement: Historical and critical.* New York: Russell Sage.

Edwards, A. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.

Elzey, R. (1998). *The construct validity of the principles of edification as measures of edifying teaching in the LDS Church Educational System.* Unpublished Dissertation, Brigham Young University, Provo, Utah.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Lawrence Erlbaum.

Farson, R. (1996). *Management of the absurd: Paradoxes in leadership.* New York: Simon & Schuster.

Feeley, T. H. (2002). Comment on halo effects in rating and evaluation research. *Communication Education, 51* (3), 225-236.

Fink, A. (1995). *The survey kit 2: How to ask survey questions.* Thousand Oaks, CA: Sage.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests (Introduction to mental test theory.).* Bern, Switzerland: Huber.

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*, 238.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*(419-429).

Floyd, F. J., & Widaman, K. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19*, 149-161.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Hambleton, R. K., & Swaminathan, H. (1985). A look at psychometrics in the Netherlands. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 40*(7), 446-451.

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

Howell, G. B. (1995). *Teacher evaluation: An examination of the congruence of purpose and practice in LDS CES.* Unpublished dissertation, Arizona State University, Tempe, AZ.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403-424.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Karabatsos, G. (1999). *Axiomatic measurement theory as a basis for model selection in item-response theory.* Paper presented at the 32nd annual conference of the Society for Mathematical Psychology, Santa Cruz, CA.

Kavanagh, M. J., MacKinney, A. C., & Wollins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin, 75*, 34-49.

Keaveny, T. J., & McGann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. *Journal of Applied Psychology, 60*, 695-703.

Kenny, D., & Berman, J. (1980). Statistical approaches to the correction of correlational bias. *Psychological Bulletin, 88*, 288-295.

Kerlinger, F. (1986). *Foundations of behavioral research.* New York: Holt, Rinehart and Winston.

Kiker, D. S., & Motowidlo, S. J. (1998). Effects of rating strategy on interdimensional variance, reliability, and validity of interview ratings. *Journal of Applied Psychology, 83*, 763-768.

King, L., Hunter, J., & Schmidt, F. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 65*, 507-516.

Kraut, A. I. (1975). Prediction of managerial success by peer and training-staff ratings. *Journal of Applied Psychology, 60*, 14-19.

Lance, C. E., Fisicaro, S. A., & LaPointe, J. A. (1990). An examination of negative halo error in ratings. *Educational and Psychological Measurement, 50*, 545-554.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 7*, 356-366.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(No. 140).

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2004a). FACETS: Rasch measurement computer program. (Version 3.54). Chicago: Winsteps.com.

Linacre, J. M. (2004a). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2004b). *A user's guide to FACETS: Many-facet Rasch analysis.* Chicago: MESA Press.

Linacre, J. M. (2004b). WINSTEPS: Rasch measurement computer program. (Version 3.53). Chicago: Winsteps.com.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

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

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluation by their students. *Journal of Educational Psychology, 71*, 149-160.

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

Maughan, D. M. (1994). *Teachers' attitudes toward accountability in a religious education setting.* Unpublished dissertation, Arizona State University, Tempe, AZ.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355-383.

Moritsch, B. G., & Suter, W. N. (1988). Correlates of halo error in teacher evaluation. *Educational Research Quarterly, 12*, 29-34.

Murphy, K. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology, 67*, 161-164.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619-624.

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*, 218-225.

Myford, C. M., & Wolfe, E. W. (2004a). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 460-517). Maple Grove, MN: JAM Press.

Myford, C. M., & Wolfe, E. W. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 518-574). Maple Grove, MN: JAM Press.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage.

Netemeyer, R. G., Boles, J. S., & McMurrian, R. C. (1996). Development and validation of Work-Family Conflict and Family-Work Conflict scales. *Journal of Applied Psychology, 81*, 400-410.

Newcomb, T. (1931). An experiment designed to test the validity of a rating technique. *Journal of Educational Psychology, 22*, 279-289.

Obermiller, C., & Spangenberg, E. R. (1998). Development of a scale to measure consumer skepticism toward advertising. *Journal of Consumer Psychology, 7*(2), 159-186.

Payne, S. (1951). *The art of asking questions.* San Francisco: Jossey-Bass.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices.* Thousand Oaks, CA: Corwin Press.

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education, 40*(1), 61-86.

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

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Criteria for scale selection and evaluation. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 1-15). San Diego, CA: Academic Press.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Saal, F. E., Downey, R. G., & Layhey, M. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428.

Smith, E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement: Theory, models and applications* (pp. 93-122). Maple Grove, MN: JAM Press.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541-565.

Smith, R. M. (1998, December). *Ordered response category analysis: A best practice model.* Paper presented at the Winter 1998 Midwestern Objective Measurement Seminar, Chicago.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 73-92). Maple Grove, MN: JAM Press.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-426.

Spector, P. E. (1992). *Summated rating scale construction: An introduction.* Newbury Park, CA: Sage.

Stanley, S. J., & Popham, W. J. (Eds.) (1988). *Teacher evaluation: Six prescriptions for success.* Alexandria, VA: Association for Supervision and Curriculum Development.

Stone, M. H. (2004). Substantive scale construction. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 201-225). Maple Grove, MN: JAM Press.

Sudweeks, R. R. (1979). *An evaluation of the SES: An instrument to collect student ratings of seminary teachers.* Salt Lake City, UT: LDS Church Educational System.

Thomas, S. J. (2004). *Using web and paper questionnaires for data-based decision making: From design to interpretation of the results.* Thousand Oaks, CA: Corwin Press.

Thorndike, E. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25-29.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review, 34*, 278-286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21*, 384-400.

Thurstone, L. L. (1927c). Psychophysical analysis. *American Journal of Psychology, 38*, 368-389.

Thurstone, L. L. (1927d). Three psychophysical laws. *Psychological Review, 34*, 424-432.

Thurstone, L. L. (1931). The measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26*, 249-269.

Thurstone, L. L. (1935). *Vectors of mind.* Chicago: University of Chicago Press.

Tian, K. T., Bearden, W. O., & Hunter, G. L. (2001). Consumers' Need for Uniqueness: Scale development and validation. *Journal of Consumer Research, 28*(June), 50-66.

van der Linden, W., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item-response theory*. New York: Springer.

Warwick, D., & Lininger, C. (1975). *The sample survey: Theory and practice.* New York: McGraw-Hill.

Weiss, D. J. (Ed.). (1983). *New horizons in testing.* New York: Academic Press.

Wells, F. (1907). *A statistical study of literary merit*. Unpublished manuscript.

Wherry, R. J., & Bartlett, C. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*, 521-551.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Wolf, R. (1973). How teachers feel toward evaluation. In E. House (Ed.), *School evaluation: The politics and process* (pp. 156-168). Berkeley, CA: McCutchan.

Wothke, W. (1996). Models for multitrait-multimethod matrix analysis. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and applications* (pp. 7-56). Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 63-104). Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*, 281-294.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Wright, B. D., & Stone, M. H. (2003). *Managing observations, inventing constructs, crafting yardsticks, examining fit.* Chicago: Phaneron Press.

Yamaguchi, J. (1997). Positive vs. negative wording. *Rasch Measurement Transactions, 11*, 567.

Appendix A

Focus Group Protocol

Welcome

- Introduce moderator and assistant moderator

Topic overview

- We want to discuss what makes for *effective gospel teaching.*
- You were selected to participate because we would like feedback from students, teachers, and administrators involved in seminary and institute classes in different parts of the country.
- Your comments will be used to help us write questions to be used in seminary and institute teacher evaluations. These evaluations will be tools that teachers can use to improve their teaching. Administrators can also use the evaluations to make decisions about training.
- [Read informed consent document for focus group participants]

Ground rules

- There are no right or wrong answers, only differing points of view.
- We're tape recording our discussion, so we ask that only one person speak at a time.
- You don't need to agree with others, but you must listen respectfully as others share their views. We are just as interested in negative comments as positive comments, at times the negative comments are the most helpful. You may be assured of complete confidentiality.
- Please turn off your cellular phones or pagers.
- My role as moderator will be to guide the discussion, but we encourage you to talk to each other during the discussion.

Questions

- Do you have any questions about the purpose of this discussion or the ground rules?
- Think back to the most effective seminary or institute teacher you have seen. What do they do that makes them such effective gospel teachers? Write down your answers and then let's discuss them (need paper & pencils). [adapt questions to teacher and administrator groups]
- What are the characteristics of poor gospel teachers?
- What have your teachers done that has most influenced you to live a righteous life?
- Describe your best experiences in seminary and/or institute. Give specific examples.
- Describe your worst experiences. Give specific examples.
- Introduce the core competencies identified by the committee; transition to more focused structure:
  - Teachers should "teach students the gospel of Jesus Christ." What does that mean to you? Describe a teacher that does this well/poorly.

  - Teachers should “teach by the Spirit.” How can you tell if a teacher is teaching by the Spirit?
  - Teachers should “teach by example.” How does a teacher teach by example? Describe a teacher who does this well?
  - Teachers should “establish and maintain an appropriate setting” for teaching the gospel. What do you think an appropriate setting is? What should the physical setting be like? What should the spiritual setting be like? What should the rules be for class discussion and interaction?
  - Teachers should “help students accept their responsibility for gospel learning.” Do you think you are taking responsibility for your own gospel learning? How do teachers help you to do that?
  - Teachers should effectively decide what and how to teach. Can you recognize if a teacher has been effective in deciding what and how to teach? How can you recognize it?
  - Teachers should “effectively use scripture study and teaching skills.” Are you able to recognize when a teacher does that? How do good teachers do it?
  - Teachers should develop good relationships with students. How would you describe a good relationship between a teacher and a student? How does the teacher act? What do teachers do that make it hard for you to relate to them? What causes you to place your trust in your teacher? What causes you to lose confidence in them?
  - Teachers should “prepare young people for effective church service.” Have your teachers helped to prepare you for church service? If so, how? If not, why? What could they do better?
- Are there other attributes that teachers should have that we haven’t mentioned? Have we missed anything?

Conclusion

- Summarize salient points with confirmation.
- Review the purpose of the session and ask if anything has been missed.

Thank the participants and dismiss.

Appendix B

Informed Consent Form for Focus Group Participants

The objective of this research is to develop teacher evaluations that CES students, teachers, and administrators can use to improve teaching effectiveness. Eric Rogers, a research analyst for CES and a graduate student at BYU, is conducting the study. You were selected for participation because of your involvement in CES programs.

You will be asked to spend approximately 50 minutes in answering questions and participating in a group discussion. Part of this time may be class time. You will be asked questions about what you consider effective teaching in a religious education setting. Your group will include 4-15 people in addition to the moderator and his assistant. The purpose of the focus group is to get you talking with others about what makes for effective gospel teaching. Your comments will be recorded and transcribed, but your identity will not be recorded or reported in any fashion.

There are minimal disadvantages for participation in this study. You should feel comfortable sharing your honest feelings. Rude comments by focus group participants will not be allowed. Differing points of view, however, are encouraged. The moderator will make sure that you are not teased or made fun of for your comments. He will also make sure that the discussion doesn't take longer than what you have agreed to. There are no known benefits to you from participation in the study.

Participation in this research is voluntary. You have the right to refuse to participate. Strict confidentiality will be maintained. No individual identifying information will be disclosed. All data collected in this research study will be stored in a secure area and access will only be given to personnel associated with the study.

In order to participate you must give your signed consent to be a research subject. If you are under 18 years of age you must also obtain a parent's signature. If you are willing to be a research subject please sign below.

____ ____________

Participant's signature Date

____ ____________

Parent's signature (if participant is under 18 years of age) Date

If you have any questions regarding this research project, you may contact:

Eric Paul Rogers
50 East North Temple, 821
Salt Lake City, Utah 84150
801-240-7832
RogersEP@ldschurch.org

If you have questions regarding your rights as a participant in a research project, you may contact:

Dr. Shane S. Schulthies
Chair of the Institutional Review Board
120 B RB, Brigham Young University
Provo, Utah 84602
801-422-5490

Appendix C

Consent To Be a Research Subject For Seminary Instructors

The objective of this research is to develop teacher evaluations that CES students, teachers, and administrators can use to improve teaching effectiveness. Eric Rogers, a research analyst for CES and a graduate student at BYU, is conducting the study. You were selected for participation because of your involvement in CES programs.

There are minimal disadvantages for participation in this study. The main disadvantage is the time needed to complete the survey. There are no known benefits to you from participation in the study.

Participation in this research is voluntary. You have the right to refuse to participate. Strict confidentiality will be maintained. No individual identifying information will be disclosed. All data collected in this research study will be stored in a secure area and access will only be given to personnel associated with the study.

The purpose of the study is not to draw conclusions about your effectiveness as a teacher, but rather to gather information about how the items on the survey function statistically. The data resulting from the administration of the survey in your classes will be used to refine individual items and groups of items, or scales.

In order to participate you must give your signed consent to be a research subject. By signing this form you consent to allow students, peers, and supervisors to respond to the items on the survey based on their observations of your teaching practices.

____ _____________
Instructor's signature Date

If you have any questions regarding this research project, you may contact:

Eric Paul Rogers
50 East North Temple, 9th Floor
Salt Lake City, Utah 84150
801-240-7832
RogersEP@ldschurch.org

If you have questions regarding your rights as a participant in a research project, you may contact:

Dr. Shane S. Schulthies
Chair of the Institutional Review Board
120 B RB, Brigham Young University
Provo, Utah 84602
801-422-5490

Appendix D

Student Consent Form for Scale Administration for Student Survey Participants

The objective of this research is to develop teacher evaluations that CES students, teachers, and administrators can use to improve teaching effectiveness. Eric Rogers, a research analyst for CES and a graduate student at BYU, is conducting the study. You were selected for participation because of your involvement in CES programs.

You will be asked to spend up to 30 minutes answering questions about your current teacher. Your name will not be recorded and your answers will not be reported to anyone.

There are minimal disadvantages for participation in this study. The main disadvantage is the time needed to complete the survey. There are no known benefits to you from participation in the study.

Participation in this research is voluntary. You have the right to refuse to participate. Strict confidentiality will be maintained. No individual identifying information will be disclosed. All data collected in this research study will be stored in a secure area and access will only be given to personnel associated with the study.

In order to participate you must give your signed assent to be a research subject and one of your parents must also give signed consent (if you are under 18). If you are willing to be a research subject please sign and obtain a parental signature below if required.

____ ____________
Student signature Date

____ ____________
Parent’s signature (if participant is under 18) Date

If you have any questions regarding this research project, you may contact:

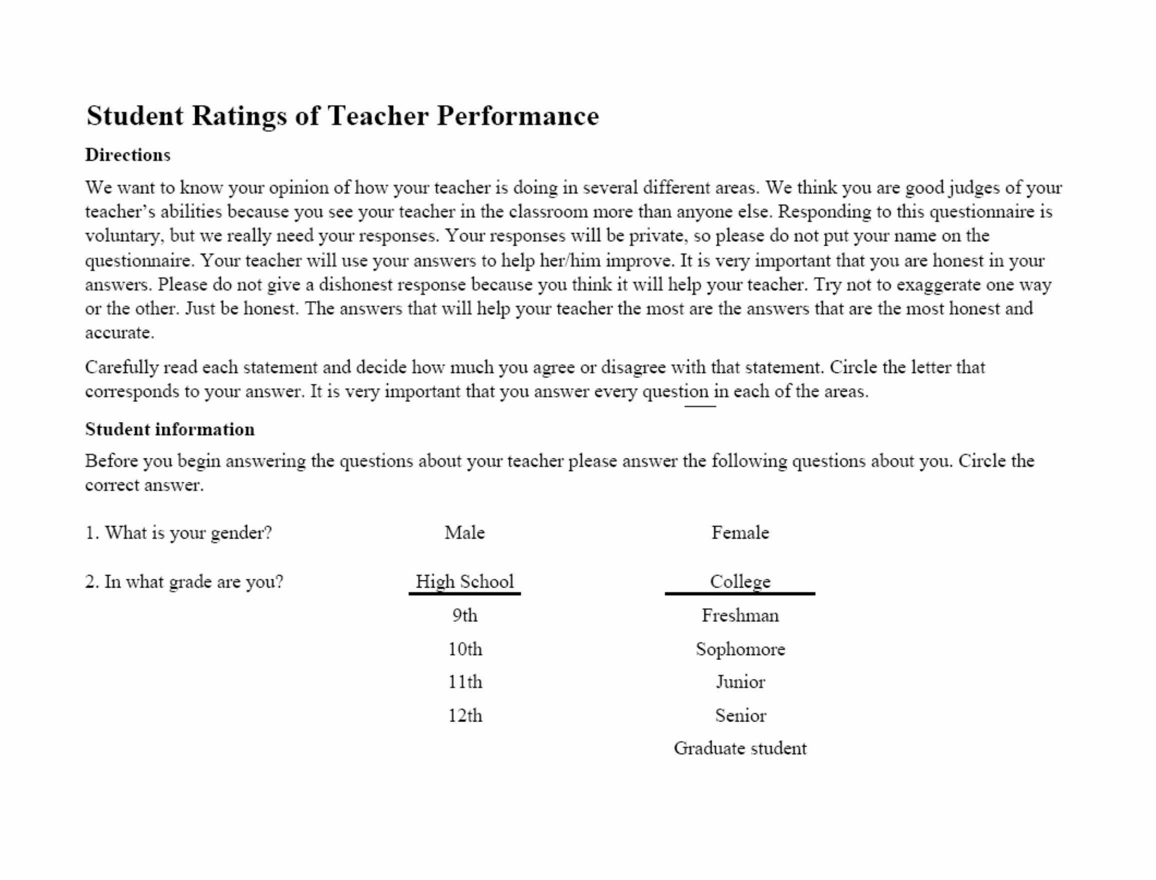
Eric Paul Rogers
50 East North Temple, 9th Floor
Salt Lake City, Utah 84150
801-240-7832
RogersEP@ldschurch.org

If you have questions regarding your rights as a participant in a research project, you may contact:

Dr. Shane S. Schulthies
Chair of the Institutional Review Board
120 B RB, Brigham Young University
Provo, Utah 84602
801-422-5490

Appendix E

Student Ratings of Teacher Performance



Student Ratings of Teacher Performance

Directions

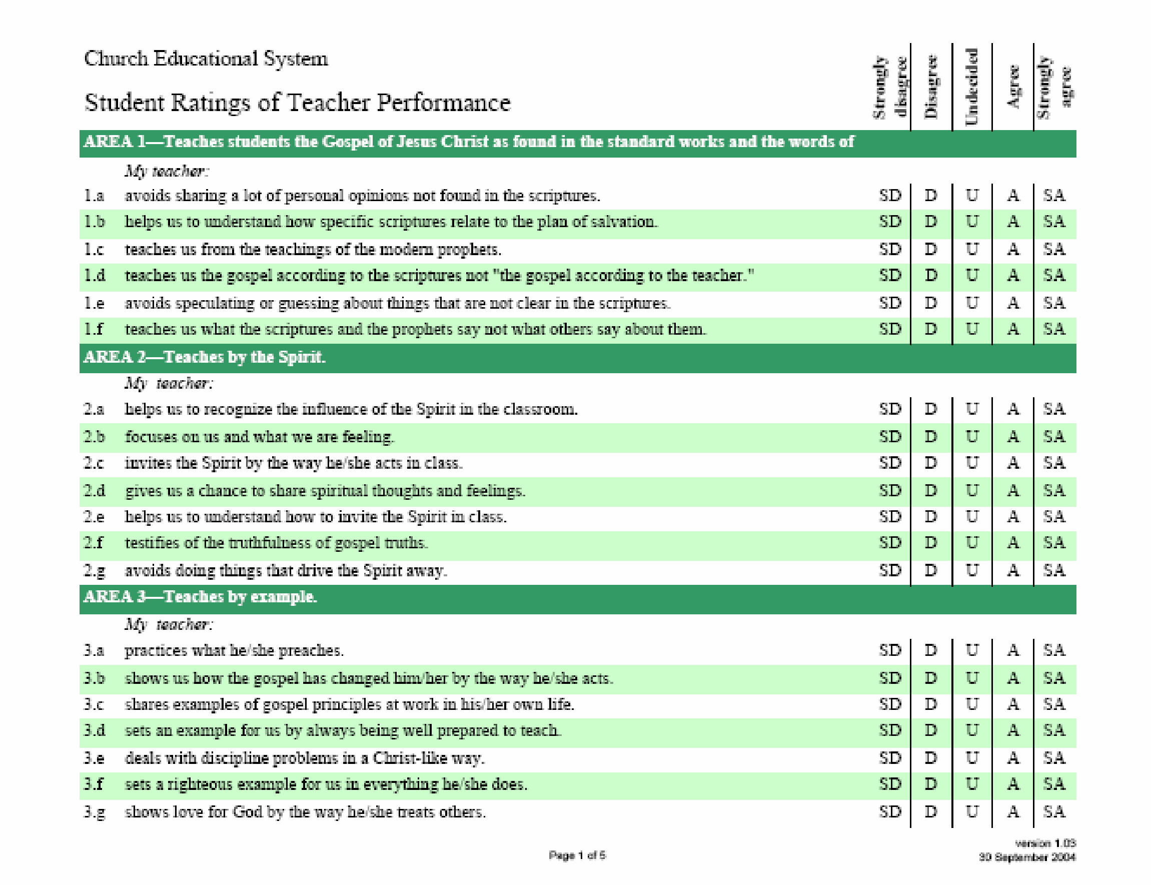
We want to know your opinion of how your teacher is doing in several different areas. We think you are good judges of your teacher's abilities because you see your teacher in the classroom more than anyone else. Responding to this questionnaire is voluntary, but we really need your responses. Your responses will be private, so please do not put your name on the questionnaire. Your teacher will use your answers to help her/him improve. It is very important that you are honest in your answers. Please do not give a dishonest response because you think it will help your teacher. Try not to exaggerate one way or the other. Just be honest. The answers that will help your teacher the most are the answers that are the most honest and accurate.

Carefully read each statement and decide how much you agree or disagree with that statement. Circle the letter that corresponds to your answer. It is very important that you answer every question in each of the areas.

Student information

Before you begin answering the questions about your teacher please answer the following questions about you. Circle the correct answer.

1. What is your gender?	Male	Female
2. In what grade are you?	High School	College
	9th	Freshman
	10th	Sophomore
	11th	Junior
	12th	Senior
		Graduate student



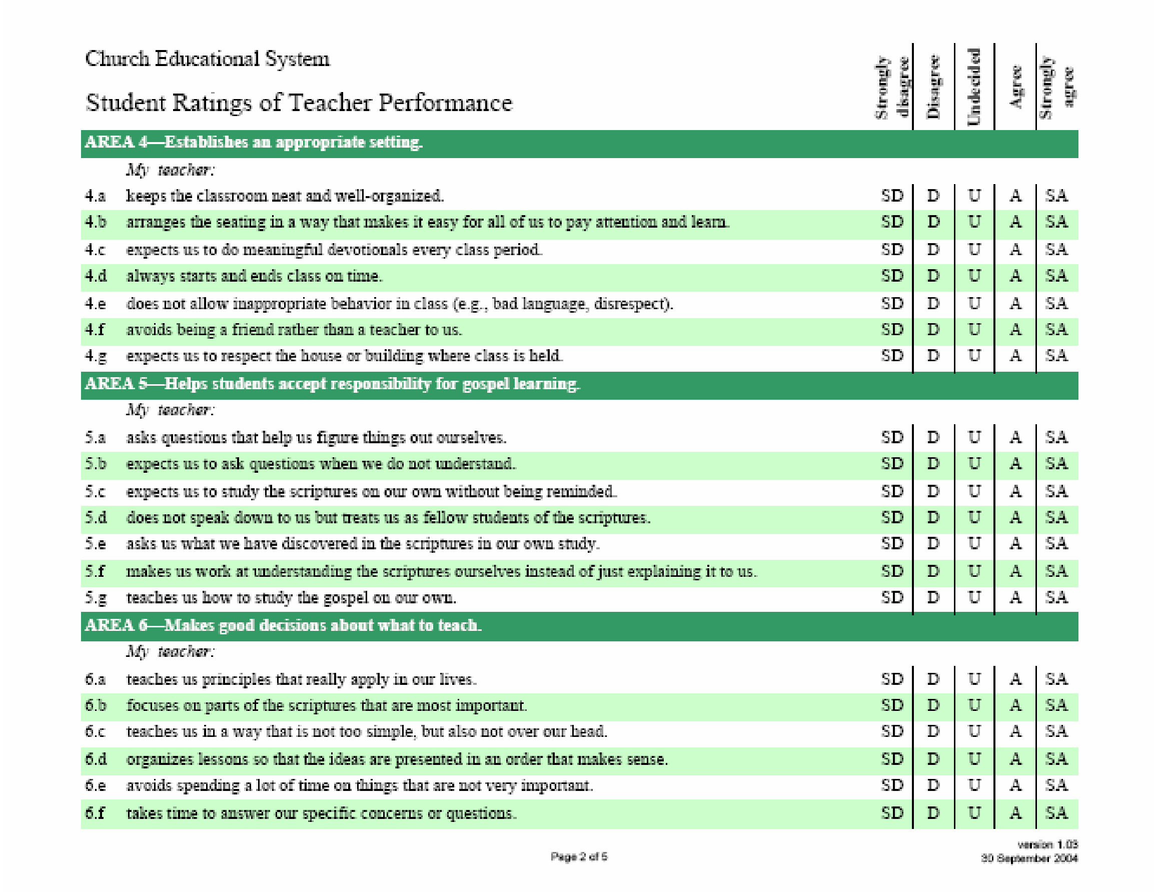
Church Educational System

Student Ratings of Teacher Performance

		Strongly disagree	Disagree	Undecided	Agree	Strongly agree
AREA 1—Teaches students the Gospel of Jesus Christ as found in the standard works and the words of						
	My teacher:					
1.a	avoids sharing a lot of personal opinions not found in the scriptures.	SD	D	U	A	SA
1.b	helps us to understand how specific scriptures relate to the plan of salvation.	SD	D	U	A	SA
1.c	teaches us from the teachings of the modern prophets.	SD	D	U	A	SA
1.d	teaches us the gospel according to the scriptures not "the gospel according to the teacher."	SD	D	U	A	SA
1.e	avoids speculating or guessing about things that are not clear in the scriptures.	SD	D	U	A	SA
1.f	teaches us what the scriptures and the prophets say not what others say about them.	SD	D	U	A	SA
AREA 2—Teaches by the Spirit.						
	My teacher:					
2.a	helps us to recognize the influence of the Spirit in the classroom.	SD	D	U	A	SA
2.b	focuses on us and what we are feeling.	SD	D	U	A	SA
2.c	invites the Spirit by the way he/she acts in class.	SD	D	U	A	SA
2.d	gives us a chance to share spiritual thoughts and feelings.	SD	D	U	A	SA
2.e	helps us to understand how to invite the Spirit in class.	SD	D	U	A	SA
2.f	testifies of the truthfulness of gospel truths.	SD	D	U	A	SA
2.g	avoids doing things that drive the Spirit away.	SD	D	U	A	SA
AREA 3—Teaches by example.						
	My teacher:					
3.a	practices what he/she preaches.	SD	D	U	A	SA
3.b	shows us how the gospel has changed him/her by the way he/she acts.	SD	D	U	A	SA
3.c	shares examples of gospel principles at work in his/her own life.	SD	D	U	A	SA
3.d	sets an example for us by always being well prepared to teach.	SD	D	U	A	SA
3.e	deals with discipline problems in a Christ-like way.	SD	D	U	A	SA
3.f	sets a righteous example for us in everything he/she does.	SD	D	U	A	SA
3.g	shows love for God by the way he/she treats others.	SD	D	U	A	SA

Page 1 of 5

version 1.03
30 September 2004



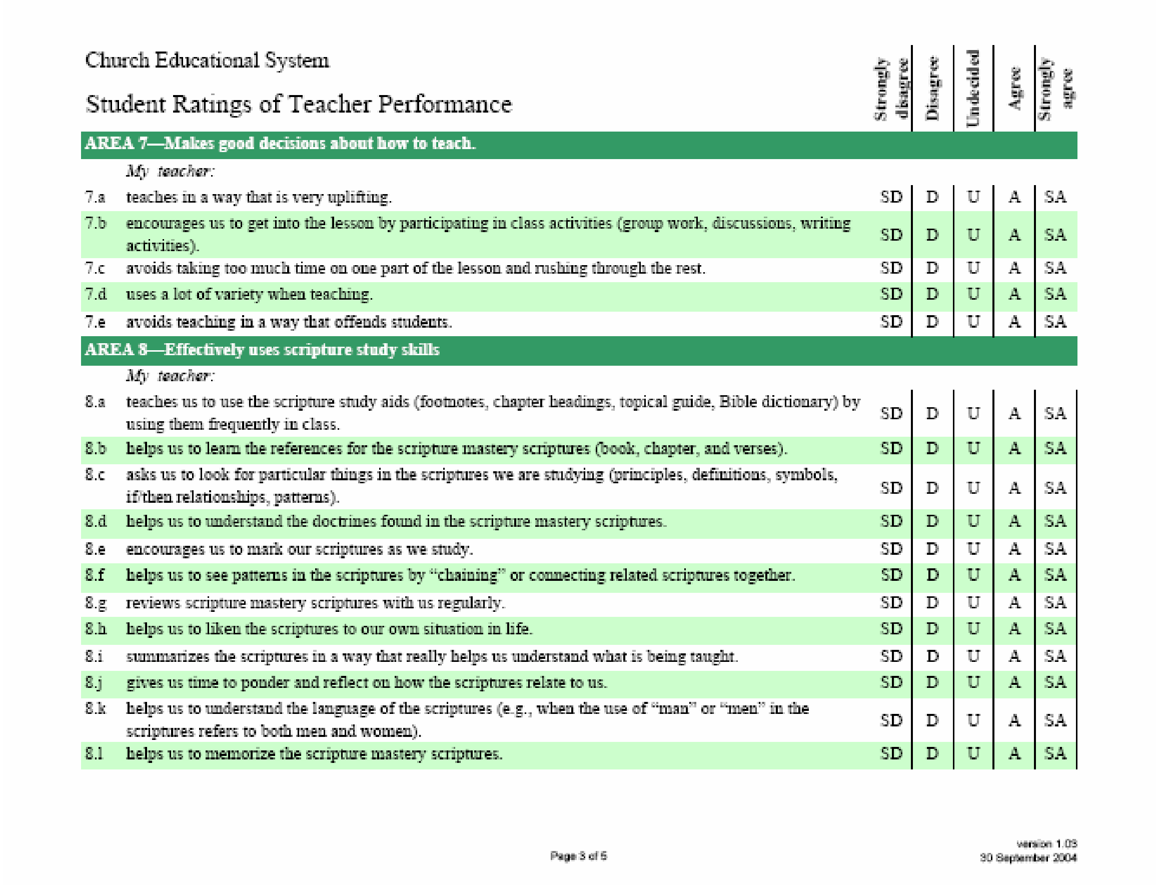
Church Educational System

Student Ratings of Teacher Performance

		Strongly disagree	Disagree	Undecided	Agree	Strongly agree
AREA 4—Establishes an appropriate setting.						
	My teacher:					
4.a	keeps the classroom neat and well-organized.	SD	D	U	A	SA
4.b	arranges the seating in a way that makes it easy for all of us to pay attention and learn.	SD	D	U	A	SA
4.c	expects us to do meaningful devotionals every class period.	SD	D	U	A	SA
4.d	always starts and ends class on time.	SD	D	U	A	SA
4.e	does not allow inappropriate behavior in class (e.g., bad language, disrespect).	SD	D	U	A	SA
4.f	avoids being a friend rather than a teacher to us.	SD	D	U	A	SA
4.g	expects us to respect the house or building where class is held.	SD	D	U	A	SA
AREA 5—Helps students accept responsibility for gospel learning.						
	My teacher:					
5.a	asks questions that help us figure things out ourselves.	SD	D	U	A	SA
5.b	expects us to ask questions when we do not understand.	SD	D	U	A	SA
5.c	expects us to study the scriptures on our own without being reminded.	SD	D	U	A	SA
5.d	does not speak down to us but treats us as fellow students of the scriptures.	SD	D	U	A	SA
5.e	asks us what we have discovered in the scriptures in our own study.	SD	D	U	A	SA
5.f	makes us work at understanding the scriptures ourselves instead of just explaining it to us.	SD	D	U	A	SA
5.g	teaches us how to study the gospel on our own.	SD	D	U	A	SA
AREA 6—Makes good decisions about what to teach.						
	My teacher:					
6.a	teaches us principles that really apply in our lives.	SD	D	U	A	SA
6.b	focuses on parts of the scriptures that are most important.	SD	D	U	A	SA
6.c	teaches us in a way that is not too simple, but also not over our head.	SD	D	U	A	SA
6.d	organizes lessons so that the ideas are presented in an order that makes sense.	SD	D	U	A	SA
6.e	avoids spending a lot of time on things that are not very important.	SD	D	U	A	SA
6.f	takes time to answer our specific concerns or questions.	SD	D	U	A	SA

Page 2 of 5

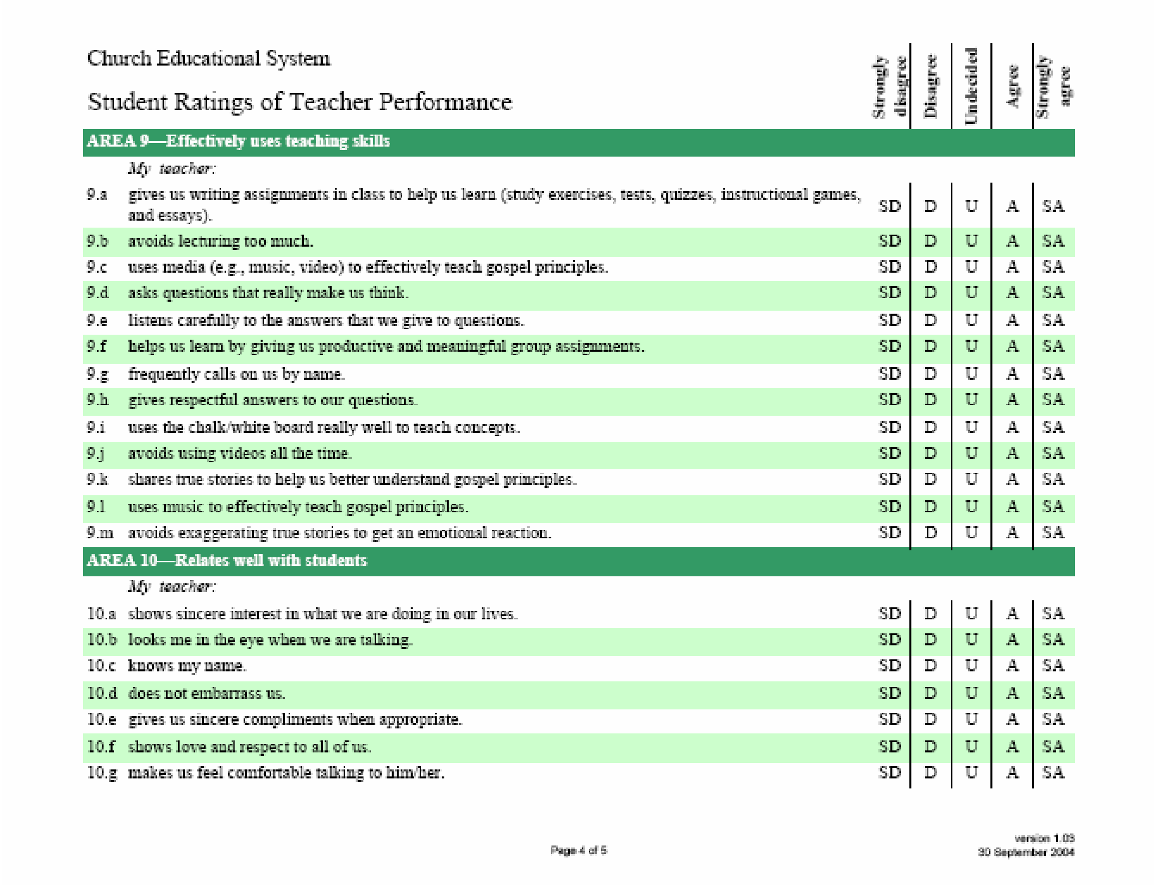
version 1.03
30 September 2004



Church Educational System

Student Ratings of Teacher Performance

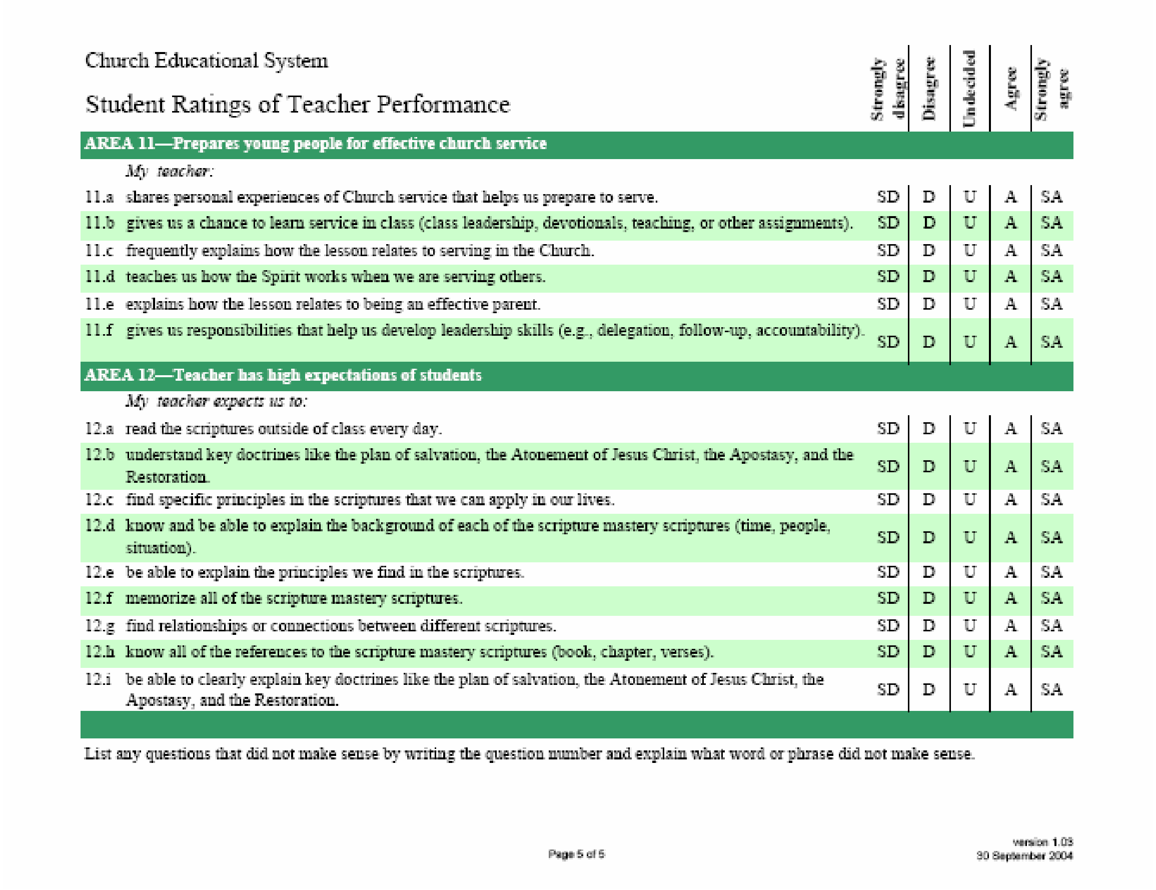
		Strongly disagree	Disagree	Undecided	Agree	Strongly agree
AREA 7—Makes good decisions about how to teach.						
	My teacher:					
7.a	teaches in a way that is very uplifting.	SD	D	U	A	SA
7.b	encourages us to get into the lesson by participating in class activities (group work, discussions, writing activities).	SD	D	U	A	SA
7.c	avoids taking too much time on one part of the lesson and rushing through the rest.	SD	D	U	A	SA
7.d	uses a lot of variety when teaching.	SD	D	U	A	SA
7.e	avoids teaching in a way that offends students.	SD	D	U	A	SA
AREA 8—Effectively uses scripture study skills						
	My teacher:					
8.a	teaches us to use the scripture study aids (footnotes, chapter headings, topical guide, Bible dictionary) by using them frequently in class.	SD	D	U	A	SA
8.b	helps us to learn the references for the scripture mastery scriptures (book, chapter, and verses).	SD	D	U	A	SA
8.c	asks us to look for particular things in the scriptures we are studying (principles, definitions, symbols, if/then relationships, patterns).	SD	D	U	A	SA
8.d	helps us to understand the doctrines found in the scripture mastery scriptures.	SD	D	U	A	SA
8.e	encourages us to mark our scriptures as we study.	SD	D	U	A	SA
8.f	helps us to see patterns in the scriptures by "chaining" or connecting related scriptures together.	SD	D	U	A	SA
8.g	reviews scripture mastery scriptures with us regularly.	SD	D	U	A	SA
8.h	helps us to liken the scriptures to our own situation in life.	SD	D	U	A	SA
8.i	summarizes the scriptures in a way that really helps us understand what is being taught.	SD	D	U	A	SA
8.j	gives us time to ponder and reflect on how the scriptures relate to us.	SD	D	U	A	SA
8.k	helps us to understand the language of the scriptures (e.g., when the use of "man" or "men" in the scriptures refers to both men and women).	SD	D	U	A	SA
8.l	helps us to memorize the scripture mastery scriptures.	SD	D	U	A	SA



Church Educational System

Student Ratings of Teacher Performance

		Strongly disagree	Disagree	Undecided	Agree	Strongly agree
AREA 9—Effectively uses teaching skills						
	My teacher:					
9.a	gives us writing assignments in class to help us learn (study exercises, tests, quizzes, instructional games, and essays).	SD	D	U	A	SA
9.b	avoids lecturing too much.	SD	D	U	A	SA
9.c	uses media (e.g., music, video) to effectively teach gospel principles.	SD	D	U	A	SA
9.d	asks questions that really make us think.	SD	D	U	A	SA
9.e	listens carefully to the answers that we give to questions.	SD	D	U	A	SA
9.f	helps us learn by giving us productive and meaningful group assignments.	SD	D	U	A	SA
9.g	frequently calls on us by name.	SD	D	U	A	SA
9.h	gives respectful answers to our questions.	SD	D	U	A	SA
9.i	uses the chalk/white board really well to teach concepts.	SD	D	U	A	SA
9.j	avoids using videos all the time.	SD	D	U	A	SA
9.k	shares true stories to help us better understand gospel principles.	SD	D	U	A	SA
9.l	uses music to effectively teach gospel principles.	SD	D	U	A	SA
9.m	avoids exaggerating true stories to get an emotional reaction.	SD	D	U	A	SA
AREA 10—Relates well with students						
	My teacher:					
10.a	shows sincere interest in what we are doing in our lives.	SD	D	U	A	SA
10.b	looks me in the eye when we are talking.	SD	D	U	A	SA
10.c	knows my name.	SD	D	U	A	SA
10.d	does not embarrass us.	SD	D	U	A	SA
10.e	gives us sincere compliments when appropriate.	SD	D	U	A	SA
10.f	shows love and respect to all of us.	SD	D	U	A	SA
10.g	makes us feel comfortable talking to him/her.	SD	D	U	A	SA



Church Educational System

Student Ratings of Teacher Performance

		Strongly disagree	Disagree	Undecided	Agree	Strongly agree
AREA 11—Prepares young people for effective church service						
	My teacher:					
11.a	shares personal experiences of Church service that helps us prepare to serve.	SD	D	U	A	SA
11.b	gives us a chance to learn service in class (class leadership, devotionals, teaching, or other assignments).	SD	D	U	A	SA
11.c	frequently explains how the lesson relates to serving in the Church.	SD	D	U	A	SA
11.d	teaches us how the Spirit works when we are serving others.	SD	D	U	A	SA
11.e	explains how the lesson relates to being an effective parent.	SD	D	U	A	SA
11.f	gives us responsibilities that help us develop leadership skills (e.g., delegation, follow-up, accountability).	SD	D	U	A	SA
AREA 12—Teacher has high expectations of students						
	My teacher expects us to:					
12.a	read the scriptures outside of class every day.	SD	D	U	A	SA
12.b	understand key doctrines like the plan of salvation, the Atonement of Jesus Christ, the Apostasy, and the Restoration.	SD	D	U	A	SA
12.c	find specific principles in the scriptures that we can apply in our lives.	SD	D	U	A	SA
12.d	know and be able to explain the background of each of the scripture mastery scriptures (time, people, situation).	SD	D	U	A	SA
12.e	be able to explain the principles we find in the scriptures.	SD	D	U	A	SA
12.f	memorize all of the scripture mastery scriptures.	SD	D	U	A	SA
12.g	find relationships or connections between different scriptures.	SD	D	U	A	SA
12.h	know all of the references to the scripture mastery scriptures (book, chapter, verses).	SD	D	U	A	SA
12.i	be able to clearly explain key doctrines like the plan of salvation, the Atonement of Jesus Christ, the Apostasy, and the Restoration.	SD	D	U	A	SA

List any questions that did not make sense by writing the question number and explain what word or phrase did not make sense.

Appendix F

WINSTEPS Tables for the Student-Teacher Rapport Scale

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

     SUMMARY OF 298 MEASURED (NON-EXTREME) students
+------------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      24.3       6.0        3.99    1.00       .92    -.1    .94    -.1 |
| S.D.       4.0        .4        2.92     .22       .69    1.1    .77    1.1 |
| MAX.      29.0       6.0        7.96    2.29      3.26    2.7   3.65    2.6 |
| MIN.       8.0       2.0       -5.06     .60       .00   -2.7    .00   -2.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE    1.13  ADJ.SD     2.70  SEPARATION  2.39  studen RELIABILITY  .85 |
|MODEL RMSE    1.02  ADJ.SD     2.74  SEPARATION  2.69  studen RELIABILITY  .88 |
| S.E. OF student MEAN = .17                                                  |
+-----------------------------------------------------------------------------+
  MAXIMUM EXTREME SCORE:     78 students
      LACKING RESPONSES:      1 students
                DELETED:     29 students
        VALID RESPONSES:  99.2%

     SUMMARY OF 6 MEASURED (NON-EXTREME) items
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN    1207.2     295.7         .00     .13       .98    -.3    .94    -.6 |
| S.D.      38.4       1.1         .62     .00       .17    1.9    .21    2.0 |
| MAX.    1249.0     298.0        1.25     .14      1.18    1.9   1.18    1.5 |
| MIN.    1131.0     295.0        -.57     .12       .69   -3.7    .60   -3.9 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .14  ADJ.SD      .61  SEPARATION  4.43  item   RELIABILITY  .95 |
|MODEL RMSE     .13  ADJ.SD      .61  SEPARATION  4.58  item   RELIABILITY  .95 |
| S.E. OF item MEAN = .28                                                     |
+-----------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
```

Figure F1. WINSTEPS Table 3.1 for the STRS.

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------
student: REAL SEP.: 2.39  REL.: .85 ... item: REAL SEP.: 4.43  REL.: .95

         item STATISTICS:  MISFIT ORDER

+-------------------------------------------------------------------------+
|ENTRY    RAW                          MODEL|   INFIT  |  OUTFIT  |PTMEA|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| item |
|------------------------------------+----------+----------+-----+------|
|     2   1222    296     -.24     .13|1.18   1.9|1.18   1.5|A .80| 1.b  |
|     3   1249    298     -.57     .14|1.13   1.4|1.15   1.2|B .81| 1.c  |
|     1   1131    295     1.25     .12|1.05    .6|1.07    .7|C .87| 1.a  |
|     6   1215    295     -.17     .13| .93   -.7| .86  -1.3|c .87| 1.f  |
|     4   1191    295      .26     .13| .87  -1.5| .78  -2.1|b .85| 1.d  |
|     5   1235    295     -.53     .14| .69  -3.7| .60  -3.9|a .87| 1.e  |
|------------------------------------+----------+----------+-----+------|
| MEAN  1207.2  295.7      .00     .13| .98   -.3| .94   -.6|     |      |
| S.D.    38.4    1.1      .62     .00| .17   1.9| .21   2.0|     |      |
+-------------------------------------------------------------------------+
```

Figure F2. WINSTEPS Table 10.1 for the STRS.

```
 INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

DIF specification is: DIF=$S5W1

+------------------------------------------------------------------------------------+
| student    DIF    DIF    student    DIF    DIF       DIF     JOINT          item       |
| GROUP    MEASURE S.E.   GROUP    MEASURE S.E.   CONTRAST  S.E.    t  d.f. Number  Name |
|------------------------------------------------------------------------------------|
| 1          1.13   .17  2          1.44   .18       -.31   .25 -1.25 279      1 1.a   |
| 1           .00   .19  2          -.52   .20        .52   .28  1.89 280      2 1.b   |
| 1          -.29   .19  2          -.98   .21        .69   .28  2.46 282      3 1.c   |
| 1           .31   .18  2           .25   .19        .06   .27   .21 279      4 1.d   |
| 1          -.73   .20  2          -.43   .20       -.29   .28 -1.04 279      5 1.e   |
| 1          -.43   .19  2           .15   .19       -.58   .27 -2.13 280      6 1.f   |
+------------------------------------------------------------------------------------+
```

Figure F3. WINSTEPS Table 30.1 for the STRS.

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

SUMMARY OF CATEGORY STRUCTURE.  Model="R"
+------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN| MEASURE|
|-------------------+------------+------------++---------+--------+
|  1   1      19   1| -4.30 -4.08|   .69   .67||  NONE   |( -5.30)| 1
|  2   2      48   3| -2.05 -2.16|   .96   .86||   -4.07 |  -3.29 | 2
|  3   3     202  11|   .15   .22|   .98   .84||   -2.45 |  -1.03 | 3
|  4   4    1003  56|  3.78  3.76|  1.00  1.05||     .30 |   3.27 | 4
|  5   5     502  28|  6.88  6.90|  1.01   .92||    6.22 |(  7.32)| 5
|-------------------+------------+------------++---------+--------+
|MISSING      14   1|  2.50      |            ||         |        |
+------------------------------------------------------------------
AVERAGE MEASURE is mean of measures in category. It is not a parameter estimate.

+---------------------------------------------------------------------------------------------+
|CATEGORY    STRUCTURE   |  SCORE-TO-MEASURE   | 50% CUM.| COHERENCE|ESTIM| OBSERVED-EXPECTED |
| LABEL    MEASURE  S.E. | AT CAT. ----ZONE----|PROBABLTY| M->C C->M|DISCR|RESIDUAL DIFFERENCE|
|------------------------+---------------------+---------+----------+-----+-------------------|
|   1      NONE          |( -5.30) -INF   -4.43|         |  81%  47%|     |   -1.1%        -.2 | 1
|   2       -4.07    .31 |  -3.29  -4.43  -2.26|    -4.23 |  54%  37%| 1.29|    -.5%        -.3 | 2
|   3       -2.45    .18 |  -1.03  -2.26    .45|    -2.35 |  61%  53%| 1.00|    -.1%        -.2 | 3
|   4         .30    .10 |   3.27    .45   6.23|      .36 |  78%  83%| 1.04|     .1%         .9 | 4
|   5        6.22    .07 |(  7.32)  6.23  +INF |     6.23 |  74%  72%|  .98|    -.1%        -.3 | 5
+---------------------------------------------------------------------------------------------+
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      ++-------+-------+-------+-------+-------+-------+-------++
R  1.0 +                                                         +
O      |                                                         |
B      |                                  4444444                |
A      |11                              44       44             5|
B   .8 +  1                            4           4          55 +
I      |   1                          4             4        5   |
L      |    1                        4               4      5    |
I      |     1            33333     4                 4    5     |
T   .6 +      1          3     3   4                   4  5      +
Y      |      1    2    3       3 4                     4 5      |
    .5 +       1 22 22 3         *                       *       +
O      |        *     *          43                     5 4      |
F   .4 +      221    3 2        4  3                   5  4      +
       |     2   1   3  2      4    3                 5    4     |
R      |    2     1 3    2    4      3               5      4    |
E      |   2       *      2  4        3             55       4   |
S   .2 +  2       3 1      24          3           5          44 +
P      |22       3   1     42           3         5             4|
O      |        3     1  44  22          333   555               |
N      |     333      4**1     222         5***3                 |
S   .0 +**************5555***************************************+
E      ++-------+-------+-------+-------+-------+-------+-------++
    -6      -4      -2       0       2       4       6       8
```

Figure F4. WINSTEPS Table 3.2 for the STRS.

```
 INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

       students MAP OF items
               <more>|<rare>
    8 .############  +
                  .  |
                     |
                     |
    7          .### S+
                     |
                     |
                .##  |
    6                +
                  .  |
               .###  |
                     |
    5                +
                     |
                .##  |
                     |
    4               M+
                     |
                     |
           .#######  |
    3                +
                  .  |
                     |
                     |
    2           .##  +
                     |
                     |
                     |T 1.a
    1           .## S+
                     |
                  .  |S
                     |  1.d
    0             .  +M
                     |  1.b    1.f
                  .  |S 1.c    1.e
                     |
   -1            .#  +
                     |T
                  .  |
                  . T|
   -2                +
                     |
                     |
                     |
   -3                +
                  .  |
                     |
                  .  |
   -4             .  +
                     |
                  .  |
                     |
   -5             .  +
               <less>|<frequ>
 EACH '#' IS 9.
```

Figure F5. WINSTEPS Table 1.1 for the STRS.

```
 INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT) (BY OBSERVED CATEGORY)
-6    -4     -2     0      2      4      6      8     10
|------+------+------+------+------+------+------+------|  NUM   item
1      1  :   2  :    3    :         4         :   5    5    1  1.a
|                                                       |
|                                                       |
1  1  :   2   :   3    :         4          :   5       5    4  1.d
|                                                       |
1 1  :   2  :    3    :         4         :   5         5    6  1.f
1 1  :   2  :    3    :         4         :   5         5    2  1.b
11  :   2  :    3    :         4         :   5          5    5  1.e
1   :  2   :   3     :        4          :   5          5    3  1.c
|------+------+------+------+------+------+------+------|  NUM   item
-6    -4     -2     0      2      4      6      8     10

                  1      1  2    6   2  2  2 3   3    7
   1 1 211     54 13 5 518  6  2 6   1  81 6 2  18   25    students
               T           S           M           S
```

Figure F6. WINSTEPS Table 10.1 for the STRS.

```
 INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

 FACTOR 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
  STANDARDIZED RESIDUAL CORRELATIONS FOR items (SORTED BY LOADING)
       Factor 1 extracts 1.6 units out of 6 units of item residual variance noise.
       Yardstick (variance explained by measures)-to-This Factor ratio: 7.2:1
       Yardstick-to-Total Noise ratio (total variance of residuals): 1.9:1

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                             Empirical    Modeled
Total variance in observations     =         17.4  100.0%  100.0%
Variance explained by measures     =         11.4   65.4%   64.5%
Unexplained variance (total)       =          6.0   34.6%   35.5%
Unexpl var explained by 1st factor =          1.6    9.1%

+------------------------------------------------+
|      |       |       INFIT OUTFIT| ENTRY      |
|FACTOR|LOADING|MEASURE  MNSQ MNSQ |NUMBER ite |
|------+-------+-------------------+-----------|
|  1   |   .71 |    -.24 1.18 1.18 |2     2 1.b |
|  1   |   .69 |    -.57 1.13 1.15 |3     3 1.c |
|      |-------+-------------------+-----------|
|  1   |  -.50 |    1.25 1.05 1.07 |1     1 1.a |
|  1   |  -.41 |    -.17  .93  .86 |6     6 1.f |
|  1   |  -.32 |    -.53  .69  .60 |5     5 1.e |
|  1   |  -.30 |     .26  .87  .78 |4     4 1.d |
+------------------------------------------------+
```

Figure F7. WINSTEPS Table 23.3 for the STRS.

Appendix G

WINSTEPS Tables for the Scripture Mastery Expectation Scale

```
INPUT: 406 students, 6 items  MEASURED: 378 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

     SUMMARY OF 335 MEASURED (NON-EXTREME) students
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      23.5       5.9        2.67     .78       .98    -.2    .98    -.2 |
| S.D.       4.2        .5        1.95     .13       .73    1.3    .74    1.3 |
| MAX.      29.0       6.0        6.01    1.37      3.33    2.7   3.21    2.7 |
| MIN.       5.0       2.0       -2.54     .59       .04   -2.8    .04   -2.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .88  ADJ.SD    1.74  SEPARATION  1.97  studen RELIABILITY  .80 |
|MODEL RMSE    .79  ADJ.SD    1.78  SEPARATION  2.26  studen RELIABILITY  .84 |
| S.E. OF student MEAN = .11                                                  |
+-----------------------------------------------------------------------------+
  MAXIMUM EXTREME SCORE:     43 students
      LACKING RESPONSES:      1 students
                DELETED:     27 students
        VALID RESPONSES:  98.7%

     SUMMARY OF 6 MEASURED (NON-EXTREME) items
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN    1313.2     330.7         .00     .10      1.00     .0    .98    -.2 |
| S.D.      53.3       1.4         .52     .00       .08    1.0    .09    1.1 |
| MAX.    1403.0     333.0         .72     .11      1.12    1.4   1.15    1.7 |
| MIN.    1240.0     329.0        -.92     .10       .88   -1.5    .88   -1.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .10  ADJ.SD     .51  SEPARATION  4.98  item   RELIABILITY  .96 |
|MODEL RMSE    .10  ADJ.SD     .51  SEPARATION  5.08  item   RELIABILITY  .96 |
| S.E. OF item MEAN = .23                                                     |
+-----------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
TABLE 10.1 Scripture Mastery                         v2out.txt Feb 12 15:35 2005
INPUT: 406 students, 6 items  MEASURED: 378 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------
student: REAL SEP.: 1.97  REL.: .80 ... item: REAL SEP.: 4.98  REL.: .96
```

Figure G1. WINSTEPS Table 3.1 for the SMES.

```
item STATISTICS:  MISFIT ORDER

+-----------------------------------------------------------------------+
|ENTRY    RAW                          MODEL|   INFIT  |  OUTFIT  |PTMEA|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| item |
|------------------------------------+----------+----------+-----+------|
|     2   1332    333    -.10     .10|1.10   1.2|1.15   1.7|A .77| 2.b  |
|     1   1403    332    -.92     .11|1.12   1.4|1.06    .7|B .75| 2.a  |
|     4   1336    330    -.23     .10|1.00    .0| .96   -.4|C .74| 2.d  |
|     3   1240    330     .72     .10| .96   -.5| .94   -.7|c .80| 2.c  |
|     6   1307    330     .07     .10| .96   -.5| .91  -1.1|b .77| 2.f  |
|     5   1261    329     .46     .10| .88  -1.5| .88  -1.5|a .82| 2.e  |
|------------------------------------+----------+----------+-----+------|
| MEAN  1313.2  330.7     .00     .10|1.00    .0| .98   -.2|     |      |
| S.D.    53.3    1.4     .52     .00| .08   1.0| .09   1.1|     |      |
+-----------------------------------------------------------------------+
```

Figure G2. WINSTEPS Table 10.1 for the SMES.

```
 INPUT: 406 students, 6 items  MEASURED: 378 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

DIF specification is: DIF=$S5W1

+-----------------------------------------------------------------------------------+
| student   DIF   DIF   student   DIF   DIF      DIF    JOINT          item         |
| GROUP   MEASURE S.E.  GROUP   MEASURE S.E.  CONTRAST  S.E.   t  d.f. Number  Name |
|-----------------------------------------------------------------------------------|
| 1         -.83   .15  2        -1.02   .16       .19   .22   .86 315      1 2.a   |
| 1          .02   .14  2         -.24   .15       .26   .21  1.27 316      2 2.b   |
| 1          .55   .14  2          .88   .14      -.33   .20 -1.67 313      3 2.c   |
| 1         -.20   .14  2         -.24   .15       .03   .21   .16 313      4 2.d   |
| 1          .52   .14  2          .41   .15       .11   .20   .54 312      5 2.e   |
| 1         -.05   .14  2          .16   .15      -.22   .21 -1.05 313      6 2.f   |
+-----------------------------------------------------------------------------------+
```

Figure G3. WINSTEPS Table 30.1 for the SMES.

```
INPUT: 406 students, 6 items  MEASURED: 378 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

SUMMARY OF CATEGORY STRUCTURE.  Model="R"
+------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN| MEASURE|
|-------------------+------------+------------++---------+--------+
|  1   1       9   0| -1.66 -2.09|  1.28  1.19||  NONE   |( -5.17)| 1
|  2   2     117   6|  -.88  -.77|   .93   .91||   -4.04 |  -2.53 | 2
|  3   3     323  16|   .96   .93|   .95   .94||    -.95 |   -.13 | 3
|  4   4    1008  50|  2.71  2.71|  1.03  1.03||     .68 |   2.52 | 4
|  5   5     527  26|  4.58  4.57|  1.03   .99||    4.31 |(  5.42)| 5
|-------------------+------------+------------++---------+--------+
|MISSING      26   1|   .90      |            ||         |        |
+------------------------------------------------------------------
AVERAGE MEASURE is mean of measures in category. It is not a parameter estimate.

+--------------------------------------------------------------------------
|CATEGORY    STRUCTURE   |  SCORE-TO-MEASURE   | 50% CUM.| COHERENCE|ESTIM|
| LABEL    MEASURE  S.E. | AT CAT. ----ZONE----|PROBABLTY| M->C C->M|DISCR|
|------------------------+---------------------+---------+----------+-----+
|   1      NONE          |( -5.17) -INF   -4.15|         |   0%   0%|     | 1
|   2       -4.04    .36 |  -2.53  -4.15  -1.18|   -4.08 |  66%  40%|  .91| 2
|   3        -.95    .12 |   -.13  -1.18    .95|   -1.06 |  49%  47%| 1.07| 3
|   4         .68    .07 |   2.52    .95   4.38|     .81 |  68%  80%|  .99| 4
|   5        4.31    .06 |(  5.42)  4.38  +INF |    4.33 |  74%  59%| 1.01| 5
+--------------------------------------------------------------------------
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      ++---------+---------+---------+---------+---------+---------++
R  1.0 +                                                             +
O      |                                                             |
B      |1                                                            |
A      | 11                                                         5|
B   .8 +   1                                                      55 +
I      |    1                                    44444           5   |
L      |     11         2222                   44     44        5    |
I      |       1      22    22               44         4      5     |
T   .6 +        1   22        2             4            4    5      +
Y      |         1 2           2     3     4              4  5       |
    .5 +          *             2  33 333 4                45        +
O      |         2 1             *3      *3                54        |
F   .4 +        2   1           3 2     4  3              5  44      +
       |       2     1         3   2    4   3            5     4     |
R      |      2       1       3     2  4     3          5       4    |
E      |    22         1    33       24       33      55         4   |
S   .2 +   2            1  3        442         3    5            44 +
P      | 22              **        4   22        3355               4|
O      |2              33  11    44      2       5533                |
N      |           3333      1***         22**555    33333           |
S   .0 +**********************555***********11***********************+
E      ++---------+---------+---------+---------+---------+---------++
       -6        -4        -2         0         2         4         6
```

Figure G4. WINSTEPS Table 3.2 for the SMES.

```
 INPUT: 406 students, 6 items  MEASURED: 378 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

       students MAP OF items
               <more>|<rare>
    7     .########  +
                     |
                    T|
                     |
                     |
    6        .#####  +
                     |
                     |
                     |
                     |
    5        .#####  +
                     |
                    S|
            .######  |
                     |
    4             .  +
             ######  |
                     |
                     |
             .#####  |
    3                +
                  .  |
      ############# M|
                     |
                     |
    2        .#####  +
                     |
                  .  |
               ####  |
                  .  |
    1           ###  +T
                    S|  2.c
                .##  |S
                  .  |  2.e
                 .#  |
    0                +M 2.f
                ###  |  2.b    2.d
                 .#  |
                  .  |S
                  .  |
   -1                +T 2.a
                  . T|
                     |
                  .  |
                  .  |
   -2             .  +
                     |
                     |
                  .  |
                     |
   -3                +
               <less>|<frequ>
 EACH '#' IS 5.
```

Figure G5. WINSTEPS Table 1.1 for the SMES.

```
INPUT: 406 students, 6 items  MEASURED: 378 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT) (BY OBSERVED CATEGORY)
-7     -5      -3      -1       1       3       5       7
|-------+-------+-------+-------+-------+-------+-------|  NUM   item
1         1   :      2    :   3    :     4      :    5  5    3  2.c
1        1   :      2    :   3    :     4      :    5   5    5  2.e
|                                                       |
1       1   :     2     :   3   :     4       :   5     5    6  2.f
1      1   :     2     :   3   :      4      :   5      5    2  2.b
1     1   :      2    :    3   :     4       :   5      5    4  2.d
|                                                       |
|                                                       |
1   1   :     2     :   3   :     4       :   5         5    1  2.a
|-------+-------+-------+-------+-------+-------+-------|  NUM   item
-7     -5      -3      -1       1       3       5       7

                            1 1 1 2 2 6  2 3  3 2   2   4
                  4 2134 351564 520 7 51 9 03 2 7   6   3  students
                      T        S         M        S
```

Figure G6. WINSTEPS Table 10.1 for the SMES.

```
INPUT: 406 students, 6 items  MEASURED: 378 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

 FACTOR 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
  STANDARDIZED RESIDUAL CORRELATIONS FOR items (SORTED BY LOADING)
       Factor 1 extracts 1.9 units out of 6 units of item residual variance noise.
       Yardstick (variance explained by measures)-to-This Factor ratio: 5.1:1
       Yardstick-to-Total Noise ratio (total variance of residuals): 1.6:1

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                             Empirical    Modeled
Total variance in observations     =         15.6  100.0%  100.0%
Variance explained by measures     =          9.6   61.4%   61.1%
Unexplained variance (total)       =          6.0   38.6%   38.9%
Unexpl var explained by 1st factor =          1.9   12.1%

+-----------------------------------------------+
|      |       |        INFIT OUTFIT| ENTRY     |
|FACTOR|LOADING|MEASURE  MNSQ MNSQ  |NUMBER ite |
|------+-------+--------------------+-----------|
|  1   |   .71 |     -.10 1.10 1.15 |2    2 2.b |
|  1   |   .65 |     -.92 1.12 1.06 |1    1 2.a |
|  1   |   .17 |      .46  .88  .88 |5    5 2.e |
|      |-------+--------------------+-----------|
|  1   |  -.66 |     -.23 1.00  .96 |4    4 2.d |
|  1   |  -.62 |      .72  .96  .94 |3    3 2.c |
|  1   |  -.30 |      .07  .96  .91 |6    6 2.f |
+-----------------------------------------------+
```

Figure G7. WINSTEPS Table 23.3 for the SMES.

Appendix H

WINSTEPS Tables for the Spiritual Learning Environment Scale

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

     SUMMARY OF 332 MEASURED (NON-EXTREME) students
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      25.2       6.0        3.90     .90       .97    -.1    .98    -.1 |
| S.D.       3.4        .1        2.27     .13       .62    1.1    .66    1.1 |
| MAX.      29.0       6.0        6.76    1.11      3.65    2.7   3.72    2.6 |
| MIN.       7.0       5.0       -5.27     .57       .04   -2.7    .03   -2.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE   1.00  ADJ.SD    2.04  SEPARATION  2.04  studen RELIABILITY  .81 |
|MODEL RMSE    .91  ADJ.SD    2.08  SEPARATION  2.28  studen RELIABILITY  .84 |
| S.E. OF student MEAN = .12                                                  |
+-----------------------------------------------------------------------------+
  MAXIMUM EXTREME SCORE:     44 students
      LACKING RESPONSES:      2 students
                DELETED:     28 students
        VALID RESPONSES:  99.7%

     SUMMARY OF 6 MEASURED (NON-EXTREME) items
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN    1396.7     331.0         .00     .12       .98    -.2    .98    -.2 |
| S.D.      32.9       1.4         .48     .00       .06     .8    .07     .8 |
| MAX.    1451.0     332.0         .69     .13      1.08     .9   1.05     .6 |
| MIN.    1349.0     328.0        -.77     .11       .89   -1.4    .85   -1.9 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .12  ADJ.SD     .46  SEPARATION  3.91  item   RELIABILITY  .94 |
|MODEL RMSE    .12  ADJ.SD     .47  SEPARATION  3.95  item   RELIABILITY  .94 |
| S.E. OF item MEAN = .21                                                     |
+-----------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
```

Figure H1. WINSTEPS Table 3.1 for the SLES.

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------
student: REAL SEP.: 2.04  REL.: .81 ... item: REAL SEP.: 3.91  REL.: .94

         item STATISTICS:  MISFIT ORDER

+-------------------------------------------------------------------------+
|ENTRY    RAW                         MODEL|   INFIT  |  OUTFIT  |PTMEA|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| item |
|------------------------------------+----------+----------+-----+------|
|     1   1418    332     -.22     .12|1.08    .9|1.03    .4|A .76| 3.a  |
|     6   1391    331      .10     .12|1.03    .4|1.05    .6|B .78| 3.f  |
|     3   1349    332      .69     .11| .97   -.4|1.02    .3|C .81| 3.c  |
|     2   1402    328     -.23     .12| .99   -.1| .98   -.2|c .76| 3.b  |
|     5   1451    331     -.77     .13| .94   -.7| .97   -.3|b .76| 3.e  |
|     4   1369    332      .44     .11| .89  -1.4| .85  -1.9|a .81| 3.d  |
|------------------------------------+----------+----------+-----+------|
| MEAN  1396.7  331.0      .00     .12| .98   -.2| .98   -.2|     |      |
| S.D.    32.9    1.4      .48     .00| .06    .8| .07    .8|     |      |
+-------------------------------------------------------------------------+
```

Figure H2. WINSTEPS Table 10.1 for the SLES.

```
 INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------
DIF specification is: DIF=$S5W1
+-------------------------------------------------------------------------------------+
| student    DIF    DIF   student    DIF    DIF       DIF     JOINT          item        |
| GROUP    MEASURE S.E.   GROUP    MEASURE S.E.   CONTRAST  S.E.    t  d.f. Number  Name |
|-------------------------------------------------------------------------------------|
| 1          -.10   .16  2          -.35    .18       .25    .24  1.03 316      1 3.a   |
| 1          -.33   .17  2          -.15    .18      -.18    .25  -.73 312      2 3.b   |
| 1           .79   .16  2           .55    .17       .24    .23  1.06 316      3 3.c   |
| 1           .39   .16  2           .55    .17      -.16    .23  -.67 316      4 3.d   |
| 1          -.76   .18  2          -.94    .19       .17    .26   .67 315      5 3.e   |
| 1          -.02   .16  2           .29    .17      -.32    .24 -1.33 315      6 3.f   |
+-------------------------------------------------------------------------------------+
```

Figure H3. WINSTEPS Table 30.1 for the SLES.

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

SUMMARY OF CATEGORY STRUCTURE.  Model="R"
+------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN| MEASURE|
|-------------------+------------+------------++---------+--------+
|  1   1      16   1| -4.30 -4.29|   .70   .70||  NONE   |( -4.72)| 1
|  2   2      34   2| -1.30 -1.31|   .98   .97||   -3.49 |  -2.72 | 2
|  3   3     171   9|   .80   .80|  1.04   .99||   -1.85 |   -.74 | 3
|  4   4    1042  52|  3.46  3.47|   .98  1.02||     .26 |   2.68 | 4
|  5   5     723  36|  5.70  5.69|   .97   .95||    5.08 |(  6.18)| 5
|-------------------+------------+------------++---------+--------+
|MISSING       6   0|  3.07      |            ||         |        |
+------------------------------------------------------------------
AVERAGE MEASURE is mean of measures in category. It is not a parameter estimate.
```

```
+---------------------------------------------------------------------------------------------+
|CATEGORY    STRUCTURE    |  SCORE-TO-MEASURE    | 50% CUM.| COHERENCE|ESTIM| OBSERVED-EXPECTED |
| LABEL    MEASURE  S.E.  | AT CAT. ----ZONE----|PROBABLTY| M->C C->M|DISCR|RESIDUAL DIFFERENCE|
|------------------------+--------------------+---------+----------+-----+-------------------|
|   1      NONE          |( -4.72) -INF   -3.84|         |  83%  62%|     |   -1.5%       -.2 | 1
|   2       -3.49    .40 |  -2.72  -3.84  -1.76|   -3.65 |  50%  32%| 1.06|    -.7%       -.2 | 2
|   3       -1.85    .21 |   -.74  -1.76    .49|   -1.80 |  54%  43%| 1.01|    -.2%       -.3 | 3
|   4         .26    .10 |   2.68    .49   5.10|     .36 |  74%  75%| 1.00|     .0%        .5 | 4
|   5        5.08    .06 |(  6.18)  5.10  +INF |    5.09 |  71%  75%| 1.02|     .0%        .3 | 5
+---------------------------------------------------------------------------------------------+
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?
```

```
        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      ++---------+---------+---------+---------+---------+---------++
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |                                    444444                 55|
B   .8 +11                                44      44             55  +
I      |  1                             44          44          5    |
L      |   1                           4              4        5     |
I      |    1                         4                4     55      |
T   .6 +     1               33      4                  4   5        +
Y      |      1    2       33  33   4                    4 5         |
    .5 +       1 22 222  33      334                      *          +
O      |       2*      23         43                     5 4         |
F   .4 +      2 1      322       4  3                   5   4        +
       |     2   1    3   2     4    3                 5     44      |
R      |   22     1  3     2   4      3               5        4     |
E      |  2        13       2 4        33            5          4    |
S   .2 +22         31        *           3         55           44  +
P      |         33  11    44 22          33     55                44|
O      |       33      1144     22          33355                    |
N      |    333       444111      2222    5555533333                 |
S   .0 +**************555555*****************************************+
E      ++---------+---------+---------+---------+---------+---------++
       -5        -3        -1         1         3         5         7
```

Figure H4. WINSTEPS Table 3.2 for the SLES.

```
 INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

       students MAP OF items
               <more>|<rare>
    7   ###########  +
      .############  |
                     |
                    S|
    6                +
      .############  |
                  .  |
                     |
    5    .#########  +
                     |
        .##########  |
                     |
    4               M+
           ########  |
                     |
                     |
    3                +
      .############  |
                     |
                     |
    2                +
              ##### S|
                     |
                     |
    1          ####  +T
                     |  3.c
                .##  |S 3.d
                  .  |
    0            .#  +M 3.f
                     |  3.a    3.b
                     |S
                 .# T|  3.e
   -1             .  +T
                     |
                  .  |
                     |
   -2             .  +
                     |
                     |
                  .  |
   -3             .  +
                     |
                     |
                     |
   -4                +
                     |
                     |
                     |
   -5                +
                  .  |
                     |
                     |
   -6                +
               <less>|<frequ>
 EACH '#' IS 4.
```

Figure H5. WINSTEPS Table 1.1 for the SLES.

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT) (BY OBSERVED CATEGORY)
-6     -4      -2       0       2       4       6       8
|-------+-------+-------+-------+-------+-------+-------|  NUM   item
1       1  :    2   :   3    :       4         :   5    5    3  3.c
|                                                       |
1      1  :    2   :   3    :       4         :   5     5    4  3.d
|                                                       |
1     1  :    2  :   3    :        4         :   5      5    6  3.f
|                                                       |
1   1   :   2   :   3    :        4         :   5       5    1  3.a
1   1   :   2   :   3    :        4        :    5       5    2  3.b
|                                                       |
|                                                       |
1 1   :   2   :   3    :        4        :    5         5    5  3.e
|-------+-------+-------+-------+-------+-------+-------|  NUM   item
-6     -4      -2       0       2       4       6       8

                          1 1  2   5   3  4 3  5   5    4
   2        11  1 2 17  611 6  0   1   2  1 7 12   0    3  students
                     T         S          M         S
```

Figure H6. WINSTEPS Table 10.1 for the SLES.

```
INPUT: 406 students, 6 items  MEASURED: 376 students, 6 items, 5 CATS       3.53
--------------------------------------------------------------------------------

 FACTOR 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
  STANDARDIZED RESIDUAL CORRELATIONS FOR items (SORTED BY LOADING)
       Factor 1 extracts 1.5 units out of 6 units of item residual variance noise.
       Yardstick (variance explained by measures)-to-This Factor ratio: 6.5:1
       Yardstick-to-Total Noise ratio (total variance of residuals): 1.6:1

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                             Empirical    Modeled
Total variance in observations     =         15.6  100.0%  100.0%
Variance explained by measures     =          9.6   61.5%   61.9%
Unexplained variance (total)       =          6.0   38.5%   38.1%
Unexpl var explained by 1st factor =          1.5    9.5%

+-----------------------------------------------+
|      |       |        INFIT OUTFIT| ENTRY     |
|FACTOR|LOADING|MEASURE  MNSQ MNSQ  |NUMBER ite |
|------+-------+--------------------+-----------|
|  1   |   .66 |    -.23  .99   .98 |2    2 3.b |
|  1   |   .65 |    -.22 1.08 1.03 |1    1 3.a |
|  1   |   .05 |    -.77  .94   .97 |5    5 3.e |
|      |-------+--------------------+-----------|
|  1   |  -.52 |     .44  .89   .85 |4    4 3.d |
|  1   |  -.50 |     .10 1.03 1.05 |6    6 3.f |
|  1   |  -.31 |     .69  .97 1.02 |3    3 3.c |
+-----------------------------------------------+
```

Figure H7. WINSTEPS Table 23.3 for the SLES.