12-2000

Issues of Factorial Invariance Inherent in Conceptual Change: Teachers' Evolving Perceptions of Classroom Practice

Cynthia C. Phillips
Western Michigan University

ISSUES OF FACTORIAL INVARIANCE INHERENT IN CONCEPTUAL CHANGE: TEACHERS' EVOLVING PERCEPTIONS OF CLASSROOM PRACTICE

by

Cynthia C. Phillips

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Educational Studies

Western Michigan University
Kalamazoo, Michigan
December 2000

ISSUES OF FACTORIAL INVARIANCE INHERENT IN CONCEPTUAL CHANGE: TEACHERS' EVOLVING PERCEPTIONS OF CLASSROOM PRACTICE

Cynthia C. Phillips, Ph.D.

Western Michigan University, 2000

This study explored the extent to which confirmatory factor analysis (CFA) can be used to address the measurement challenges faced by evaluators engaged in the assessment of change; in particular, the interpretation of self-report survey data collected under quasi-experimental conditions. The psychometric principles behind the instruments used to measure change are built on the assumptions that the constructs of interest remain stable and that error and score magnitude alone may vary. This study examined the complications that arise, with respect to the valid use of change scores, when the constructs of interest reflect conceptual change.

CFA techniques are available enabling structural comparison of the equivalence, or invariance, among factors across groups, situations, and/or time applicable to situations where issues of construct coherence and stability threaten the valid use of survey data. In that factor structure reflects the "mental model" expressed by a group of respondents for a given construct, these techniques more importantly can be utilized to provide as yet untapped evidence of conceptual change, widely theorized to precede behavioral outcomes. This investigation of factorial invariance served as the means to examine the extent to which systemic reform-minded professional development was associated with the structural evolution of teachers' perceptions with respect to the multi-dimensional nature of classroom practice (traditional, investigative culture, investigative practice factors).

The findings from this study provide evidence that teachers who have participated in reform-minded professional development envision their teaching practice in different ways

than teachers that have not yet been reached. Although treatment exposure was not associated with extensive alterations in the measurement structure for any of the three teaching practice factors, these data do provide evidence of conceptual change in the relationships among factors in that higher levels of treatment exposure were found to be associated with reform factors both more distinct from each other and from the traditional practice factor. In addition, interpretation of these results presents clear implications and suggestions for improved evaluation practice and a deeper understanding of the challenge of change.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

UMI Number: 9988434

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGMENTS

I wish to warmly express my appreciation to several individuals for their contributions to the completion of this dissertation. I thank my husband for his technical support and my children for their patience. I am grateful to the many friends and colleagues who thoughtfully listened to my struggles giving me valuable feedback when it was most needed. I thank my committee members Dr. Zoe Barley, who first exposed me to the field of evaluation and Dr. Jianping Shen, who showed me a matrix is worth a thousand words. In particular, I am profoundly fortunate that my committee chairperson, Dr. Mary Anne Bunda, was tireless in her insistence that I was indeed a "measurement person."

Cynthia C. Phillips

TABLE OF CONTENTS

Table of Contents—continued

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This dissertation describes an application of confirmatory factor analysis (CFA) that delves into the measurement issues conceptual change presents for the interpretation of self-report survey data. The psychometric principles behind the instruments used to measure change are built on the assumptions that the constructs of interest remain stable and that error and score magnitude alone may vary. This study examined the measurement complications that arise, with respect to the valid use of change scores, when the constructs of interest are theorized to evolve in response to conceptual change.

Although rarely used in the field of evaluation, CFA techniques are available that enable comparison of the relative coherence and equivalence among factors and their structural relationships across groups, situations, and/or time. In that factor structure reflects the "mental model" expressed by a group of respondents for a given construct, exploration of factorial invariance using CFA methodology presents a plausible approach to the capture and interpretation of conceptual change. Indication of conceptual change, widely theorized to precede behavioral outcomes, provides as yet untapped evidence for evaluators to consider when assessing the efficacy of change initiatives, such as systemic educational reform (Argyris & Schön, 1974; Evans, 1996; Halford, 1995; Taylor, 1990). Alternatively, under those conditions where the radical reconstitution of constructs does occur, the valid use of change scores becomes compromised. Specifically, this study draws on the CFA methods established to investigate factorial invariance as the means to examine the extent to which systemic reform-minded professional development is associated with the structural evolution of teachers' perceptions with respect to the multi-dimensional nature of classroom practice.

This dissertation contains five chapters. Given the nature of the investigation, the contents and scope of the first two chapters of this dissertation depart somewhat from the expected format. The present chapter provides an overview to the research conducted and lays out the case for the problem to be addressed. It consists of (a) a statement of the problem of change and its bearing on the practice of evaluation and (b) a brief explanation of the rationale and relevance behind this method-focused research. The second chapter consists of a literature review that provides background on the topic of factorial invariance and its application to the study of change—and as such lays out a proposed way to address the problem. The third chapter describes the CFA methodology employed in the execution of the research. The fourth chapter puts forth the results and the fifth chapter presents the discussion and conclusions drawn. The assumption is made throughout that the reader grasps the mathematics and mechanics behind CFA—those readers less familiar with basic concepts are referred to the excellent structural equation modeling texts currently available (Byrne, 1998; Kelloway, 1998; Maruyama,1998).

Statement of the Problem

Measurement Challenges Associated With the Analysis of Change

Surveys are among the most important data collection tools available in evaluation for the assessment of change. In particular, self-report surveys are widely used by evaluators to assess the prevalence of attitudes, beliefs, and behavior; to track long-term change; as well as to identify and examine differences between treatment and control or comparison groups (Braverman, 1996; Weisberg, Krosnick, & Bowen, 1996). However, the practical reality of survey implementation in the field frequently presents evaluators with challenges related to fundamental measurement principles that threaten the valid use of self-report survey data. The structure, clarity, and stability of the mental picture or "nomological net" respondents use

to describe a given construct is of foremost concern (Cronbach & Meehl, 1955). Treatment influence, research design, and analysis constraints, common to most longitudinal evaluation studies, threaten construct validity and thus the credible application of survey methodology, in particular, to the study and analysis of change (Cook & Campbell, 1979; Murnane, Singer, & Willett, 1988; Porras & Berg, 1978). Each of these constraints is discussed further below.

Three decades ago, Cronbach and Furby (1970) concluded that the measurement of change was a compound and challenging venture. Subsequent research pointed out that self-report survey data makes the measurement of change more complex and problematic than originally thought (Golembiewski, Billingsley, & Yeager, 1976; Howard & Dailey, 1979). Under some conditions, for example Cronbach and Furby's (1970) recommendation that comparison of post-intervention scores be used to assess change when possible is often inappropriate with self-report data (Howard, Schmeck, & Bray, 1979). The use of self-report survey data presents the prospect that an intervention or treatment may change the composition of, or relationship among, concepts that respondents use to describe behavior (Lindell & Drexler, 1979). This potential to produce alteration in the "mental model" held by respondents for the construct being assessed is a distinct possibility when the treatment affects abilities or knowledge structures (Aiken & West, 1990; Senge, 1990). Yet the evaluator is frequently in the position of having to measure and report on change, be it change measured as conceptual or mean differences in the construct of interest, under conditions such as these—or worse. To address this first measurement challenge, evaluators need a way to establish that the treatment or intervention under investigation has not changed the way survey respondents "see" the constructs of interest.

Despite an increasing emphasis placed on outcome measurement and the assessment of program effectiveness, most programs are neither planned nor implemented in such a way as to enable evaluators to make use of powerful experimental research designs that deliver definitive data on causal relationships. Evaluators charged with making

summative judgment on the value or worth of a program frequently must rely, at best, on quasi-experimental designs—such as, non-equivalent control groups and observational studies—to assess program effects (Cook & Campbell, 1979). Under these constrained conditions lack of random assignment and the potential nonequivalence of the groups complicate, and frequently compromise, the valid use of evaluation findings. To address this second challenge, the evaluator needs the means to determine—at least in terms of the constructs under investigation—the extent to which the groups being compared "see" constructs the same way.

Most experts agree that panel studies, where the same individuals once sampled from the population of interest are surveyed across multiple points in time, provide the best evidence of the extent of change (Collins, 1991). Yet they are rarely used in evaluation practice because of cost and implementation limitations. Alternatively, successive cross-sections, where a different sample is drawn across multiple points in time from the same population, are more frequently used to assess change in evaluation studies. As with the previous challenges, here the evaluator also must use caution in interpreting results because of the possibility of conceptual differences across groups; but, here the concerns are the effect of time and/or developmental processes not related to the treatment under investigation.

When addressing these construct validity challenges in the context of evaluation studies, it is important to note that they arise, in part, because two distinct classes of variables are encountered—static and dynamic. *Dynamic* variables, which involve systematic intra-individual change over time, figure most prominently in the study of change; whereas, *static* variables are most frequently grounded in theory that does not hypothesize that change will occur. The rationale behind traditional measurement approaches (i.e., classical test theory) to instrument development focuses on static variables and is "based on the idea of unchanging true scores, with any change in observed scores directly attributable to measurement error"

(Collins, 1991, p. 138).

Although evaluators do rely on the internal consistency reliability and factor structure of such measures to establish the integrity of the survey tool, it is rare that the dynamic nature of the variables under investigation is taken into consideration. Where evaluators are most likely to tread on questionable measurement ground, despite these precautions, is when dynamic variables are involved. With a dynamic variable—whether studied across treatment groups, situations, and/or over time—change in observed score may be attributable to sources other than measurement error. Changes in the structural relationships for a given construct may have profound effects on the interpretation of differences across groups, situations, and/or over time. In addition to the valid use of self-report survey data, the evaluator must also be concerned with data reliability—the stability and consistency of the measure of respondents' perceptions with respect to the constructs and conditions under investigation.

Factorial Invariance is Prerequisite to Valid and Reliable Evaluative Inference

Clearly, issues of validity and reliability are central to the measurement challenges faced by evaluators that choose to rely on self-report survey data. First, evaluators must speak to the substantive meaning of the constructs under investigation. Construct validation evidence establishes that the same constructs are likely being measured across each aspect of the situation under investigation. Second, evaluators must attest to the immutable nature, or reliability of the construct being measured (Pitts et al., 1996). For evaluators to be able to compare results across groups, situations, and/or over time with confidence and rigor, it is essential to first establish that an invariant relationship exists for each construct across the conditions pertinent to the investigation conducted (Pitts et al., 1996).

Using the widely accepted definition proposed by Tisak and Meredith (1991) *measurement invariance* addresses the extent to which the same constructs are being

measured for each group, under each condition, and/or for each measurement wave. When measurement invariance is explored within a factor analytic model it is referred to as *factorial invariance*. Evaluators can use factorial invariance to address the challenges of working with self-report survey data that are quasi-experimental and/or longitudinal.

The extent to which factorial invariance can be demonstrated in these instances describes the degree to which respondents share the same perception, or "mental model" for a given construct such that it is comparable, equivalent, and stable across groups, conditions, and/or time. The determination of factorial invariance for the constructs under investigation serves as an essential requirement for making valid evaluative inferences about the effects of a treatment or intervention. Thus, factorial invariance provides the common thread connecting the challenges that arise during the assessment of program effects with a practical, methodological solution (Aiken, Stein, & Bentler, 1994; Horn & McArdle, 1992; McArdle & Nesselroade, 1994; Reise, Widaman, & Pugh, 1993; Schaubroeck & Green, 1989).

Once instrument developers have addressed basic issues of the internal consistency reliability, dimensionality, and the valid use of scores in initial cross sectional investigations (Feldt & Brennan, 1989; Messick, 1989; Nunnally & Bernstein, 1993) attention should then turn toward the determination of factorial invariance. It is important that the instrument developer identify such changes in factor structure, that "would be most detrimental to useful score interpretation" (Crocker & Algina, 1986, p. 132). In longitudinal studies with quasi-experimental and non-experimental designs evaluators should be concerned with whether the structure of their measures change across treatment, samples, and time (Pitts et al., 1996).

While sufficient for the static nature of some cross-sectional studies—reluctance to probe deeper into the factor structure of survey instruments used to study the dynamics of change limits the quality and utility of the evidence evaluators can compile for the purposes of program improvement and accountability. If the structures of factor scores are not invariant

across groups, then differences between groups in mean levels or in the pattern of correlations among factors are potentially artifactual and may be misleading (Meredith, 1993; Widaman & Reise, 1997).

Depending on whether factorial invariance can be established, the evaluator stands better able to assess the nature and extent of change. If a measure demonstrates factorial invariance across conditions commonly encountered in the practice of program evaluation, the strength of the argument that can be made for the effectiveness of the intervention—"quantity" of change, based on the comparison of mean scores, is greatly augmented. Yet alternatively, should the measure fail to demonstrate factorial invariance across conditions—and muddy the valid interpretation of mean differences—this in itself is powerful and as yet untapped evidence of the effectiveness of the intervention to produce conceptual or "quality" change (Widaman, 1991).

Factorial Invariance Identifies Undercurrents of Conceptual Change

"Evaluator's choices of what to count and what to study affect what they are likely to find out about what works" (Schorr, 1997, p.141). Change is usually only studied quantitatively, where change scores and mean differences are used with mixed success to indicate the relative amount or standing of a person or group on a particular variable or factor. The magnitude and interpretation of change scores and mean differences used to gauge the effectiveness of many initiatives are influenced by the measurement challenges, described herein, that arise throughout the evaluation process. Factorial invariance, which establishes the "quality" of change—in terms of substance and stability—should be considered a prerequisite diagnostic method. This type of evidence that supports the valid and reliable use of factor scores under a variety of situations also could be used to point to progress toward desired results, in terms of movement (conceptual change) or stasis in the way respondents have come to perceive the constructs of interest.

The psychometric properties of the linear composite scores frequently used to monitor change may well provide supporting evidence of the conceptual undercurrents that characterize change processes—even when little or no net change in the magnitude of a construct may be indicated across groups or over time. "The psychological structures underlying many types of behavior undergo important changes in kind, as well as exhibiting changes in level" (Widaman, 1991, p. 205). These structural modifications in the mental model of a given behavior undergoing change are the hallmark of conceptual change that promote observable behavior change (Senge, 1990). Thus, investigation of factorial invariance provides insight into the illusive process of change and in addition contributes interpretative power for both static and dynamic concepts not yet widely exploited by the field of evaluation (Millsap & Hartog, 1988).

What could conceivably contribute to a large body of psychometrically sound and convincing alternative evidence of conceptual change to date has been largely avoided and underutilized in evaluation practice. If factorial invariance cannot be demonstrated this in and of itself provides substantive evidence of dramatic shifts in the mental models held by groups of respondents for a given construct. Also, differences in factor variances/covariances or error variances across groups provide evidence of treatment induced "quality" change in terms of how respondents "see" and "interpret" similarities and differences in the relationships within and among factors even when composite scores fail to demonstrate a "quantity" difference.

For the most part, however, evaluators have yet to follow the recent lead by researchers in the training and organizational development field that capitalizes on the use of factorial invariance as an interpretative lens for change-focused research (Pitts et al., 1996; Taris et al., 1998). The analytic techniques used to investigate hypotheses of factorial invariance have also been shown to lessen the implications of the measurement challenges faced by evaluation professionals with respect to the utility, feasibility, and credibility of their

findings and recommendations (Taris et al., 1998). Given the frequency with which the measurement challenges described arise in evaluation studies, evaluators should be encouraged to consider conducting appropriate tests of factorial invariance on the surveys, questionnaires, and other instruments used to assess the quantity and quality of change.

In summary, the problem addressed by this dissertation is that *if* factorial invariance is not established, each of the measurement challenges described above can profoundly limit the valid use of evaluation data, particularly when analyzing change. Alternatively, changes in the conceptualization and reconstitution of multi-item concepts across groups, situations, and/or time may represent legitimate effects that *should* be investigated in their own right (Cunningham, 1991; Taris, et al., 1998; Widaman & Reise, 1997).

Research Rationale and Relevance

The purpose of this research was to explore the utility of employing CFA techniques to investigate factorial invariance as a means to improve the ways in which the assessment of change is approached. The assessment of factorial invariance not only provides evidence that supports improving the valid and reliable use of change scores but also bolsters the repertoire of methodological techniques currently available for the field of evaluation and brings a new lens to bear on the assessment of change. This study used CFA techniques to detect the reconstitution of constructs, as measured by changes in factors and the relationships among factors.

Quite simply, this study was an attempt to determine the extent to which teachers that have participated in reform-minded professional development envision their teaching practice in the same or different ways as teachers that have not yet been reached. This study framed factorial invariance as an opportunity to determine the extent to which evidence of conceptual change can be detected by CFA and factorial invariance methods. The overarching rationale for the study was to focus on the extent to which treatment, designed to evoke conceptual

change, would be associated with measurement structure alteration as hypothesized. The goal of this investigation was not so much to establish factorial invariance, but to explore the value of CFA techniques as alternative evaluation methods and the means to support the valid use of change scores under conditions when conceptual change occurs.

The Research Question

This study used an existing data set from the national evaluation of a systemic science educational reform initiative to model the influence of reform-minded professional development on the evolution of K-8 science teaching practice from a traditional to a more student-centered, constructivist approach. The overarching research question for this dissertation was as follows: To what extent was the factor structure for a self-report teaching practice frequency scale invariant across increased exposure to reform-minded professional development?

Research Audiences and Applications

Although this work was primarily intended to serve to inform evaluation methods, its findings are substantively grounded in systemic science educational reform. Thus the audiences and applications for this research are two-fold. The innovative methodological approach targets evaluators as an audience with its intent in application to encourage evaluation practice to include factorial invariance testing as a means to address the measurement challenges associated with the analysis of change. Whereas, the content aspects of the work that speak to the evolving structural "mental model" of teaching practice target educational practitioners as an audience with the intent of increasing awareness of the conceptual underpinnings of reform and how evidence of these changes might be measured.

In addition to perhaps reducing the influence of treatment, design, and analysis constraints on the valid use of information generated from longitudinal self-report survey data,

tests for factorial invariance add value above and beyond strengthening the psychometric properties of evaluative measures. The CFA techniques described here disentangle the qualitative and quantitative aspects of change. Through the procedures illustrated by this method research, evaluators can begin to assess the influence of conceptual change, or reconstitution, on the valid use and interpretation of mean differences observed across treatment groups. By moving the field of practice to include tests for factorial invariance as evidence that supports the valid use of change scores, evaluators also stand to gain valuable insights into the very nature of the change process.

As such, although advancing systemic science educational reform content knowledge was of secondary importance to this methods study, it is of practical interest to national, state, and project level evaluators and the decision-makers they serve. Additional project level audiences include the principal investigators and project staff, district and building administrators, and teachers actively pursuing science education reform. This examination of the relative influence of reform-minded professional development on the relationships among intermediate outcome indicators—those that describe the perceptions of teaching practice—effectively illustrates the conceptual change processes antecedent to reform.

CHAPTER 2

LITERATURE REVIEW

This chapter, which presents a review of literature used to construct the rationale for this investigation focused on the assessment of change, is comprised of two sections: (a) background on the topic of factorial invariance and (b) what the results from factorial invariance hypothesis testing can mean when applied to the study of change. Given that the case has been made for the study in Chapter One via its description of the measurement challenges facing evaluators charged with the assessment of change, the primary purpose of this literature review is to provide the reader with sufficient background on factorial invariance and the CFA techniques used to be able to appreciate the support these strategies bring to the measurement challenges associated with the assessment of change. CFA techniques provide straightforward and unequivocal ways to test the crucial hypotheses related to factorial invariance and thus serve as a valuable a tool to detect, distinguish, and assess both types of change—conceptual, which precludes the use of change scores, and those of magnitude alone. In addition upon application, CFA techniques serve as a flexible and potent new lens for assuring that any evaluative interpretation of change processes relies on the assertion of similarities and differences across groups from a compelling measurement position.

The Relevance and Importance of Factorial Invariance to the Study of Change

The concept of factorial invariance is central to understanding the methods and findings reported in this dissertation. Cursory definitions were provided in the introductory chapter but a more through discussion of historical background and synthesis of the factorial

invariance and conceptual change literature is presented here. Further explanation is necessary to assist the reader to make the connection between the CFA method used to determine factorial invariance and its practical application as an appropriate means to reduce the limitations of a set of wide spread measurement problems encountered by evaluators. Factorial invariance can be used to look deeper and more judiciously into the qualitative and quantitative aspects of the change process. When measurement structures fail to be equivalent, this in and of itself may provide evidence of change in the qualities of an idea, for example conceptual change. On the other hand, when structures are found to be equivalent group differences are more succinctly interpretable when group, contextual, and/or temporal comparisons are required.

A Concrete Example

An applied, concrete illustration of the measurement implications of factorial invariance may be beneficial at this point before engaging in a review of the pertinent literature on factorial invariance and its application to the analysis of change. Loosely following the excellent example provided by Horn (1991), suppose that an evaluator seeks to compare the quality of life for 20, 40, and 60-year-old women. This evaluator, from the literature hypothesizes that "quality of life" could be measured by summing the number of yes answers to the following three questions:

Z_1: Do you think you are more attractive than the average person?

Z_2: Do you think you are wealthier than the average person?

Z_3: Do you think you are healthier than the average person?

Hypothetical factor pattern and factor score results for this example follow in the diagram below (Figure 1). These factor analytic results reveal that "quality of life" was conceived differently in the minds of young, middle-aged, and older women. In young women beauty was the best indicator, yet for middle-aged women it was weaker, and for older women

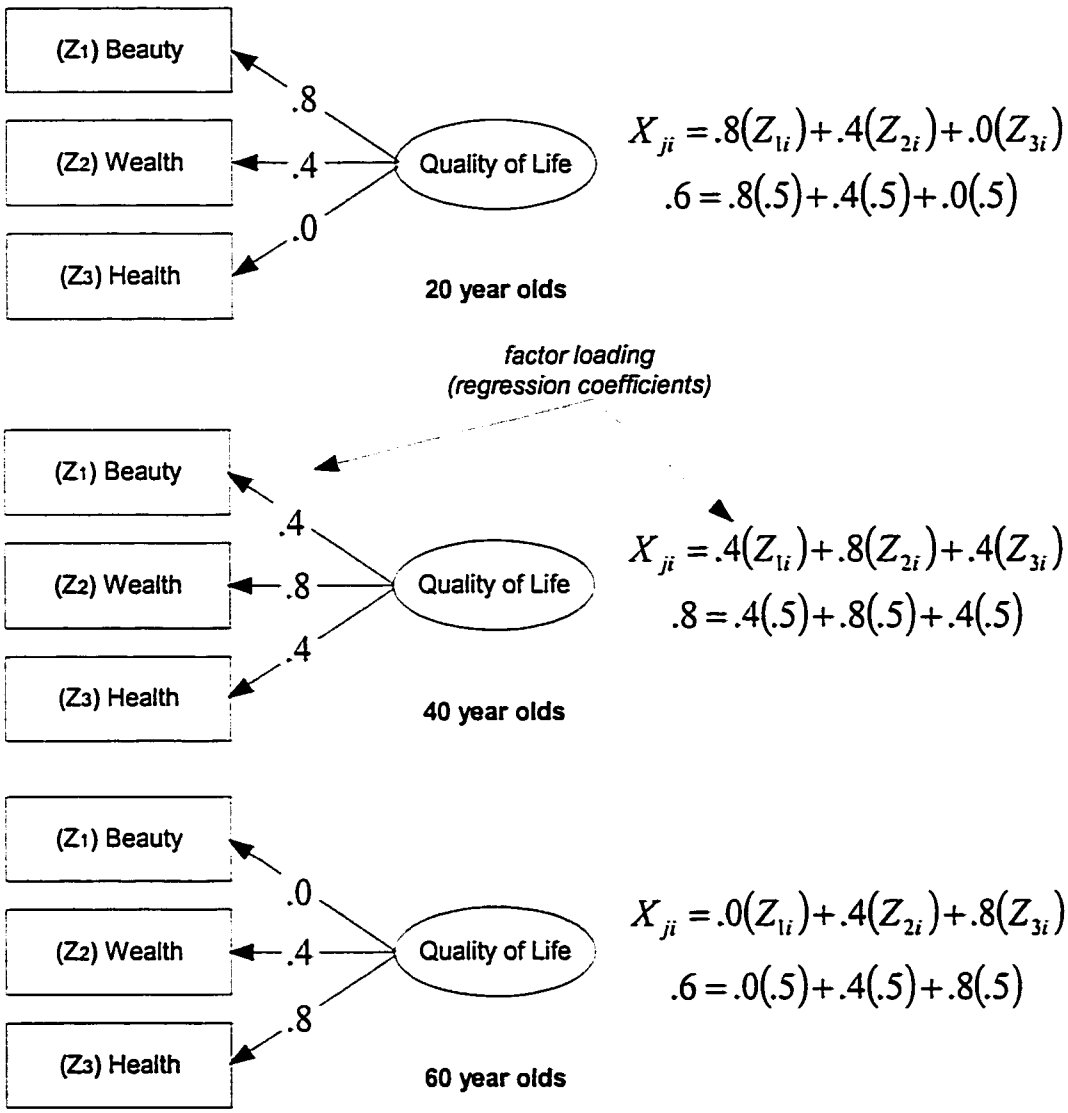


Figure 1. Factor Pattern Example for Items Measuring Quality of Life.

Note. (X_{ji}) are measures of "quality of life," (Z_{ki}) are quality of life item sub-scores.

it was not an indicator for quality of life at all. Alternatively, for middle-aged women wealth was the best indicator, but it contributed less to the construct for both young and older women. In older women, health was the best indicator, yet it was weaker for middle-aged women and not an indicator at all of quality of life for the younger women. Thus, the

constitution of the "mental model" for quality of life, measured as a linear composite, was quite different across age groups as evidenced by the differences observed in factor loading. Fore example, a "yes" response to the question about beauty increases the "quality of life" score for a young woman by weight of .8, that of a middle-aged woman by .4, and does not increase the score for an older woman at all. Structural differences, such as those illustrated in this example, indicate the kind of haziness that arises when attempting to compare qualitatively different constructs. If the sample means on each item were .5 for each of the three groups compared, the evaluator in this example might wrongly assert that middle-aged women perceive the highest quality of life and that younger and older women perceive the same quality of life. Reflecting on this example further, it is clear that the item mean for each group could be substantial, but that does not speak to whether each item contributes equally to the "quality of life" construct—item means do not contribute to item weight.

"If statements about quantitative change are to be unambiguous, it is important that the elements of the composite measurements be invariant across the situations over which change is said to occur" (Horn, 1991, p.118). In this simple one-factor example the issue was that the data failed to demonstrate invariance of the factor pattern matrix and thus the loadings across groups varied. As in this example, when evidence of factorial invariance was lacking, the conclusions of the study may be seriously flawed. To make an accurate assessment and interpretation the evaluator needs to demonstrate that the construct(s) of interest can be measured by the same items and that the units of measure are equivalent across groups, situations, and/or time as determined by the context of the comparison. "Patterns in which loadings of an item change over time [or condition] indicate changes in the *meaning* of the underlying construct" (Pitts et al., 1996, p. 337). Clearly, the generally implicit assumption that the relations among a set of measured items and a given construct are invariant is central to all research involving comparisons or relations among multi-item constructs (Taris et al., 1998).

Background on Factorial Invariance

Factorial invariance has long been used in the field of psychology to investigate the issues of structure (validity) and stability (reliability) associated with longitudinal and cross-cultural data (Byrne, 1998; Drascow, 1987; Drascow & Kanfer, 1985; Frederiksen, 1987; Linn & Harnisch, 1981). The application of factorial invariance as a method, as with the example above, has been primarily directed toward the comparison of groups of individuals on their level of a trait or linear composite construct and to determine whether such scores have different correlates across groups. As shown in the example, for a linear composite score to be comparable across groups, the observed items must have the same relationship with the latent variables for each group of interest, so that the units of measure, or the scale and the scale's interpretation are assured to be the same (Meredith, 1993). Over time two main methods, exploratory and confirmatory factor analysis, have emerged to address the measurement issues encompassed. Although the field has begun to explore factorial invariance using item response theory, discussion of this most recent method was beyond the scope of this dissertation (Flannery, Reise, & Widaman, 1995; Reise et al., 1993).

Historical Emphasis on Exploratory Factor Analytic Techniques

Historically, theorists regarded factor invariance as a criterion to establish the validity of the factor analytic method and as such, were concerned with the problem of equivalence among factors identified in separate studies or across sub-groups in the same study (Thurstone, 1935, 1947; Ahmavaara, 1954). Under simple structure restrictions, factorial invariance studies were aimed initially to provide the foundation for more consistent factor analytic results. Factor structure, particularly item loading, was expected to hold equal across measurement waves, conditions, and/or groups.

The need to compare factor structure over samples and sub-samples necessitated

the development of a vast range of comparison methods (see reviews by Pinneau & Newhouse, 1964; Mulaik, 1972; Alwin & Jackson, 1981). Prior to the introduction of advanced computer programs, various heuristic strategies were employed to study invariance between two or more factor structures. The most widely used *ad hoc* methods were developed primarily for results obtained from exploratory factor analysis (EFA). These early methods were for the most part variations on the theme of an index of factor similarity for factors given estimates from two or more samples—such as the coefficient of similarity (Burt, 1939), the coefficient of congruence (Tucker, 1951), and the coefficient of pattern similarity (Cattell, 1947).

Interest in these early methods has declined with the introduction of the means to explore item-factor relationships in a more confirmatory fashion. Since the 1970s factorial invariance has been assessed though the use of confirmatory factor analytic (CFA) techniques which include the "study of similarities and differences in the covariation patterns of item-factor relations" (Windle, Iwawaki, & Learner, 1988, p. 551). A comparative assessment of different exploratory and confirmatory procedures demonstrated that covariance structure analysis was the preferred technique for investigating changes in factor structure (Schmitt, Pulakos, & Lieblein, 1984). The use of CFA as a tool to explore and address measurement issues has become increasingly common in the social and behavioral sciences (Bollen & Long, 1993).

Contemporary Emphasis on Confirmatory Factor Analytic Techniques

The primary benefit of using confirmatory factor analysis (CFA) methods over EFA is that EFA can only be used to compare basic factor structure. In addition, EFA uses a correlation matrix as a starting point—this implies an *a priori* standardization of variables—which results in an underestimation of the differences across groups or situations (Widaman & Reise, 1997). With CFA the evaluator can compare factor structures—as well as the

variances, covariances, and item score reliability among latent variables--thus, conduct a more detailed psychometric analysis and comparison.

Jöreskog (1973) presented a common factor model and later introduced the software application, LISREL, that enabled the investigation of similarities and differences among factor structures across groups using information about the parameters contained in covariance matrices. A detailed application of this early methodology was presented in McGaw and Jöreskog (1971). CFA differs from other multivariate statistical procedures in that it compares the observed covariance matrix with the covariance matrix implicit in a proposed model. The analyst draws on theory to develop the basic structure of the model, then LISREL is used to estimate the parameters describing the relationships between observed indicators and the latent constructs proposed in the model.

In this confirmatory approach, the analyst can assign arbitrary values or constrain parameters to be invariant across particular conditions or groups—such as, factor loadings, factor variances, factor covariances, or error variances—and thus estimate the equivalence of the relationships among variables and factors proposed by a simple model and their fit to the data. What is freely estimated and what is specified as fixed, is subjective and related to the parameters of greatest interest to the study at hand. The number of estimable parameters is related to the issue of identification. The analyst develops the measurement model with the following constraints in mind: (a) scale and interpretation considerations, as well as, (b) the relative importance of variance, covariance, and regression coefficients to the analysis (Maruyama, 1998).

In CFA a measurement model is specified for each of the latent variables proposed by the instrument under investigation. The end result is a model that reflects the "theory" behind the relationships as proposed in the literature and closely agrees with the observed relations between selected indicators and the constructs of interest. Given that the measurement model is focused on the extent to which measured variables are linked to their

underlying latent variables, or factors, it pertains directly to the investigation of factorial invariance. CFA enables the analyst to examine a wide range of degrees of invariance across a variety of parameters from the perspective of hypothesis testing and serves to move factor analysis away from a purely exploratory technique (Millsap & Hartog, 1988).

Testing for factorial invariance has come to be applied in measurement situations that require a range of rigor and interpretation broader than that described in the "quality of life" example. There are several types of factorial invariance with progressively more stringent restrictions that may be tested using CFA techniques. Each type places an increasing number of equality constraints on the parameter estimates derived across groups. Placing additional equality constraints on the parameter estimates increases the strength of the comparative statements that can be made about qualitative and quantitative differences and similarities among factors across groups.

These types of factorial invariance—*configural* (simple structure), *weak* (factor pattern), *strong* (factor pattern and intercept), and *strict* (factor pattern, intercept, and error variance)—can be investigated using LISREL and the CFA model (Horn, McArdle & Mason, 1983; Meredith, 1993; Widaman & Reise, 1997). There are two overlapping dimensions to this factorial invariance testing hierarchy: (a) model form and (b) similarity of parameter estimates (Bollen, 1989). In addition to the configural, weak, strong, and strict typology emphasized by Meredith (1993), several other types of factorial invariance such as variance/covariance and factor mean level can be investigated across groups. Additionally, given that each of these types of factorial invariance connote the extent to which groups share a mental model for the construct(s) under investigation, it is also possible to use these techniques to gather evidence of conceptual change (Golembiewski et al., 1976).

The Analysis of Change Using CFA Techniques to Determine Factorial Invariance

The developmental psychology and organizational development literature suggests

that studies of factorial invariance can be deployed to investigate the nature and extent to which systematic changes in how individuals conceptualize their work occur (Millsap & Hartog, 1988; Schaubroeck & Green, 1989). Studies of factorial invariance have been used to identify and describe three types of change: (a) *alpha* change—changes in factor score indicate that the magnitude or level of a phenomenon has changed, (b) *beta* change—the magnitude of factor loadings and factor variances can indicate that a variation or recalibration across a conceptual domain has resulted in a change in the weight or clarity of perception, and (c) *gamma* change—a shift in the pattern of factor loadings or relationships among factors can indicate a redefinition of the conceptual domain resulting in a different frame of reference for a given domain (Schaubroeck & Green, 1989; Taris et al., 1998).

Historical Background for Alpha, Beta, and Gamma Conceptual Change

The theoretical framework for the idea that the "mental model" (structure or organization of thought) driving a given stage of development undergoes a transformation to become a more mature structure that embodies the next stage began in the late 1970's with the work of Golembiewski, Billingsley, & Yeager (1976). This work sprang from the observation that interventions often attempt to change both organizational functioning and the individual's perceptions or conceptualizations of this functioning (Millsap & Hartog, 1988). In particular, Schmitt's (1982) study demonstrated that experience in a work environment could systematically shift or transform response patterns in ways that alter the meaning of work-related concepts over time. There is a relative dearth of studies from the late 1980s through the mid-1990s where upon the role and importance of factorial invariance studies are experiencing a revival as evidenced by the intriguing papers by Taris et al. (1998), and Pitts et al. (1996). These most recent applications of CFA techniques to the study of factorial invariance and change processes served as the catalyst to initiate interest in the evaluation specific methods investigation detailed in this dissertation. It is important to acknowledge

which type of change has occurred as a result of an intervention or treatment if effectiveness is to be unambiguously assessed (Terborg, Howard, & Maxwell, 1980).

Two CFA Approaches—Covariance Structure and Moment Structure Modeling

The most frequent approach to factorial invariance testing found in the literature relies on the use of CFA techniques to model and test the equivalence of covariance structures (Jöreskog, Sörbom, du Toit, & du Toit, 1999). The traditional approach relies on the common factor model where each item or measured variable, y_{ji} is represented as the raw score deviation for person i from the mean of variable j. In addition, each measured variable is defined as a linear function of one or more latent variables, η_k (factors) and stochastic error, ε_{ji}. In the traditional CFA approach to factorial invariance the relationship of a measured variable to respective latent variables is described in equation (1) below; whereas the matrix equation describing the aggregate condition for p measured variables is described in equation (2):

$$y_{ji} = \lambda_{j1}\eta_{1i} + \lambda_{j2}\eta_{2i} + \Lambda + \lambda_{jm}\eta_{mi} + \varepsilon_{ji} \qquad (1)$$

$$y = \Lambda\eta + \varepsilon \qquad (2)$$

Equation (3) below describes the multiple-group linear covariance structure used with traditional CFA modeling where S is the $(p \times p)$ observed sample covariance matrix for measured variables and the $\hat{\Lambda}_g$, $\hat{\Phi}_g$, and $\hat{\Theta}_{\delta g}$, and $\hat{\Sigma}_g$ matrices contain sample estimates of the population parameters. $\hat{\Lambda}_g$ is the $(p \times m)$ matrix of the loadings of p measured variables on m latent variables and $\hat{\Lambda}_g'$ is this matrix transposed. $\hat{\Phi}_g$ is the $(m \times m)$ matrix of covariances among the factor scores and $\hat{\Theta}_{\delta g}$ is the $(p \times p)$ matrix of covariances among the measurement residuals. $\hat{\Sigma}_g$ describes the $(p \times p)$ matrix of covariances among the population estimates of p measured variables. The g subscript

indicates that the matrices described were derived from the g th group (Widaman & Reise, 1997).

$$S_g \cong \hat{\Lambda}_g \hat{\Phi}_g \Lambda'_g + \hat{\Theta}_{\delta g} = \hat{\Sigma}_g \qquad (3)$$

Typically, this approach to the investigation of factorial invariance does not include the $\hat{\tau}_g$ (measured variable intercepts) nor the $\hat{\kappa}_g$ (factor mean) matrices; however, Meredith (1993) drew the distinction that failing to include these matrices enabled only the testing of the less stringent forms of factorial invariance. Inclusion of these matrices requires that moment structure models be employed. Measurement models based on moment matrices are analyzed using LISREL in the same manner as covariance structure models, except that moment matrices are "raw-score cross-products matrices among measured variables" (Widaman & Reise, 1997, p. 290). The respective mean and standard deviations for the measured variables are input to LISREL and the software calculates and evaluates the moment matrices (Jöreskog, et al.,1999).

These more stringent forms of factorial invariance require that the item intercepts and factor means be considered. To do this, one includes τ_j which is the intercept for predicting the observed variable y_{ji} from the latent variables η and κ_k is the mean for factor k. The y_{ji} score is retained in its raw form, rather than as a deviation score, and equation (4) and matrix equation (5) are then rewritten as:

$$y_{ji} = \tau_j + \lambda_{j1}(\kappa_1 + \eta_{1i}) + \lambda_{2j}(\kappa_2 + \eta_{2i}) + \Lambda\ \lambda_{jm}(\kappa_m + \eta_{mi}) + \varepsilon \qquad (4)$$

$$y = \tau\mu'_\tau + \Lambda(\kappa\mu'_\alpha + \eta) + \varepsilon \qquad (5)$$

The general equation for estimating parameters and assessing factorial invariance across groups using moment structure model CFA as proposed by Meredith (1993) is as follows in equation (6):

$$M_g \cong \hat{\tau}_g \hat{\tau}'_g + \hat{\Lambda}_g \left(\hat{\kappa}_g \hat{\kappa}'_g + \hat{\Phi}_g \right) \hat{\Lambda}'_g + \hat{\Theta}_{\delta g} = \hat{M}_g \qquad (6)$$

The addition of the $\hat{\tau}_g$ and $\hat{\kappa}_g$ matrices to the multiple group analysis enables the testing of more rigorous forms of factorial invariance. When traditional approaches are used, although the $\hat{\Lambda}_g$ matrices may be found to be invariant, without the inclusion of a test for equality of the $\hat{\tau}_g$ matrices the evaluator is unable to ascertain from among a number of possible linear combinations whether that identified for each group is equivalent.

The Interpretative Implications of Factorial Invariance Using the General Linear Model

A simple illustration of the general linear model for a single factor is presented in Figure 2. As can be seen in Figure 2, under the conditions of *configural* factorial invariance,

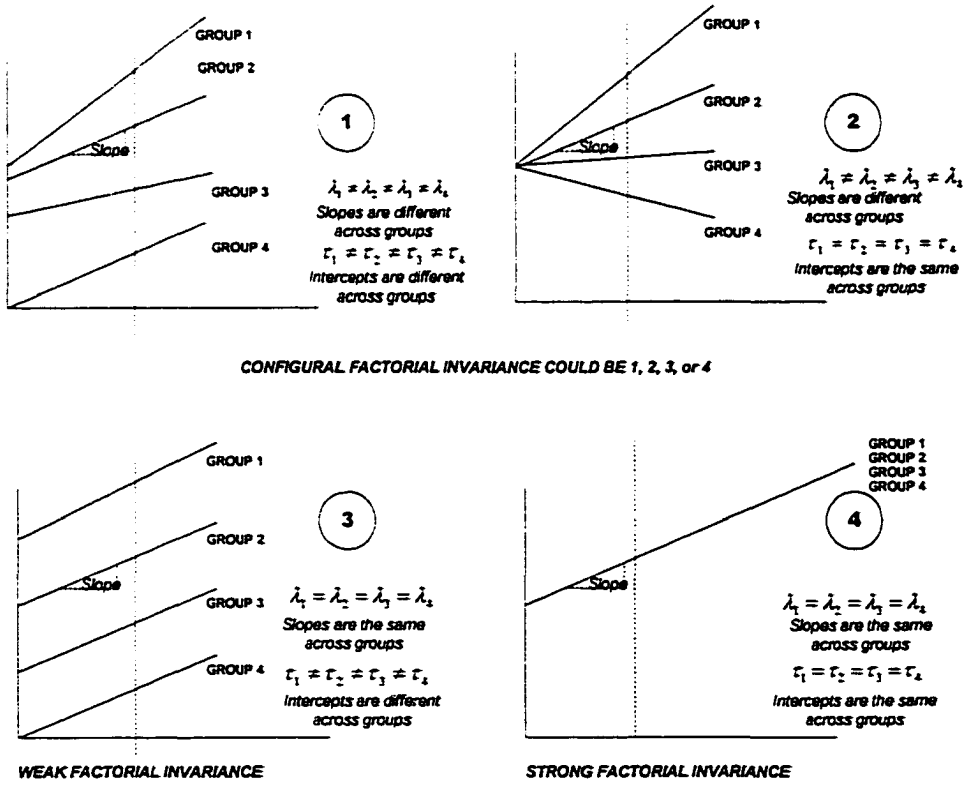


Figure 2. Strong Factorial Invariance Supports Unambiguous Interpretation of Group Differences.

one can only be assured that the items load on the same factors, it is uncertain whether the loadings (λ) and/or intercepts (τ) are the same across groups. Accordingly, with *weak* factorial invariance, as described in quadrant 3, the loadings are equivalent but there is uncertainty as to the equivalence of the intercepts. Only under conditions of *strong* factorial invariance (quadrant 4) or higher can one be certain that the linear equations are equivalent. This condition is also referred to as ARF invariance, or appropriate rescaling factors, such that "any method of identifying a model will provide substantively invariant interpretations of across-group differences in factor means and variances" (Widaman & Reise, 1997, p. 295). Thus, the interpretation of similarities or differences across groups under more ambiguous conditions is severely compromised. Addition of the $\hat{\kappa}_g$ matrices allows for comparison of factor means across groups when the conditions of *strong* factorial invariance are established, not possible using the traditional covariance structure method.

Configural Factorial Invariance Addresses Issues of Construct Meaning in Terms of Gamma Change

When testing for *configural* invariance, only the patterns of zero and non-zero loadings that comprise the $\hat{\Lambda}_g$ matrices are constrained to equality across groups, whereas the elements of the intercept, factor loading, variance/covariance, error variance, and factor mean matrices are free to vary across groups (Horn, et al., 1983; Widman & Reise, 1997). This is the same as achieving simple structure across groups where latent variables are similar but not identical. Under the conditions of configural factorial invariance items load on the same factors across the groups and/or conditions being compared. Should the test for this type of factorial invariance fail, as it most likely would for the "quality of life" example provided earlier, the interpretation of group differences to any extent is severely impaired in that this breach of simple structure provides evidence that the constructs being measured are not perceived the same across groups.

This shift in the pattern of factor loadings, which is indicative of a different frame of reference, or a redefinition/reconstitution of the conceptual domain, is one aspect of *gamma* change (Golembiewski et al., 1976; Schaubroeck & Green, 1989; Taris et al., 1998). Gamma change results when respondents adjust their understanding of the criterion being measured such as, a major change in perspective or a shift in their frame of reference for classifying the relevance of a construct (Golembiewski et al., 1976). This type of change is a reconceptualization of what a given behavior includes (Van de Vliert, Huismans, & Stok, 1985). It would be impossible to compare respondents on a phenomenon that changes from one dimension to multi-dimensional over time. An abstract construct may evolve to mean different things to respondents over time, especially if the intervention being evaluated included sessions intended to increase the respondents' understanding of the concept. If the respondents have come to redefine the construct during treatment exposure, their survey responses before intervention may have little relation to their responses after the intervention (Zmud & Armenakis, 1978). In addition, *gamma* change, as detected by failing to uphold the *configural* invariance (number of common factors for a construct) hypothesis, could occur in the absence of treatment as well due to maturation or environmental influences (Millsap & Hartog, 1988). Thus, *gamma* change is what Taris et al. (1998) refer to as "big bang" change—when it occurs comparison across situations has little meaning. There is agreement that *alpha* and *beta* change can neither be empirically measured nor substantively interpreted if *gamma* change has been demonstrated (Van de Vliert et al., 1985).

Weak Factorial Invariance Addresses Issues of Construct Scaling in Terms of Beta Change

When testing for *weak* factorial invariance, the loading elements that comprise the $\hat{\Lambda}_g$ matrices are constrained to equality across groups, whereas the elements of the intercept, variance/covariance, error variance, and factor mean matrices are free to vary across groups (Meredith, 1993; Widman & Reise, 1997). When factor loadings are equivalent

across comparison groups it means that the groups weight the items the same. Should the test for this form of factorial invariance fail, the interpretation of group differences, other than those with respect to variance/covariances, are limited in that changes in the magnitude of factor loadings across groups indicate a variation or recalibration of scale across the conceptual domain (Van de Vliert et al., 1985).

This kind of conceptual change is one aspect of *beta* change (Golembiewski et al., 1976; Schaubroeck & Green, 1989; Taris et al., 1998). Most often *beta* change refers to the situation where respondents experience a limited change in perspective of some kind. "People may make different estimates of reality, given clearer (or just different) perceptions of what is happening, or they may highlight different aspects of this reality" (Taris et al., 1998, p. 302).

Beta change, an internal threat to validity, has been referred to as instrumentation bias by Campbell and Stanley (1966). In the face of *beta* change comparison of pre- and post intervention survey data will present a biased picture of the effectiveness of the intervention. Howard and Dailey (1978) demonstrated that this response-shift bias frequently occurs as a result of treatment where subjects are more able to accurately assess their real level of functioning on a given construct. In such cases, changes in measurement scale result in respondents' relative overestimation of their level of functioning at pre-test (Schaubroeck & Green, 1987; Schmitt, 1982).

Strong Factorial Invariance Also Addresses Issues of Construct Scale in Terms of Beta Change

When testing for *strong* factorial invariance, the loading elements that comprise the $\hat{\Lambda}_g$ matrices and the measured variable intercepts (item means) that make up the $\hat{\tau}_g$ matrices are constrained to equality across groups, whereas the elements of the variance/covariance, error variance, and factor mean matrices are free to vary across groups

(Meredith, 1993; Widman & Reise, 1997). *Strong* factorial invariance supports the hypothesis that the entire linear model that describes the relationship among latent variables to a given measured variable, in terms of both the regression weight (loading) and the intercept term is invariant across conditions compared.

For most substantive research and evaluation questions, constraints on the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices are considered crucial in that this condition establishes that the same latent variables or factors are identified for each group under comparison (Meredith, 1993; Widaman & Reise, 1997). Should the test for this form of factorial invariance fail, the interpretation of group differences, other than those with respect to variance/covariances, are limited.

Evidence of this type, when the *strong* factorial invariance hypothesis fails, reflects the kind of conceptual change referred to as *beta* change (Golembiewski et al., 1976). As was the case with the weak factorial invariance hypothesis example presented previously, this type of conceptual change is also one of measurement scale recalibration.

Strict Factorial Invariance Addresses Issues of Construct Reliability

Testing for *strict* factorial invariance requires that the $\hat{\Theta}_{\delta g}$ matrices be constrained to equality across groups, in addition to the previous constraints prescribed by the *strong* factorial invariance condition. Invariance of the diagonal elements of the $\hat{\Theta}_{\delta g}$ matrices determines the extent to which measurement error is equivalent across groups. When this condition holds, any differences observed across groups in means and variances on the measured variables are a function only of the differences across groups in the means and variances of the latent variables. This condition is not often met with most data sets and it is reasonable to expect that the $\hat{\Theta}_{\delta g}$ matrices will vary across groups under sampling from a population (Meredith, 1964, 1993). Failing to exhibit *strict* factorial invariance does not present serious interpretation problems because group differences are still ARF invariant if

strong factorial invariance holds (Widaman & Reise, 1997). Meeting the condition of *strict* factorial invariance is not required for substantive interpretation of group differences.

Under those rare conditions where variances are also equivalent across groups, the information from the test for *strict* factorial invariance can indicate whether item reliability is also equivalent across groups. This factorial invariance test provides evidence of another form of *beta* change where scales exhibit different error that may be dependent on situation—such that respondents may not be equally well able to understand and provide answers to the items across comparison groups (Taris et al., 1998).

Covariance/Variance Factorial Invariance Addresses Issues of Construct Boundaries

Under those factorial invariance testing conditions where the minimum condition of *strong* factorial invariance has been met, it is possible to proceed to investigate additional forms of factorial invariance of interest to the evaluator. Testing for *covariance/variance* factorial invariance requires that the $\hat{\Phi}_g$ matrices be constrained to equality across groups, in addition to the previous constraints prescribed by the *strong* factorial invariance condition ($\hat{\Lambda}_g$ and $\hat{\tau}_g$ invariant). Factorial invariance of the $\hat{\Phi}_g$ matrices should not be expected, nor is it a precondition for interpretation of mean and other parameter differences across groups (Meredith, 1964, 1993).

Invariance of the off-diagonal elements of the $\hat{\Phi}_g$ matrices determines the extent to which the covariances among the factors are equivalent across groups. When this factorial invariance hypothesis fails it means that respondents may have come to see a greater integration (covariance increase) or dissonance (covariance decrease) among the components of the conceptual domain. This shift in the boundary of meaning for the constructs under investigation is another instance of *gamma* conceptual change (Taris et al., 1998).

Invariance of the diagonal elements of the $\hat{\Phi}_g$ matrices determines the extent to which the variances among the factors are equivalent across groups. When this factorial invariance hypothesis fails it means that respondents have come to perceive more (variance increase) or less (variance decrease) of a difference in the constructs across groups. This signal of difference in the amount of disagreement across groups or recalibration of scale intervals is another instance of *beta* conceptual change (Taris et al., 1998).

Factor Mean Factorial Invariance Addresses Issues of Magnitude

Lastly, factorial invariance constraints may be placed on the $\hat{\kappa}_g$ matrices. This type, which requires the precondition of *strong* factorial invariance ($\hat{\Lambda}_g$ and $\hat{\tau}_g$ invariant), tests the equivalence of the factor means across groups. As with the $\hat{\Phi}_g$ matrix situation described previously, investigation of $\hat{\kappa}_g$ group differences under conditions that are not ARF invariant will have as Widaman & Reise (1997) assert, "no direct substantive interpretation" (p. 298).

Under those conditions where the invariance of the $\hat{\kappa}_g$ matrices fails to hold, some degree of *alpha* change has occurred. *Alpha* change involves variations in the reported level or magnitude of a construct that are neither related to any shift in respondents' understanding of the meaning for the construct nor changes in the measurement scale along which the construct is gauged (Golembiewski et al., 1976). This type of change is actual increase or decrease in a particular attitude, trait, or behavior as determined by an examination of mean differences across groups (Schaubroeck & Green, 1989; Van de Vliert et al., 1985).

Table 1 summarizes the four types of factorial invariance presented in the literature and which CFA matrices must be equivalent across groups. In addition, each type of factorial invariance encountered is matched to the appropriate type of conceptual change posited.

Literature Relevance to the Research Conducted

Thus, the CFA approach to the investigation of factorial invariance reviewed here

Table 1

Comparison of the Four Types of Factorial Invariance Discussed in the Literature

Type	Definition	CFA Model (Equality Contraints)	Interpretation If Invariance Not Established		References
			Group Differences	Conceptual Change	
Configural	Simple structure is met. Latent variables are similar, but not identical.	The same pattern of zero and nonzero loading. Values vary.	Severely compromised. Used as a baseline.	Gamma	Widaman & Reise (1997)
Weak	λ are equal for all items on their respective factors.	$\hat{\Lambda}_g$ matrices.	Variance/covariance on the latent variables only.	Beta	Meredith (1993)
Strong	λ and τ for each of the measured variables are equal.	$\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices.	Variance/covariance *and* level of means on the latent variables.	Beta	Meredith (1993)
Strict	λ, τ, and θ for each of the measured variables are equal.	$\hat{\Lambda}_g$, $\hat{\tau}_g$, and $\hat{\Theta}_g$ matrices.	Differences on the measured variables attributable to group differences on the common factors.	Beta	Meredith (1993)
Covariance/ Variance	Complex constraint on factor variances and covariances.	$\hat{\Lambda}_g$, $\hat{\tau}_g$, and $\hat{\Phi}_g$ matrices.	Variance/covariance *and* level of means on the latent variables.	Gamma/Beta	McArdle & Nesselroade (1994)
Factor Means	Factor means constrained to equality.	$\hat{\Lambda}_g$, $\hat{\tau}_g$, and $\hat{\kappa}_g$ matrices.	Variance/covariance *and* level of means on the latent variables.	Alpha	Widaman & Reise (1997)

Note. Shaded row denotes the minimum factorial invariance condition required for substantive interpretation of group mean differences.

provides a unique perspective for detecting and exploring the range of conceptual changes—reconstitution and recalibration—that are theorized to accompany shifts in science teaching practices desired by systemic educational reform. The procedures described here build on prior work examining conceptual change (Golembiewski et al., 1976; Millsap & Hartog, 1988; Schmitt, 1982; Taris et al., 1998; Thompson & Hunt, 1996) and extend it into areas of the reconstruction of meaning that have not been widely tested (Louis, 1980; Senge, 1990).

CFA provides a powerful tool for evaluators to portray a richer view of the transformation that occurs during change directed initiatives such as systemic reform (Mayer, 1999; Spillane & Zeuli, 1999). In addition to investigating the quantitative changes in the magnitude of relevant factors, it will be possible to examine changes in the qualitative meaning of those factors and the boundaries of their inter-relationships. The application of methods such as these will illuminate the role of factor structure modification in interpreting changes in factor score means built from self-report variables (or the lack of such changes). In addition, these techniques detect mean differences while controlling for changes in intercepts, loadings, error, variances, and covariances across groups. Thus, evaluators will be able to assess the impact of the reconstruction of concepts on the interpretation of mean differences, disentangle the different quantitative and qualitative aspects of the change process, and increase the explanatory power of their findings.

CFA models are not ends in themselves. Even if one detects differences between groups in crucial CFA model parameters, the CFA models do not indicate *why* these differences occur. These models however, can be used to isolate the ways in which groups differ on variables, providing a concise statistical representation of group differences and thus serve as a springboard for additional research designed to identify the sources of group differences on the latent and measured variables. The next chapter describes the methodology for the nested CFA approach to the investigation of factorial invariance applied in the execution of this study.

CHAPTER 3

METHODOLOGY

This chapter contains seven sections that present and explain the methods used to execute this research: (a) orientation to the methods investigation; (b) rationale for data set selection; (c) properties of the data set selected for secondary analysis--including description of data collection procedures and the sample, as well as content and psychometric properties of the survey instrument; (d) an overview of the three-part investigation—which includes, (e) preparatory steps required to conduct the analysis, (f) the determination of the model of teaching practice to be used as a baseline for comparison, and (g) the nested set of factorial invariance hypotheses and associated analytic strategies.

Orientation to the Methods Investigation

There are two streams of thought that contribute extensively to the design and execution of the research as presented in this chapter. The primary idea was that changes in measurement structure capture evidence of conceptual change when it has occurred. The second notion was that interventions, such as reform-minded professional development, are intended to evoke conceptual and behavioral changes in participants. This study was based on the confluence of these two ideas. The premise being that should conceptual change occur as a result of participation in professional development—it can be captured by evidence of alterations in measurement structure determined by CFA techniques. As asserted in Chapter 2, the rejection of factorial invariance hypotheses that severely compromise the comparison of mean factor scores across groups and situations provide an as yet unexplored opportunity to apply these rather abstract measurement notions to the evaluation of change.

Changes in Measurement Structure Capture Evidence of Conceptual Change

The secondary analysis of cross-sectional survey data conducted by this study utilized confirmatory factor analytic techniques to investigate the issues and implications of factorial invariance. Under those conditions where measurement structures fail to demonstrate factorial invariance (i.e., specific model parameters constrained to equality across groups or situations) it can be said that some form of conceptual change has occurred.

Conceptual change of the two general types discussed in Chapter 2 was to be identified by alterations in the measurement structure of a survey instrument as compared across treatment and control groups. As mentioned in Chapter 2, *beta* change consists of a recalibration—stretching or shrinking—of the measurement scale for a given construct as observed through altered factor loadings (item emphasis influenced by situation), factor variances (variability influenced by situation), or error variances (reliability influenced by situation). Whereas, *gamma* change represents a reconceptualization of the construct as observed through alteration in factor patterns (number of factors influenced by situation) and/or factor covariances (relationship among factors influenced by situation). Under those conditions where *gamma* change has not compromised the ability of the evaluator to compare factor means across groups, *alpha* change, or change in factor magnitude, is also captured by CFA factorial invariance methods.

Structural evidence of conceptual change would be particularly useful for treatments and interventions such as professional development and training that target changes in attitudes and practices across a variety of settings. In that, in addition to *alpha* change, which has traditionally examined and assessed during evaluation, *beta* and/or *gamma* change can also be regarded as a treatment effect. This structural evidence could serve as a valuable, but as yet untapped, intermediate or interim outcome indicator for interventions whose intent

is to evoke conceptual as well as behavioral change. One such venue known to attempt to stimulate and induce conceptual and behavioral change is systemic educational reform-minded professional development (Spillane & Zeuli, 1999).

Changes in the measurement structure of an instrument can be expected to arise most frequently under specific conditions such as those where treatments "explicitly target abilities or knowledge related to the constructs of interest" (Pitts et al., 1996, p. 346). Clearly, professional development is just such a treatment, in that it targets both knowledge and abilities. Changes in teachers' perception of classroom practice—from an emphasis on traditional, teacher-centered methods to an emphasis on those that are more constructivist and student-centered—are an anticipated outcome of systemic science educational reformed-minded professional development (Elmore, Peterson, & McCarthy, 1996; Spillane & Jennings, 1997; Spillane & Zeuli, 1999). Given the emphasis of this specific professional development approach on fostering perceptions and practices aligned with the systemic reform agenda, there is reason to expect that such an intervention may indeed evoke changes in the way teachers perceive and report on their classroom practice (Mayer, 1999; Smithson & Porter, 1994).

Application to the Evaluation of Change

When perceptions about a specific construct change, like teaching practice, the mental model held by respondents changes accordingly and thus may result in the kinds of alterations in measurement structure described in the previous chapter. For the purposes of this study, mental model was defined as the deeply entrenched assumptions, generalizations, and metaphors an individual holds about a given construct which result from the interpretation of past experience and which influence behavior (Senge, 1990). In that the measurement structure for a given set of respondents reflects the mental model they hold for the object or behavior under investigation, comparison of measurement structures across groups should

be indicative of the extent to which groups "see" the object or behavior in the same way. This study assessed the extent to which changes in the measurement structure of a self-report measure of teaching practice were associated with increasing levels of exposure to reform-minded professional development.

Quite simply, this study was an attempt to determine the extent to which teachers that have participated in reform-minded professional development envision their teaching practice in the same or different ways as teachers that have not yet been reached. This study framed factorial invariance as an opportunity to determine the extent to which evidence of conceptual change can be detected by CFA and factorial invariance methods.

Rationale for Data Set Selection

The previous chapter established the rationale for using a confirmatory factor analytic approach to investigate measurement issues, such as factorial invariance, and extending its application to the evaluation of change. Given the measurement emphasis of this study it was not only important to identify a data set likely to have captured evidence of the types of conceptual change discussed, but to assure that the data set selected came from a strong evaluation design bolstered by rigorous attention to psychometrically-sound instrument development, as well as be large and representative enough to support the proposed CFA and its interpretation.

Qualitative Differences in Teacher Will and Capacity are Anticipated Reform Outcomes

The first step in executing an analysis of secondary data such as this was to identify a data set that is likely to have captured evidence of the *beta* and/or *gamma* aspects of the conceptual change processes described previously. Should reform-minded professional development impact the will and capacity of teachers as anticipated, these changes could result in *qualitatively* different thinking about teaching and teaching practice.

If so, then it was probable that structural evidence of *beta* and/or *gamma* change would be found in this arena using CFA techniques. It was considered highly likely at the outset of this study that some degree of these types of conceptual change would be associated with the transition from one pedagogical philosophical position to another. Recent research on the transition from old to new ways of thinking about teaching practice indicates that the various aspects of teaching practice come to be weighted and/or organized differently as reform proceeds (Spillane & Zeuli, 1999).

Reform Influences the Balance Between Traditional and Constructivist Practices

In theory, the inquiry-based, constructivist approach to science education sought by systemic reform encourages a balance between content and process but, "because both teachers and the system are learning as they are reforming, the balance between the old and the new may shift as the reform evolves and practice changes" (Goertz et al., 1995, p. 45). Knapp (1997), Spillane (1994), and others report that as teachers come to embrace the tenets of the reform agenda they tend to add new practices to their existing repertoire of traditional methods. These authors assert that teachers come to perceive their classroom practice to include both traditional and constructivist methods but that each are weighted differently in terms of importance, emphasis, and relevance.

At the outset, before exposure to reform-minded professional development, teachers hold mental models that place more emphasis on the frequent use of traditional methods. However, the mental model of classroom practice held by teachers is thought to be subject to change as they become familiar and more comfortable with the reform pedagogy and as new practices are folded into current classroom routines (Knapp, 1997; Spillaine & Zeuli, 1999).

Recalibration and/or Reconceptualization as Outcomes of Reform

As teachers respond to the influence of reform-minded professional development and

begin to embrace the reform agenda and its qualitatively different constructivist teaching practices their mental models may change as well. The frequency with which each type of instructional strategy is used will most likely shift as the transition from primarily traditional practice evolves over time toward a practice that includes an increasing proportion of constructivist and inquiry-based strategies. If conceptual change does occur, the measurement structure for teaching practice may exhibit specific modifications such as those that recalibrate (*beta* change—same model, different emphasis) or reconceptualize (*gamma* change—different models) certain aspects of the classroom culture of inquiry and the investigative practices encouraged by the reform movement.

Recalibration would be considered evidence of *beta* change and could conceivably be measured as differences in the factor loadings of various items that describe and delve into the traditional and constructivist aspects of science teaching practice. On the other hand, as new practices are accepted and included, the ways in which teachers group and relate the various aspects of their classroom practice from among traditional and reform-minded instructional strategies may result in different groupings and/or differences in the strength and direction of the relationships among established groupings. This second case is one of reconceptualization and would be considered evidence of *gamma* change as measured by differences in the number of factors or correlations among the factors that describe the mental model of science classroom practice.

Thus, a data set from a systemic reform initiative that features an emphasis on reform-minded professional development would serve as an ideal candidate for this study. Classroom practice—as an indicator of intermediate systemic reform outcomes—is at the core of systemic reform initiatives at the national level. Many national systemic reform initiatives, primarily those addressing science and mathematics K-12 education sponsored by the National Science Foundation, have engaged in an evaluation process that includes large survey samples of teachers and administrators from actively reforming districts. These

national initiatives provide a number of excellent candidates for the present study.

The Local Systemic Change Initiative (LSC) is Selected as the Best Candidate

One such national initiative, the Local Systemic Change Initiative (LSC) sponsored through the Teacher Enhancement level of the Elementary, Secondary, and Informal Education Division of the National Science Foundation, focuses on the professional development of teachers within whole schools or school districts. The LSC initiative emphasizes the alignment of reform policy and teaching practice. The goal of this initiative is to enhance the teaching of science, mathematics, and technology by "preparing teachers to implement designated exemplary mathematics and science instructional materials in their classrooms" (Weiss, Montgomery, Ridgway, & Bond, 1998, p. 1).

This systemic science educational reform initiative has a strong national and local evaluation strategy that employs a set of clearly defined, measurable intermediate cognitive, affective, and behavioral outcomes for teaching. The rigorous, consistent, outcomes-based evaluation framework for this on-going and expanding initiative insures that high-quality data and information about teaching practice can be aggregated across the forty-six (at the time of this study) individual projects to inform policy at the national level. Thus, given its size, focus, and the integrity of its evaluative framework and methodology, the LSC initiative provided an ideal opportunity to further explore the extent to which the structural manifestations of conceptual change can be captured and interpreted.

Intended Use for Data Sought

This study was not intended to serve as an evaluation of LSC activities or specific LSC projects. LSC data was sought to be used to investigate whether conceptual changes occur in the mental models held by teachers about their teaching practice in association with participation in systemic reform-minded professional development. It was the intent of the

present study to determine if just such a shift could be detected and thus, be employed to add to the toolchest of methods available to evaluators interested in measuring, understanding, and reporting on the process of educational reform.

This study was an attempt to delve deeper into the relationship among teaching practice variables—as a function of exposure to reform-minded professional development—than is routinely possible given the constraints faced by evaluators in the field. Hopefully, this new lens will provide a valuable learning opportunity for evaluators as they become able to focus in on the process of reform and the mental models held by teachers in actively reforming schools and districts. Access to this national evaluation data set was provided through Horizon Research, Inc. (HRI) in Chapel Hill, North Carolina (National Science Foundation RED-92553690).

Properties of the Selected LSC Data Set

Accordingly, the LSC K-8 Science Teacher Questionnaire was selected as the best candidate for the secondary data analysis presented here, specifically because it was constructed to conceptually align with the outcomes sought by systemic science education reform and thus, as an evaluative tool, monitor change. Two issues arose during the initial selection process that influenced the specification of the actual LSC data set used to conduct this study. They were as follows: (a) data collection and sampling procedures used by HRI, which influenced cohort size and developmental sequence—a cohort was needed that was both large enough to support CFA techniques and likely to demonstrate some conceptual change with minimal diffusion effect; and (b) substantive content, data cleaning, and categorization issues—only those items which were most pertinent to the study demonstrating a full range of variance in responses, as well as, only those respondents with answers to all of the selected items could be included in the final data set analyzed.

Description of Data Collection and Sampling Procedures Used by HRI

This self-report survey has been administered annually since 1996 as a part of core evaluation activities for the LSC initiative at the national level. Data collection procedures were developed to ensure high quality data and to protect teacher confidentiality. Respondents were informed that their responses would only be reported in aggregate, that any information identifying individuals would be used for the purposes of administration and non-respondent follow-up, and that no information identifying individuals would be reported under any circumstance. For the purposes of this research, teacher responses were considered anonymous in that no identification information was provided to the researcher conducting the secondary analysis.

A systematic random sample of 300 K-8 science teachers was drawn from each of the projects participating in the science component of the LSC initiative. The sampling frame provided by each project included every teacher who was targeted to be served by the LSC project over the entire period of LSC funding and who was responsible for teaching science in the spring of each year. In those projects with fewer than 350 teachers in the sampling frame, the population of teachers was surveyed.

The secondary data analysis presented here focused on data obtained from Cohort 2 for the 1996-97 school year. Specifically, to avoid the confounding issues of multiple cohorts and survey administration over a period of years, these data from a single cohort and a single year were used. Cohort 2 was selected because it contained the largest number of K-8 Science projects and a CFA study such as the one conducted here needed a fairly substantial sample size. In 1996-97, Cohort 3 projects were collecting baseline data during their first year of funding, Cohort 2 projects were in their second year of funding, and Cohort 1 projects were in their third year. The year 1996-97 was selected because to investigate the nature and extent of the relationship between factor structure and exposure to professional development

some time was required to have elapsed for implementation of the treatment. These data with a one year elapsed treatment opportunity were selected to minimize, given the evaluation constraints, the possible introduction of cultural "cross-pollination" in the untreated teachers.

Data collection activities for the projects' 1996-97 Core Evaluation Reports were originally conducted from September 1, 1996 through August 31, 1997—with the Teacher Questionnaire being administered between March and May, 1997. Fifteen of the sixteen participating districts in Cohort 2 targeted between 500-3000 teachers, with one district targeting fewer than 350.

Substantive Content, Data Cleaning, and Categorization Issues

In particular, the LSC Teacher Questionnaire was comprised of four sections: (a) teacher opinions of reform and perceived preparedness (85 items), (b) teaching practice (61 items), (c) LSC professional development (9 items), and (d) teacher demographic information (5 items). These 160 items were used to construct 12 composite scales that covered the five domains pertinent to the questions posed by the LSC evaluation. However, this study excerpted only those 40 items that specifically inquired about the frequency of a variety of traditional and constructivist teaching practices (Items 10a-m and 11a-z) and that described the amount of exposure to the LSC professional development (Item 16). Given systemic educational reform theory, these items were identified as those most likely to provide evidence of conceptual change in response to treatment in the ways previously described (a photocopy of the complete instrument is provided in Appendix A and a photocopy of the permission granted from HRI to include the instrument is provided in Appendix B). The reader should now be aware that the study reported here was performed on a specific, much smaller sub-set of items abstracted from the entire LSC instrument.

The size of the archival 1997 LSC data set after list wise deletion was 2272 teachers. Only those cases with responses to each of the 40 selected items were included in the

analysis. For the purposes of this research teachers were categorized as either control or treatment group according to the amount of professional development they reported having participated in to date (Item 16): (a)Control-Group 1, 0 hours (n=666), (b) Treatment-Group 2, 1-19 hours (n=813), (c)) Treatment-Group 3, 20-39 hours (n=300), and (d)) Treatment-Group 4, 40+hours (n=493). In addition, the demographic characteristics of the teachers and their participating schools were reported by HRI to roughly approximate the national population (Weiss et al., 1998).

It is also important to note that the analytic and scoring approaches used here were different from those described and used by HRI to conduct and report on their evaluation of LSC activities. Composite scores for the HRI analysis and reporting were calculated as percentages of total points possible. "An individual teacher's composite score is calculated by summing his/her responses to the items associated with that composite and then divided by the total points possible" (Weiss et al, 1998, p. 8). Factor scores, as calculated using CFA, were used in the analysis reported here.

Description of the Survey Instrument

Teaching practice was but one of five domains covered by the 12 composite scales that comprised the LSC K-8 Science Teacher Questionnaire. In that the secondary analysis conducted here employed an existing survey instrument, it is necessary to provide the reader with some detail on the extent to which information was available on the processes used to construct and validate the instrument, the frequency response set used, and the composition of the *teaching practice* composite scales that were excerpted from the full instrument to conduct this methods study.

Construction of the HRI Instrument

To develop the teaching practice section of the survey, the HRI evaluation team

operationalized *teaching practice* to represent traditional practices as well as the extent to which the reform ideals of constructivist teaching for depth over breadth, creating a culture of inquiry, and employing investigative learning strategies were evident in a teacher's self-report of the frequencies with which various instructional strategies were employed in the classroom. In that reformed teaching practice represents one of the key intermediate outcomes sought by systemic reform (Shields, et al., 1995) a conceptual Table of Specifications[1] based on the system reform literature was used to identify the areas required to adequately describe and bound the spectrum of traditional through reformed instructional practices anticipated to be employed by science education teachers. A total of 39 items were included in the *teaching practice* section of the 1997 LSC K-8 Science Teacher Questionnaire. A brief description of the response set and the manner each teaching practice construct was specified by HRI follows.

A five-point, Likert-type, frequency scale was used for all of the teaching practice items, reflecting both teacher and student classroom activities, selected to serve as indicators of the three teaching practice constructs—traditional, investigative culture, and investigative practice). The response set for the teaching practice items was as follows: (a) never, (b) rarely (e.g., a few times a year), (c) sometimes (e.g., once or twice a month), (d) often (e.g., once or twice a week), and (e) all or almost all science lessons.

HRI operationalized *traditional teaching practice* as was defined in the systemic reform literature. Traditional teaching practice was teacher-directed with the teacher at the center of all activities (Evans, 1996). Desks in rows, set class periods, and heavy reliance on lecture and textbooks characterize what for the most part is a passive learning environment (Gabella, 1995; Rallis, 1995). For example, items developed to measure the extent to which a teacher relies on this mode ask for the teacher to report the frequency with which they: (a)

[1] Evidence of a formal Table of Specifications was not present in the Technical Report provided by HRI.

lecture-Q10a, (b) assign science homework-Q10l, (c) have students answer textbook/worksheet questions-Q11g, or (d) have students take short answer tests (e.g., multiple choice, true/false, fill-in-the-blank)-Q11x (see Appendix A).

Under a teaching and learning environment that embodies an *investigative culture*, students construct their understanding of the fundamental ideas and processes of science by direct encounter with each other, materials, resources, and experts (Brooks & Brooks, 1993; Driver, Asoko, Leach, Mortimer, & Scott, 1994; Fosnot, 1993). Student-centered learning acknowledges that learning is not passive and that students are expected to participate and contribute to their own investigative learning experience (Fullan, 1995; Khattri & Miles, 1995; Rallis, 1995). "The teachers question and probe—to help children make meaning—rather than to direct. They listen carefully, encouraging reflection and stimulating new connections and interpretations" (Rallis, 1995, p. 226).

The hallmark of *investigative culture* employed by HRI was the student-centered classroom, where the teacher provides a model of instruction that enables learners to interact with each other. For example, items developed to measure the extent to which a teacher creates an investigative culture ask for the teacher to report the frequency with which they: (a) require students to supply evidence to support their claims-Q10e, (b) encourage students to consider alternative explanations-Q10g, (c) have students work in cooperative groups-Q11c, (d) write reflections in a journal-Q11s, (e) read non-textbook reference materials-Q11f, or (f) use mathematics as a problem-solving tool-Q11u (see Appendix A).

The term "inquiry-based science education" is commonly used to describe the new vision reformers hold for the teaching of science (Gabella, 1995). Students will model the scientific method of discovery and in so doing move their learning beyond the rote storage and retrieval of factual knowledge. Movement away from textbooks and the memorization of facts toward this vision—which includes making observations; posing questions, planning and conducting investigations, using tools to gather, analyze, and interpret data, and

communicating the results—will require teachers to be knowledgeable about a wide range of pedagogy (National Science Foundation, 1995, 1997; National Science Teachers Association, 1997). *Investigative practices* are characterized by students engaging in activities designed to promote cognitive and conceptual development of scientific ways of thinking and knowing. These practices are more action oriented and aligned with contemporary constructivist thought. For example, items developed to measure the extent to which a teacher uses *investigative practices* ask for the teacher to report the frequency with which they have students: (a) design or implement their own investigation-Q11m, (b) work on models or simulations-Q11o, (c) participate in field work-Q11q, or (d) work on extended science projects-Q11p (see Appendix A).

Overview of the Psychometric Properties of the Instrument as Used and Reported by HRI

To assure psychometric quality and to simplify the reporting of large amounts of survey data, HRI used reliability statistics, including item-total correlations and Cronbach's α to determine the extent to which each composite measure was a robust measure for each teaching practice construct. For the purposes of evaluation reporting HRI retained 20 of the instrument's 39 teaching practice items in their analyses. Description of the instrument validation and scaling processes executed by Flora & Panter (1998) does not provide specific information on why 19 items were excluded but the procedures reported to have been used on the set HRI did include are provided as background attesting to the attention to psychometric detail that underlies the LSC survey. It is important that the reader not confuse the background description of the instrument properties as performed and reported by the HRI team with later analyses performed by this researcher for the purposes of the conducting the methods-focused secondary analysis.

The item-level factor analysis solution[2] for *teaching practice*, briefly described by Flora & Panter (1998) in their Technical Report: Analysis of the Psychometric Structure of the LSC Surveys, supports the *a priori* three factor dimensionality of classroom practice established in the systemic educational reform literature (Hirsch, 1996; National Science Resources Center, 1997; Regional Educational Laboratories, 1995; Rhoton & Bowers, 1996; St. John, Century, Tibbetts, & Heenan, 1995). Flora & Panter (1998) report that once the three teaching practice composites were affirmed by factor analysis, a measurement model was proposed for each construct and tested on a random subsample of the data using confirmatory factor analysis. Flora & Panter (1998) report that CFA[3] provided further evidence to the LSC evaluation team at HRI to support that each teaching practice composite represents a single factor as hypothesized *a priori* by the inferred Table of Specifications during instrument development. Flora & Panter (1998) performed two additional CFAs on random subsamples of the data to establish that the factor structure arrived at from the previous analyses could be cross-validated and to compare factor structures across the three annual administrations of the survey to date.

Given the insufficiency, due to the incomplete and summary nature of a report targeted toward non-technical readers, of the information provided in the instrument's Technical Report, only the assertions made in the report as to the quality and properties of the data are mentioned here. At this point, however, the reader should accept that that the Technical Report provides evidence that such psychometric studies, as were required by the HRI team to support the construction of three teaching practice composites, were performed in an acceptable manner with Cronbach's α for each of the three composites reported at

---

[2] Information on the EFA solution, such as the size of eigenvalues or communality estimates, was not available in the Technical Report provided by HRI.

[3] Information on the various CFAs performed such as model identification, chi-square values, *p*, nor other fit statistics was not available in the Technical Report provided by HRI.

values greater than 0.8[4]. This researcher could not attest to the specific instrument properties based on the Technical Report provided, only that more effort went into the determination of instrument quality than is generally found in the field and practice of evaluation.

HRI reports that these data, considered in aggregate, indicated that the constructs shared quite similar content over three years of survey administration; however, those items not shared by all three surveys were excluded from the comparative studies reported. Specific factorial invariance studies, such as those proposed herein, were not performed by HRI as of early 1999 when this method study was conceptualized (D. Flora, Personal communication, February, 1999).

Overview of the Investigation of Factorial Invariance

What follows in this overview is description of the researcher's access and storage of the data and an introduction to the processes used to conduct the research. A flowchart is used to illustrate the three aspects that comprised the research performed: (a) preparation, (b) measurement baseline, and (c) determination of invariance. The three parts to the study are described briefly in this overview and then in greater detail.

For the limited purposes of this secondary analysis, only the 39 teaching practice items (survey items 10 a-m and 11 a-z) and the 1 item identifying treatment amount, as measured by reported exposure to professional development (survey item 16) were provided by HRI as an electronic attachment via email to the researcher. This 40-item file was then stored as required by the HSIRB of Western Michigan University for the duration of the study (a photocopy of the Western Michigan University Human Subjects Institutional Review Board approval is provided in Appendix C). A research process flowchart is provided in Figure 3.

---

[4] See Appendix C. *Traditional Practices*-Q10l, Q11g, Q11h, Q 11x (α=0.83). *Investigative Culture*-Q10c, Q10d, Q10e, Q10f, Q10g, Q11b, Q11c, Q11j (α=0.89). *Investigative Practices*-Q11d, Q11k, Q11m, Q11o, Q11p, Q11q, Q11s, Q11z (α=0.82).

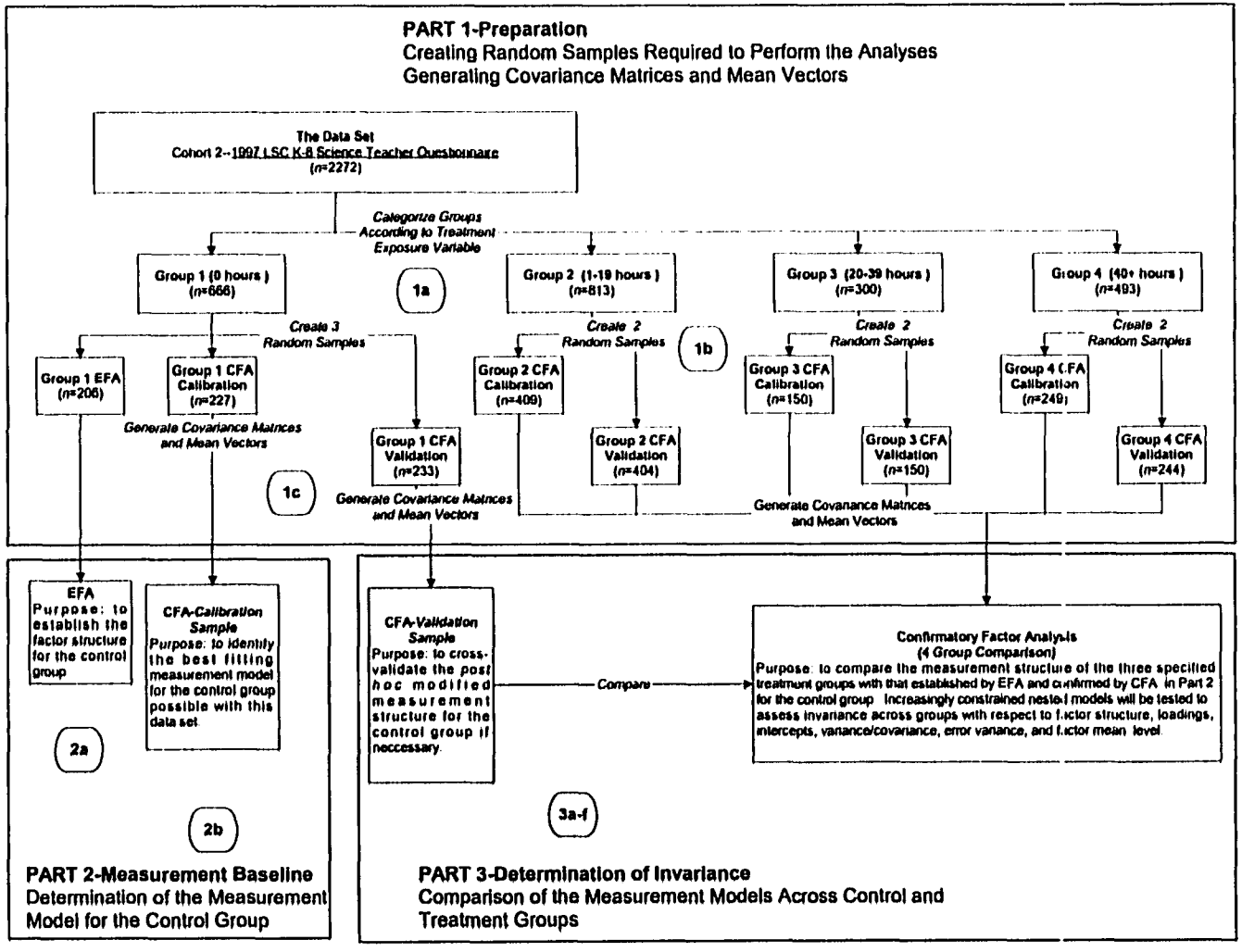


Figure 3. Flowchart for the Secondary Data Analysis.

Part 1 of this study was basically a preparatory phase performing the data screening and organizing procedures needed to execute the study. Step 1a categorized and created separate files for respondent groups according to the level of exposure to professional development reported. Step 1b created the random samples from each category needed to conduct the analyses. Step 1c generated the covariance matrices and mean vectors from each sample as well as the fixed thresholds from the entire data set needed to conduct the second and third parts of the study (see the top section of Figure 3).

Part 2 of this study consisted of a replication of the EFA and CFA performed by Flora & Panter (1998) during instrument development and described in the previous section (see the bottom left hand side of Figure 3). As seen in step 2a, a random sample of data from the control group respondents was used to perform exploratory factor analysis to determine the measurement structure for the entire set of 39 teaching practice items. In step 2b, the measurement structure specified by EFA was used to select the best fitting items for each factor. A smaller more manageable model was then confirmed using CFA for the control group of respondents that served as the baseline for the invariance study performed in Part 3.

Investigation of factorial invariance in Part 3 proceeded by determining the extent to which there was evidence to suggest that the respondents in Groups 2 through 4 shared the same mental model, or measurement structure, as those respondents in the control Group 1 (see the bottom right hand side of Figure 3). This consisted of testing the extent to which a series of nested factorial invariance hypotheses hold that constrain various model parameters to equality across control and treatment groups, as described in the literature review. In all, a set of six increasingly more restrictive CFA models were analyzed to assess invariance for factor pattern, factor loadings, intercept, variance/covariance, error variance, and factor mean level--listed in increasingly restrictive or constrained order. In CFA investigations of factorial invariance such as this, the terms *constrained* and *restrictive* are used somewhat interchangeably to refer to extent to which parameters are freely estimated or set to an

established value or equality. The more parameters that are fixed and not freely estimated the more constrained or restrictive the CFA model is said to be.

The extent to which structurally invariant factors existed across control and treatment groups was determined by placing these increasingly severe equality constraints on the baseline model established for the control group. Changes in constructs, or in the relationships among constructs, were detected by examining shifts in the variance-covariance matrices across groups. Next, the three parts of the study will be described in greater detail.

Part One of the Secondary Data Analysis—Preparation

Initially, the data set obtained from HRI as a 40-item Microsoft Excel spreadsheet was exported to SPSS (version 7). The categorization of treatment groups and the subsequent creation of random samples were both executed in SPSS prior to exporting the files to PRELIS (version 2.3) for generation of the covariance matrices and mean vectors needed.

Categorization of Treatment Groups

The independent variable, exposure to reform-minded professional development (LSC survey item 16), was used to set apart the four treatment groups examined in this study. The "select cases" and "delete unfiltered" commands in SPSS were used to sequentially isolate and save each of the four groups in separate files. Group 1, identified as the control group (0 hours) for this study had an *n* of 666. Group 2, identified as the lowest level of exposure to professional development (1-19 hours) had an *n* of 813. Group 3, identified as the moderate level of exposure to professional development (20-39 hours) had an *n* of 300 and group 4, identified as the highest level of exposure (40+ hours) had an *n* of 493.

Creation of Random Samples

Next the "compute" command in SPSS was used to create filter variables that were

then used to create the random samples[5] for each of the four groups needed to conduct the study. Typically, in CFA studies such as this, some degree of *post hoc* modification is required to achieve a baseline measurement model with acceptable fit to proceed through the factorial invariance investigation. According to convention, initial CFA was performed on what is referred to as the calibration sample; whereas, when post hoc modifications are made on the calibration sample to improve model fit, cross-validation is performed on what is referred to as the validation sample. As recommended by Jöreskog and Sörbom (1996), *post hoc* modification of structural equation models should be followed by analysis using an independent validation sample. These smaller data sets were created so that each segment of the analysis can be performed on an independent sample.

For this reason for each group, *calibration* samples were created to confirm the measurement model suggested by EFA and *validation* samples were created to allow the researcher to affirm that, should *post hoc* modification be necessary, the new models are as representative of the data as the models prior to modification and not capitalization on chance. The researcher did not know whether *post hoc* modification would be required at the outset of the investigation but had to plan for it accordingly none the less as an option. As such, validation samples were held in reserve.

Group 1 was split into three random samples: (a) a sample to be used to perform EFA in step 2a (n=206), (b) a calibration sample to be used to determine the baseline measurement model for the control group in step 2b (n=227), and (c) a validation sample to be used in steps 2b and 3a-f (n=233). Group 2 was split into two random samples: (a) a calibration sample (n=409) and (b) a validation sample (n=404). Next, Group 3 was split into two random samples: (a) a calibration sample (n=150) and (b) a validation sample (n=150). Group 4 was split into two random samples: (a) a calibration sample (n=249) and (b) a

[5] TRUNC(UNIFORM(n))+1 was the formula used to generate *n* random samples.

validation sample (n=244).

Preparation of Covariance Matrices and Mean Vectors for Use in CFA

The following procedures were used to prepare the matrices and mean vectors needed as input for both Part 2 and Part 3 of the secondary data analysis. Jöreskog and Sörbom (1993) recommended the use of PRELIS 2 to perform the first stage of preparing to conduct CFA. Information about the distribution characteristics and quality of the raw data set to be analyzed using CFA was necessary to prevent the selection of an inappropriate modeling method for the data that would result in abnormal or biased estimation of LISREL parameters. A total of 9 items were excluded from the analysis for lack of full range variance across groups (see Appendix A-Q10d, Q10e, Q10f, Q10g, Q10h, Q10i, 11b, Q11c, Q11k). Thus, 30 teaching practice items were available for inclusion in the baseline part of the study.

Given the ordinal and grouped nature of these data, du Toit (Personal communication, January 16, 2000) recommended that fixed thresholds and asymptotic covariance matrices be calculated and used to prepare the covariance matrices and vectors needed to perform the CFA. Fixed thresholds were computed in PRELIS for the 30 items retained in the data set for use as a common scale for the CFA performed on the four groups (n=2272). The asymptotic covariance matrices computed individually for each group were used to correct for any violation of normal distribution in the samples analyzed as recommended (Jöreskog, 1990, 1994). Calculation of the covariance matrices and parameter estimation was performed using the Weighted Least Squares (WLS) method taking advantage of the information contained in the asymptotic covariance matrices. The WLS method has been found to be quite robust (Chou & Bentler, 1995). PRELIS 2.3 was used to compute the covariance matrices and mean vectors needed as input for LISREL 8.

Part Two of the Secondary Data Analysis—Measurement Baseline Determination

The second part of the study (see Figure 3 above) was performed using two of the

three random samples of the data from the control group respondents (i.e., those respondents reporting 0 hours participation in LSC professional development). The purpose of this second part of the analysis was to identify the best fitting CFA model for these data. The intent was to use the baseline as a standard against which to gauge the strength and direction of any changes in factor structure and/or latent mean levels related to increased exposure to reform-minded professional development. The following sections describe the manner in which the two control group samples were processed. One sample was processed using EFA techniques to determine the number of factors and relationship of items to factors, followed by the second sample which was processed using LISREL to perform CFA in order to specify and confirm the measurement model indicated by factor analysis. These techniques were used to arrive at the baseline measurement model used in the third, factorial invariance, part of the study that examined differences in factor structure across groups.

EFA Used to Determine the Factor Structure for the Control Group

As illustrated in Figure 2, the 1997 LSC K-8 Science Teacher Questionnaire data set was split to isolate only those cases that report having received 0 hours of LSC professional development. The control group cases were further split into three random sub-samples reserved for the three sequential analyses required to pose a baseline measurement structure for the control group with adequate fit to support an investigation of factorial invariance.

A completely exploratory approach was used to determine the number of factors, instrumental, and reference variables. The principal axis factoring command in SPSS was used to determine the appropriate number of oblique factors to extract from the 30 teaching practice items retained. A sample of group 1 data (n = 206) was used to perform the factor analysis. Eight factors were found to have eigenvalues over 1.0 accounting for 66% of the variance. Examination of the scree plot for these data indicated that no more than 3 factors should be extracted. The first three factors extracted accounted for 45% of the variance.

Given that SPSS uses correlation matrices to perform factor analysis, the factor analysis function of LISREL was next used to estimate the baseline three-factor solution. This approach was selected because it was possible to calculate correct standard errors from the covariance matrix and thresholds for Group 1 using Two-Stage Least Squares (TSLS). The advantage of the TSLS solution was that it made it easier to determine simple structure, in that items with statistically significant loadings (t-value ≥ 2.0) were considered to load on that factor. On the basis of the TSLS 3-factor solution the researcher formulated an hypothesis for the baseline CFA model that specified all non-significant loadings as zero.

Bagozzi and Heatherton (1994) reported that measurement models frequently demonstrate unsatisfactory fit when there are more than four or five items per factor and sample sizes are large. As a result four items per factor were retained for inclusion in the baseline measurement model of teaching practice. The twelve items retained (four for each of three factors) were those with the highest loading and the simplest structure. In Table 2 the loadings estimated for each of the twelve retained items using TSLS are presented in bold type, the standard errors for these estimates are presented in parentheses below the loading estimates, and below that the t-values for each estimated loading are provided in italics.

The Measurement Model for Teaching Practice

In LISREL the measurement model "specifies how latent variables or hypothetical constructs depend upon or are indicated by the observed variables. It describes the measurement properties of the observed variables" (Jöreskog & Sörbom, 1993, p. 1). Jöreskog's (1993) suggested protocol for the specification and testing of measurement models was followed. Each of the three teaching practice factors as determined via factor analysis in the previous step, were specified and tested separately, and then in pairs, prior to combining the three factors to create the full measurement model.

Table 2

The 12 Items Retained for the Baseline Model From the TSLS Factor Analysis Solution

LSC Item #	Short Description	Item Stem Text	Reference Variable Factor Loadings Estimated by TSLS: Factor 1	Factor 2	Factor 3	Unique Variance
Q10A	LECTURE	Introduce content through formal presentations.	**0.383** (0.07) *5.239*	**0.176** (0.08) *2.100*	**-0.065** (0.08) *-0.832*	**0.808**
Q10L	HMWRK1	Assign science homework.	**0.573** (0.07) *8.509*	**0.193** (0.08) *2.499*	**0.192** (0.07) *2.681*	**0.517**
Q11E	READTXT	Read from a science textbook in class.	**0.738** (0.06) *12.360*	**0.010** (0.07) *0.149*	**0.020** (0.06) *0.333*	**0.449**
Q11G	WRKSHT	Answer textbook/worksheet questions.	**0.916**	**0.000**	**0.000**	**0.161**
Q10J	PREASMT	Use assessment to find out what students know before or during a unit.	**0.066** (0.08) *0.851*	**0.329** (0.09) *3.717*	**0.106** (0.08) *1.294*	**0.843**
Q11M	DESEXPT	Design or implement their own experiments.	**0.053** (0.06) *0.862*	**0.665** (0.07) *9.334*	**0.216** (0.07) *3.223*	**0.393**
Q11O	MODLSIM	Work on models or simulations.	**0.000**	**0.891**	**0.000**	**0.206**
Q11P	EXTNEXPT	Work on extended science investigations or projects (a week or more in duration).	**-0.006** (0.07) *-0.080*	**0.619** (0.08) *7.588*	**-0.069** (0.08) *-0.908*	**0.643**
Q10M	READRFT	Read and comment on student reflections or journals.	**0.116** (0.06) *2.007*	**0.224** (0.07) *3.155*	**0.447** (0.08) *5.757*	**0.648**
Q11F	READOTR	Read other (non-textbook) science related materials in class.	**0.153** (0.07) *1.965*	**0.064** (0.07) *0.710*	**0.370** (0.07) *3.219*	**0.878**
Q11S	WRTREFL	Write reflections in a notebook or journal.	**0.000**	**0.000**	**0.971**	**0.057**
Q11U	MATHTOOL	Use mathematics as a tool in problem-solving.	**0.071** (0.07) *1.066*	**0.323** (0.08) *4.197*	**0.363** (0.07) *5.024*	**0.664**

Note. **Bold** = estimated loadings, (parentheses) = standard errors, and *italics* = *t*-values.

The hypotheses tested using the second control group sample (n = 227) were as follows: (a) four variables—LECTURE (Q10a), HMWRK (Q10l), READTXT (Q11e), and WRKSHTS (Q11g)—load on the Traditional Practices factor, with WRKSHTS serving as the reference variable; (b) four variables—PREASMT (Q10j), DESSXPT (Q11m), MODLSIM (Q11o), and EXTNEXPT (Q11p)—load on the Investigative Practices factor, with MODLSIM serving as the reference variable; (c) four variables—READRFLT (Q10m), READOTR (Q11f), WRTRFLT (Q11s), and MATHTOOL (Q11u)—load on the Investigative Culture factor, with WRTRFLT serving as the reference variable (see Figure 4 below).

Given that there is indeterminancy between the scale of the factor loadings (the $\hat{\lambda}_{jklm}s$ that describe the strength and direction of the relationship between the latent variable, or factor, and the measured variable) and factor (ξ) variance, the values of the factor loadings depend on the scale of the latent factors. The scale of the latent variable had to be specified to identify the scale for the item parameters or vice versa. The loadings for each of the reference variables derived from the TSLS factor analytic solution were set to 1.0 to establish the scale for the latent variables and identify the baseline model. Error variances were not allowed to covary.

As shown in Table 3 the χ^2 fit index and relative fit indices RMSEA, CFI, and NNFI support accepting the proposed single-and two-factor models as adequate representations of these data. The measurement models (4-item, 1-factor) posed for the Traditional Practice, Investigative Practice, and Investigative Culture factors individually all had non-significant χ^2 (p>0.05), RMSEA of less than or equal to 0.05, as well as NNFI and CFI values of 0.90 or greater. The three two-factor models allowed the two factors in each case tested to covary. The fit and relative fit statistics obtained for the three two-factor combinations (8-item, 2-factor) tested all had non-significant χ^2 (p>0.05), RMSEA of less than or equal to 0.05, as well as NNFI and CFI values of 0.90 or greater.

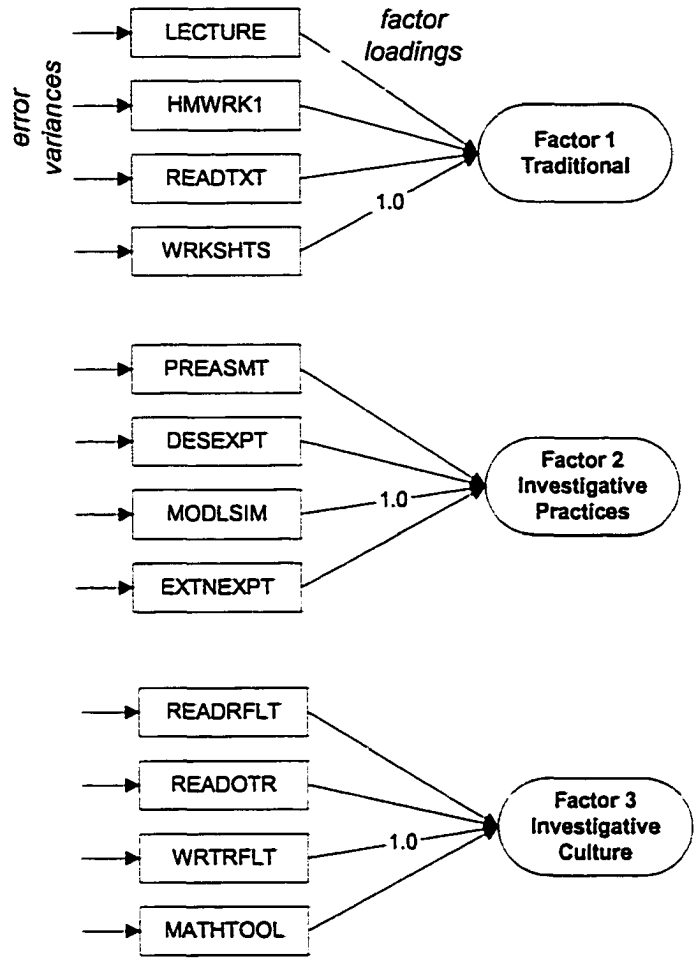


Figure 4. Measurement Models Tested for the Individual Teaching Practice Factors.

The full, 12-item, three-factor, measurement model, which allowed the three factors to covary, was not accepted on the basis of a significant χ^2 ($p<0.05$) and RMSEA greater than 0.05. Examination of the modification indices for this model indicated that *post hoc* modification was required to obtain a baseline measurement model that was an adequate representation of these control group data. The information contained in the modification indices suggested that simple structure for the three-factor model had not been reached.

Table 3

Fit Indices for Single-, Two-, and Three- Factor Baseline Teaching Practice Models Tested

Model Comparison	# Items	χ^2	*df*	$\frac{\chi^2}{df}$	RMSEA	NNFI	CFI
Group 1 Calibration Sample (*n*=227)							
Traditional Practice	4	5.12*	2	2.56	0.052	0.96	0.99
Investigative Practice	4	3.10*	2	1.55	0.000	1.00	1.00
Investigative Culture	4	1.38*	2	0.69	0.000	1.00	1.00
Traditional Practice + Investigative Practice	8	22.88*	19	1.16	0.043	0.98	0.99
Traditional Practice + Investigative Culture	8	27.33*	19	1.43	0.051	0.96	0.98
Investigative Practice + Investigative Culture	8	27.37*	19	1.44	0.044	0.97	0.98
Traditional Practice + Investigative Practice + Investigative Culture	12	92.96	51	1.82	0.060	0.93	0.95
Traditional Practice + Investigative Practice + Investigative Culture (*post hoc* modification)	9	27.86*	24	1.16	0.027	0.99	0.99
Group 1 Validation Sample (*n*=233)							
Traditional Practice + Investigative Practice + Investigative Culture (*post hoc* cross-validation)	9	30.47*	24	1.27	0.034	0.98	0.99

Note. *$p > 0.05$. RMSEA = root-mean-square error of approximation; NNFI = non-normed fit index; CFI = comparative fit index.

Two items were found to load significantly on more than one factor. The PREASMT (Q10j) item was found to load on the Traditional Practice factor in addition to the Investigative

Practice factor. The HMWRK (Q10l) item was found to load on the Investigative Practices factor in addition to the Traditional Practice factor. The READRFLT (Q10m) and WRTRFLT (Q11s) items were found to share more variance than could be accounted for by the Investigative Culture factor. As a result, the PREASMT, HMWRK, and READRFLT items were removed from the three-factor model to create a 9-item model that fit these data very well (non-significant χ^2, $p>0.05$ and RMSEA less than 0.05, as well as NNFI and CFI values of 0.90 or greater).

As recommended by Jöreskog & Sörbom (1993), the *post hoc* modification required that the three-factor 9-item model be validated using another sample of control group data to assure that the new model was not one that capitalized on chance. The third sample of control group data held in reserve (n=233) was used to cross-validate the proposed baseline model. Table 3 shows that the *post hoc* modified baseline model fit these data very well (non-significant χ^2, $p>0.05$ and RMSEA less than 0.05, as well as NNFI and CFI values of 0.90 or greater). This third sample of control group data was then carried forward to be used as the baseline for comparison in the final part of this study. The final part of this study conducted the comparisons of factor structure and mean level across groups using factorial invariance hypothesis testing techniques described previously on the 9-item set (*Traditional Practice*-Q10a, Q11e, Q11g; *Investigative Practice*-Q11m, Q11o, Q11p; *Investigative Culture*-Q11f, Q11s, Q11u).

Part Three of the Secondary Data Analysis—Determination of Invariance

The χ^2 difference test (Bentler & Bonett, 1980) and the nested hypotheses methodology (Widaman & Reise, 1997) were used to determine factorial invariance in this study. Change in fit across treatment groups with increasing exposure to reform-minded professional development was assessed against the baseline model determined for the control group in Part 2. The four groups (control plus three levels of treatment exposure)

were simultaneously compared using the multiple sample feature of LISREL 8 against increasingly restrictive factorial invariance conditions (see the flowchart in Figure 5 below for a detailed description of Part 3 of the secondary data analysis).

Establishing the Group Model for Comparison

The fit of the validation sample of control group data (n=233) to the baseline model was used as the standard against which the fit of the calibration samples for group 2 (n=409), group 3 (n=150), and group 4 (n=249) data were compared. As stated previously, samples of the treatment group data were used because the researcher did not know at the outset of the analysis whether *post hoc* modification and cross-validation in an another sample would be necessary to achieve a group model that would meet the initial condition of configural invariance.

In addition the following constraints were imposed to identify the group model: (a) the reference variable for each factor determined by TSLS was set to 1.0 and constrained to invariance across groups, and (b) the factor means were fixed to zero in the control group and freely estimated in the three treatment groups. These two constraints were sufficient to identify the remaining parameters (factor loading, intercepts, factor variance/covariance, error variance, factor mean level) estimated across groups. The three factors were allowed to covary and the error variances were not. The sequence of steps that was executed by the researcher to perform this study follows below.

The analysis performed was planned to proceed according to a predetermined framework—flowing sequentially from testing the least restricted (only the factor pattern constrained to equality across groups-all others freely estimated) model to the most highly restricted (all parameters pertinent to studies of factorial invariance constrained to equality). The investigation of factorial invariance was planned to proceed until a hypothesis detailing a specific degree of parameter equality across groups failed to be supported by these data.

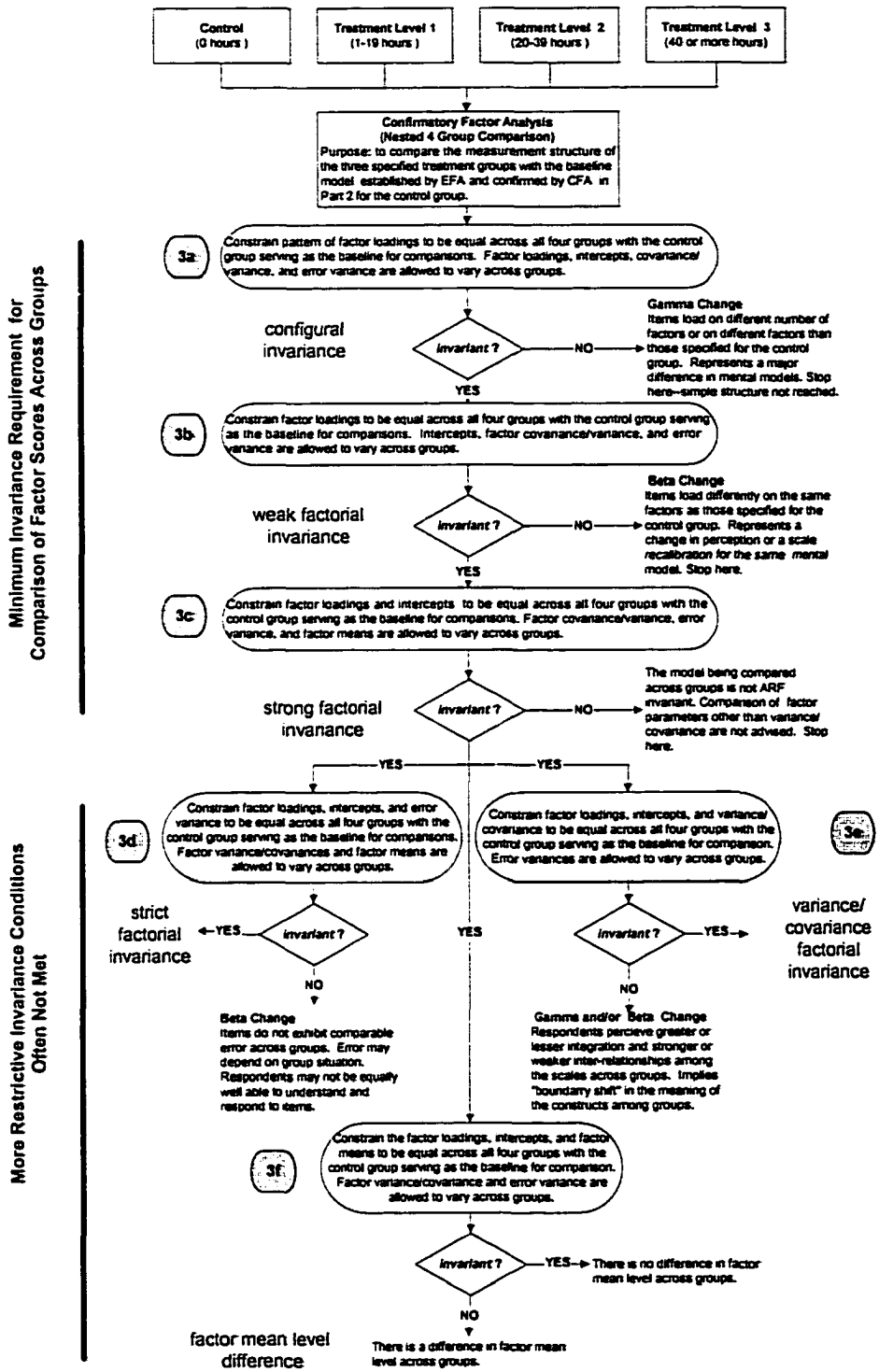


Figure 5. Decision Sequence Used for Factorial Invariance Testing.

At the outset of studies such as this the researcher would not know the extent to which factorial invariance would be demonstrated with their data, hence it was important to have a decision-tree type framework to guide the analytic sequence. In practice, most data sets are found to meet the least constrained forms of factorial invariance (factor pattern), some meet the minimum accepted requirement for comparison of factor means across groups (factor loadings), and very few meet any of the more constrained forms of factorial invariance (LaBouvie & Ruetsch, 1995).

As was the case for the previous part of the analysis, χ^2 and the three other alternative measures of practical fit (RMSEA, CFI, and NNFI) were used to establish the adequacy of model fit across groups. The multiple groups procedure in LISREL 8 was used to simultaneously determine the extent to which the covariance matrices for the control and three treatment groups were statistically equivalent for a given model and its specified parameters, both constrained to equality where appropriate and freely estimated.

Step 3a: Testing for Equality of Factor Pattern (Configural Invariance)

The first CFA task in Part 3 of this study was to determine the extent to which the proposed the factor structure (i.e., measurement model) identified by exploratory factor analysis, confirmed, and then and cross-validated by CFA in Part 2 of this study, fits the self-report teaching practice data across the four levels of treatment. At this initial stage in the analysis, only the factor pattern was constrained to equality across groups. This meant that only the number of factors and salient items were hypothesized to be the same across groups. As described in Chapter 2, testing the equality of factor pattern across groups pertained to determination of the form of factorial invariance referred to as configural invariance.

A significant difference in the variance-covariance matrices at this point in the analysis would have indicated that some alteration in the factor structure (i.e., reorganization

or gamma change) had occurred in one or more of the treatment groups compared to the control group. This test denotes the degree to which the *a priori* common factors determined for the control group or baseline model represent the data for each group. Lack of reasonable fit would have indicated that the dimensionality (either composition of factors or number of factors) of the factor model proposed differed across groups (Taris et al., 1998).

Should the hypothesis of configural invariance across groups be rejected in step 2a, the analysis was planned to proceed to identify the extent and nature of the structural differences between control and treatment groups and thus, describe the evidence of gamma change exhibited as a function of treatment exposure. When reasonable indices of fit are obtained for step 3a, the hypothesis that the conditions of configural invariance are met across groups should be accepted and the analysis should then proceed to examine the next in a series of increasingly restricted nested invariance models.

Step 3b: Testing for Equality of Factor Loadings (Weak Factorial Invariance)

Should acceptable fit values be obtained for the model specified in step 3a, this model should then be modified by adding the constraint that the factor loading matrices (Λ) be invariant across groups (see Figure 5). The χ^2 that results from this step 3b model should be compared with the value obtained for the configural invariance model tested step 3a (Bentler & Bonnett, 1980). The difference in χ^2 values between two nested models is distributed as a χ^2 with degrees of freedom equal to the difference in degrees of freedom between the two models. Should the restricted model result in a non-significant increase in χ^2 (or $\Delta\chi^2$) over the less constrained model then, the hypothesis of what was referred to as weak factorial invariance across groups should be accepted. If the hypothesis of weak factorial invariance was accepted, analysis was planned to proceed to place additional invariance constraints on the model. If the conditions of weak factorial invariance were not met, analysis was planned to proceed to determine the nature and extent of the structural differences and thus, evidence

of beta change (a recalibration or change in scaling units) exhibited across the control and treatment groups (Taris et al., 1998).

Step 3c: Testing for Equality of Intercepts (Strong Factorial Invariance)

Should acceptable fit values be obtained for the model specified in step 3b, this model will be modified by adding the constraint that the intercept matrices ($\hat{\tau}$) be invariant across groups (see Figure 5). This constraint examines the extent to which the intercept matrices among the factors being studied are invariant. Similar to step 3b above, should the more highly constrained step 3c model result in a non-significant increase in χ^2 (or $\Delta\chi^2$) over the less constrained model then, the hypothesis of equal intercepts across groups should be accepted. In addition, as described in Chapter 2, strong factorial invariance allows for ARF invariant interpretation of estimated parameters.

Step 3d: Testing for Equality of Error Variance (Strict Factorial Invariance)

Should acceptable fit values be obtained for the model specified in step 3c, this model would then be modified by adding the constraint, to all those that have been previously described, that the diagonal elements of the error matrices ($\hat{\Theta}_{\delta g}$) also be invariant across groups. Should the previous hypothesis of equal variance have been accepted, this test of equality across error variances indicates whether the measurement error is invariant across groups. Strict factorial invariance, under most conditions rarely occurs. Similar to steps 2b through 2c above, should the more highly constrained step 2d model result in a non-significant increase in χ^2 (or $\Delta\chi^2$) over the less constrained model then, the hypothesis of equal error variance across groups should be accepted. Should the constrained model be accepted, this implies equivalence across respondents in their ability to understand and provide answers to the items, regardless of group membership; however, should the hypothesis be rejected, the error variance of the items may depend on the situation (i.e. some

form of beta change) (Taris et al., 1998).

Step 3e: Testing for Equality of Factor Covariance/Variance

Should acceptable fit values be obtained for the strong factorial invariance model specified in step 3c, this model will be modified by adding the constraint that the factor covariance matrices ($\hat{\Phi}_g$) be invariant across groups (see Figure 5). This constraint examines the extent to which the covariances and variances among the factors being studied are invariant. Similar to the preceding steps 3a-d above, should the more highly constrained step 3e model result in a non-significant increase in χ^2 (or $\Delta\chi^2$) over the less constrained model then, the hypothesis of equal covariance/varaince across groups should be accepted.

If the factor covariance elements fail to exhibit invariance across groups Schmitt (1982) asserts that this is evidence of gamma change, in that the strength and/or direction of the relationship among the factors has shifted in some way. Taris et al. (1998) have indicated that should this be the case, it implies that a "shift in the boundary of meaning" among the constructs has occurred.

Alternatively, or in addition, should the factor variance elements fail to exhibit invariance across groups, Schmitt (1982) asserted that this may be evidence of a form of beta change (recalibration of the true score continua). Similarly, Taris (1998) suggested that changes in factor variances may indicate that respondents perceive more or less of a difference in the relevant constructs across groups. Thus, rejection of the equal variances hypothesis could signal that certain groups are better able to differentiate among constructs.

Step 3f: Testing for Equality of Factor Mean Level

Should acceptable fit values be obtained for the strong factorial invariance model specified in step 3c, this model would be modified by adding the constraint that the factor mean matrices ($\hat{\kappa}$) be invariant across groups (see Figure 5). This constraint examines the

extent to which the mean level among the factor scores are equivalent. Similar to the preceding steps 3a-e above, should the more highly constrained step 3f model result in a non-significant increase in χ^2 (or $\Delta\chi^2$) over the less constrained model then, the hypothesis of equal factor mean level across groups should be accepted. Thus, when the factor mean level fails to exhibit invariance across groups then the data support group differences in factor mean level. The results of factorial invariance testing (steps 3a-f) as described follow in Chapter 4.

CHAPTER 4

RESULTS

Testing Factorial Invariance Across Groups

The core function of a factorial invariance study is to determine whether the factor structures under investigation demonstrate sufficient similarity across groups, conditions, or time to support valid comparison of factor mean scores. The results that follow present the extent to which the nested set of factorial invariance hypotheses described in the previous methods chapter were upheld.

Establishing the Baseline Model for Comparison

The first step in investigating factorial invariance within a multiple-group CFA model was to specify a baseline model that fits the data satisfactorily. The baseline model was one that meets the minimal conditions of configural invariance. Configural invariance requires that the factor pattern matrices are equivalent across the groups under comparison. This meant that for each group, the measured variables (items) relate to the latent variables (factors) in the same general way. Specifically, the pattern of zero (an item does not load on a given factor) and non-zero (an item does load on a given factor) loading should be the same across groups (see Table 4).

In this study, these minimal conditions were tested when all model matrices were freely estimated for each of the four treatment groups, with the exclusion of those constraints imposed to identify the model across groups. This baseline model then served as the starting point against which the fit of more restricted forms of invariance were compared.

Table 4

Hypothesized Pattern Matrix for Configural Invariance

Measured Variable	Factor 1 Traditional Practice	Factor 2 Investigative Practice	Factor 3 Investigative Culture
LECTURE (Q10a) Lecture	λ_{11}	0	0
READTXT (Q11e) Textbook	λ_{21}	0	0
WRKSHT (Q11g) Worksheets	λ_{31}	0	0
DESEXPT (Q11m) Experiments	0	λ_{42}	0
MODLSIM (Q11o) Simulations	0	λ_{52}	0
EXTNTXPT (Q11p) Projects	0	λ_{62}	0
READOTH (Q11f) References	0	0	λ_{73}
WRTREFL (Q11s) Reflection	0	0	λ_{83}
MATHTOOL (Q11u) Problem-solving	0	0	λ_{93}

As shown in Table 5 the baseline model (Model 1) had an acceptable level of fit. Specifically, Model 1 had a χ^2/df ratio of less than 2, a RMSEA of less than 0.05, and both NNFI and CFI greater than 0.90. Thus, although the chi-square fit statistic was statistically significant at $p < 0.001$, the baseline model can be said to fit the data reasonably well considering the large sample size for all four groups relative to the small number of variables included in the model. These data provide evidence that support acceptance of the hypothesis of configural invariance. Under these conditions the researcher can assert that similar, but not identical, latent variables (factors) are present across the four groups.

Table 5

Fit Indices for Alternative Structural Models

Model	Absolute Fit Indices			Relative Fit Indices		
	χ^2	df	$\frac{\chi^2}{df}$	RMSEA	NNFI	CFI
Configural Invariance Model 1 Baseline	151.36	96	1.58	0.047	0.96	0.97
Weak Factorial Invariance Model 2 Model 1 + $\hat{\Lambda}_g$ invariant	176.91	114	1.55	0.046	0.96	0.97
Strong Factorial Invariance Model 3 Model 2 + $\hat{\tau}_g$ invariant	185.47	132	1.41	0.040	0.97	0.97
Strict Factorial Invariance Model 4 Model 3 + $\hat{\Theta}_g$ invariant	271.64	159	1.71	0.052	0.95	0.95
Variance/Covariance Invariance Model 5 Model 3 + $\hat{\Phi}_g$ invariant	263.55	150	1.76	0.054	0.95	0.95
Factor Mean Invariance Model 6 Model 3 + $\hat{\kappa}_g$ invariant	945.13	141	6.70	0.148	0.61	0.62

Note: All chi-square values were significant at the $p < 0.001$ level. RMSEA = root-mean-square error of approximation; NNFI = non-normed fit index; CFI = comparative fit index.

Using the Chi-square Difference Test with Nested Invariance Hypotheses

The absolute (χ^2 and $\chi^2 : df$) and relative fit statistics (RMSEA, NNFI, and CFI)

are reported in Table 5 for the set of five nested models tested. The chi-square values for each of the five models were significant at the $p < 0.001$ level. The influence of large sample size on the chi-square statistic used to assess absolute model fit made it necessary to rely on what Widaman & Reise (1997) refer to as comparative measures of "practical fit" rather than the statistical significance of the chi-square alone for decisions as to accept or reject each of the increasingly restricted nested models and their related invariance hypotheses. The Bentler & Bonett (1980) method of chi-square difference was used as an alternative to test the remaining invariance hypotheses for the set of nested models. This method allowed the researcher to establish the extent to which invariance exists using the strength of a statistical test to accept or reject the set of nested invariance hypotheses.

When restrictions are placed on one model to create another more constrained model, such as holding a parameter invariant, the more constrained model was said to be nested within the less constrained model. According to Bentler & Bonett (1980) under these conditions, the difference in chi-square values ($\Delta\chi^2$) for the nested model pair is distributed as a chi-square variate with degrees of freedom equal to the difference in degrees of freedom between the two models (Δdf). The $\Delta\chi^2$ value is then used to test the statistical significance of the difference in fit between the nested models. When the $\Delta\chi^2$ value is statistically significant ($p< 0.05$), the less constrained model provides a significantly better fit to the data. Thus, for each of the increasingly constrained models, the extent to which factorial invariance holds was determined by testing whether constraining a given matrix (such as $\hat{\Lambda}_g$, $\hat{\tau}_g$, $\hat{\Theta}_g$, $\hat{\Phi}_g$, or $\hat{\kappa}_g$) to invariance across the four treatment groups results in significant deterioration in model fit compared to that for the less constrained model from which it was constructed. In those instances where the nested invariance restriction did not result in degradation of model fit ($p > 0.05$), that invariance hypothesis was accepted and that level of invariance, as defined by the parameter constraint tested, was said to hold. In those instances where the nested invariance restriction did result in significant degradation of model

fit ($p < 0.05$), that invariance hypothesis was rejected and that level of invariance, as defined by the parameter constraint tested, was said to fail to hold.

Testing for Weak Factorial Invariance

Upon accepting the baseline hypothesis of configural invariance, the hypothesis of weak factorial invariance, $\hat{\Lambda}_1 = \hat{\Lambda}_2 = \hat{\Lambda}_3 = \hat{\Lambda}_4$, was then evaluated. Under the conditions of weak factorial invariance, the loading, or regression coefficient, for each of the measured variables on their respective latent variable was hypothesized to be equivalent across groups. To test this hypothesis, the baseline Model 1 was modified to create Model 2 by imposing the constraint that the $\hat{\Lambda}_g$ matrix, or factor loading, be held invariant across groups. As shown in Table 5 the chi-square value of 176.91 with 114 degrees of freedom for Model 2 was statistically significant ($p < 0.001$) which under the stringent conditions of absolute measures of fit would lead to rejection of this large model. However as seen in Table 6, using Bentler & Bonett's (1980) $\Delta\chi^2$ nested model method to compare relative fit between Model 1 and Model 2, the additional constraint of holding the $\hat{\Lambda}_g$ matrix to invariance across groups did not yield a statistically significant decrease in model fit.

As shown in Table 6, moving from configural (Model 1) to weak factorial invariance (Model 2) there was no evidence of fit deterioration—the $\Delta\chi^2$ value for this comparison was 25.55 with a Δdf of 18 ($p > 0.05$). In addition, RMSEA exhibited a small degree of improvement and NNFI and CFI remained unchanged. Consequently, Model 2 represents a relatively better fitting alternative structural model than Model 1. Thus, the hypothesis of weak factorial invariance across the four treatment groups was accepted.

Weak factorial invariance is as a prerequisite to the comparison and interpretation of differences across groups with respect to factor variance/covariance that are not subject to the indeterminancy that arises with alternative scaling strategies for latent variables. These data provided evidence to support the assertion that factor variances and covariances are

ARF (appropriate rescaling factors) invariant across the four treatment groups. As in this case when the ARF condition is met, factor variances and covariances across groups may be meaningfully compared regardless of how the model was identified/scaled.

Table 6

Differences in Fit of Alternative Structural Models

				Difference		
Model Comparison	$\Delta\chi^2$	Δdf	$\frac{\Delta\chi^2}{\Delta df}$	RMSEA	NNFI	CFI
Weak Factorial Invariance Model 1 vs. Model 2: Testing Invariance of $\hat{\Lambda}_g$	25.55*	18	1.42	-0.001	0.000	0.000
Strong Factorial Invariance Model 2 vs. Model 3: Testing Invariance of $\hat{\tau}_g$	8.56*	18	0.48	-0.006	-0.01	0.000
Strict Factorial Invariance Model 3 vs. Model 4: Testing Invariance of $\hat{\Theta}_g$	86.17**	27	3.19	+0.012	+0.02	+0.02
Variance/Covariance Invariance Model 3 vs. Model 5: Testing Invariance of $\hat{\Phi}_g$	78.08**	18	4.34	+0.014	+0.02	+0.02
Factor Mean Invariance Model 3 vs. Model 6: Testing Invariance of $\hat{\kappa}_g$	759.66**	9	84.4	+0.108	+0.36	+0.35

Note. For all model comparisons, the second-listed model was more restricted than, and was nested within, the first-listed model. Given the way in which indices of practical fit were computed and interpreted, negative difference values mean better fit for more restricted model, positive difference values mean worse fit for the more restricted model. RMSEA = root-mean-square error of approximation; NNFI = non-normed fit index; CFI = comparative fit index.

*$p > 0.05$. ** $p < 0.001$.

Testing for Strong Factorial Invariance

Next, the hypothesis of strong factorial invariance, $\hat{\tau}_1 = \hat{\tau}_2 = \hat{\tau}_3 = \hat{\tau}_4$, was evaluated. To test this hypothesis, Model 2 was modified to create Model 3 by imposing an additional constraint that the $\hat{\tau}_g$ matrix, or item intercepts, be held invariant across groups. As shown in Table 5 the chi-square value of 185.47 with 132 degrees of freedom for Model 3 was statistically significant ($p < 0.001$) which under the stringent conditions of absolute measures of fit would lead to rejection of this large model. However as seen in Table 6, using the Bentler & Bonett (1980) nested model method to compare relative fit between Model 2 and Model 3, the additional constraint of the $\hat{\tau}_g$ matrix to invariance across groups did not yield a statistically significant decrease in model fit.

As shown in Table 6, moving from weak (Model 2) to strong (Model 3) factorial invariance there was no evidence of fit deterioration—the $\Delta\chi^2$ value for this comparison was 8.56 with a Δdf of 18 ($p > 0.05$). Constraining the $\hat{\tau}_g$ matrix to invariance did not result in a statistically significant decrease in model fit from that obtained for constraining the $\hat{\Lambda}_g$ matrix alone. There was much better fit per degree of freedom difference for Model 3 (0.48) compared to Model 2 (1.42). In addition, the RMSEA and NNFI exhibited some degree of improvement and the CFI remained unchanged. These measures indicated that the strong factorial invariance model fit these data better than the less restricted, weak factorial invariance model against which fit was compared. Thus, the hypothesis of strong factorial invariance across the four treatment groups was accepted.

Strong factorial invariance serves as a prerequisite to the meaningful comparison and interpretation of differences across groups with respect to the relative level of factor mean scores and factor variance/covariance. These data provided evidence that supports the assertion that factor mean scores and factor variances are ARF invariant across the four treatment groups. Group differences in both the means and variances on the latent variables,

representing the teaching practice constructs theorized by the systemic reform literature, are captured in group differences in the means and variances on the measured variables.

Testing for Strict Factorial Invariance

Model 4 imposed the constraint that the $\hat{\Theta}_g$ matrices be held invariant across the four treatment groups in addition to the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices specified in Model 3. Strict factorial invariance specifies that the measurement residuals (error) be invariant across groups. As shown in Table 5 the chi-square value of 271.64 with 159 degrees of freedom for Model 4 was statistically significant ($p < 0.001$) which under the stringent conditions of absolute measures of fit would lead to rejection of this large model. The Bentler & Bonett (1980) nested model method was again used to compare relative fit between Model 4 and Model 3 to determine the effect of constraining the $\hat{\Theta}_g$ matrix to invariance across groups ($\hat{\Theta}_1 = \hat{\Theta}_2 = \hat{\Theta}_3 = \hat{\Theta}_4$) on model fit.

As shown in Table 6, moving from strong (Model 3) to strict (Model 4) factorial invariance there was evidence of significant fit deterioration—the $\Delta\chi^2$ value for this comparison was 86.17 with a Δdf of 27 ($p < 0.001$). Constraining the $\hat{\Theta}_g$ matrix to invariance did result in a statistically significant decrease in model fit from that obtained for constraining the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices alone. There was a large decrease in the fit per degree of freedom difference for the more constrained Model 4 (3.19) compared to the less constrained Model 3 (0.48). In addition, the RMSEA, NNFI, and CFI all exhibited some degree of deterioration. These measures indicated that the strict factorial invariance model does not fit these data better than the less restricted, strong factorial invariance model against which fit was compared. Thus, the hypothesis of strict factorial invariance across the four treatment groups was rejected.

Strict factorial invariance serves as a prerequisite to the comparison and interpretation of differences across groups with respect to the error variance of the measures.

Under those conditions where strict factorial invariance holds group differences in the mean scores and variances on the measured variables (items) are a function only of group differences in mean scores and variances on the latent variables (factors). When strict factorial invariance does not hold, group differences on the measured variables are not entirely attributable to group differences on the latent variables. In addition, these data provided evidence to support the assertion that there are differences in the reliability of measures across groups (the nature and extent of these differences will be presented in the following section on parameter estimates).

Additional Invariance Tests Across Groups

In that the condition of strong factorial invariance held with these data, it was possible to proceed to perform additional invariance tests across groups to determine the equivalence of factor variance/covariance and mean level matrices and examine similarities and differences across the complete set of parameter estimates ($\lambda, \tau, \theta, \Phi$, and κ).

Testing Invariance of the Factor Variance/Covariance

In addition to the constraints specified by Model 3, holding the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices invariant, Model 5 specifies that the variance/covariance matrices ($\hat{\Phi}_g$) be invariant across treatment groups ($\hat{\Phi}_1 = \hat{\Phi}_2 = \hat{\Phi}_3 = \hat{\Phi}_4$). As can be seen in Table 5, the comparison between Model 3 and Model 5 resulted in a $\Delta\chi^2$ value of 78.08 with a Δdf of 18 ($p < 0.001$). Constraining the $\hat{\Phi}_g$ matrix to invariance did result in a statistically significant decrease in model fit from that obtained for constraining the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices alone.

There was a moderate decrease in the fit per degree of freedom difference for the more constrained Model 5 (4.34) compared to the less constrained Model 3 (3.19). In addition, the RMSEA, NNFI, and CFI all exhibited some degree of deterioration. These measures indicated that the variance/covariance factorial invariance model does not fit these

data better than the less restricted, strong factorial invariance model against which fit was compared. Thus, the hypothesis of variance/covariance factorial invariance across the four treatment groups was rejected. In addition, these measures provided evidence to support the assertion that there are differences in the factor variances and covariances across groups (the nature and extent of these differences will be presented in the following section on parameter estimates).

Testing the Invariance of the Factor Means

In addition to the constraints specified by Model 3, holding the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices invariant, Model 6 specifies that the factor mean matrices ($\hat{\kappa}_g$) be invariant across treatment groups ($\hat{\kappa}_1 = \hat{\kappa}_2 = \hat{\kappa}_3 = \hat{\kappa}_4$). As can be seen in Table 5, the comparison between the less constrained Model 3 and the more constrained Model 6 resulted in a $\Delta\chi^2$ value of 759.66 with a Δdf of 9 ($p < 0.001$). Constraining the $\hat{\kappa}_g$ matrix to invariance did result in a statistically significant decrease in model fit from that obtained for constraining the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices alone. There was an extremely large decrease in the fit per degree of freedom difference for the more constrained Model 6 (84. 4) compared to the less constrained Model 3 (3.19). In addition, the RMSEA, NNFI, and CFI all exhibited a very large amount of deterioration. These measures indicated that the factor mean factorial invariance model does not fit these data better than the less restricted, strong factorial invariance model against which fit was compared. Thus, the hypothesis of factor mean level factorial invariance across the four treatment groups was rejected. In addition, these measures provided evidence to support the assertion that there are differences in the mean level of factor scores across groups (the nature and extent of these differences will be presented in the following section on parameter estimates).

Comparison of Parameter Estimates

Taking into account the differences in relative fit between the nested models tested it was found that Model 3, the strong factorial invariance model ($\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices held invariant across treatment groups), best represented these data among those alternative measurement models tested. Invariance of the $\hat{\Lambda}_g$ and $\hat{\tau}_g$ matrices established that the same latent variables, or factors, are identified in each of the four treatment groups evaluated. Alternatively, elements in the $\hat{\Theta}_g$, $\hat{\Phi}_g$, and $\hat{\kappa}_g$ matrices were found not to be invariant across the groups. Parameter estimates from Model 3, the strong factorial invariance model, were then examined to detect the extent to which similarities and differences were evident across groups.

Measured Variable Intercepts and Factor Loadings

Common metric completely standardized estimates of the elements in the $\hat{\tau}_g$ matrix varied from –0.01 to –0.32 (see Table 7). The intercept estimates for the Traditional Practice factor ranged from –0.01 to –0.03, whereas those for the two reform-minded factors Investigative Practice and Investigative Culture ranged from –0.27 to –0.32. Given that the $\hat{\tau}_g$ matrix was found to be invariant, these values are equivalent across the four treatment groups.

The factor loadings for the $\hat{\Lambda}_g$ matrix are provided in Table 7. Using the criteria established by Comrey & Lee (1992), 89% of the measured variables in this study were shown to be excellent to good markers for their respective factors. Specifically, 56% were shown to be excellent (textbook, worksheets, experiments, simulations, and projects) with loadings larger than 0.71 (communality estimates of 50% or greater). 33% were found to be very good to good (references, reflection, and problem-solving) with loadings between 0.63 and 0.55 (communality estimates of 40-30%). Although the *t*-value for the lecture variable

Table 7

WLS Estimates of Intercept, Factor Loading, Error Variance, and Squared Multiple Correlation (SMC) for Model 3

		Loading ($\hat{\lambda}$) *(Communality Estimate)*			Error Variance ($\hat{\Theta}_\varepsilon$) *(Squared Multiple Correlation)*			
Measured Variable	Intercept ($\hat{\tau}$)	Trad. Practice	Invest. Practice	Invest. Culture	GR1	GR2	GR3	GR4
Lecture	-0.01	0.43 *(0.19)*	0	0	0.88 *(0.18)*	0.66 *(0.24)*	0.87 *(0.18)*	0.97 *(0.13)*
Textbook	-0.03	0.80 *(0.64)*	0	0	0.55 *(0.54)*	0.43 *(0.62)*	0.30 *(0.68)*	0.12 *(0.81)*
Worksheets	-0.01	0.89 *(0.79)*	0	0	0.23 *(0.78)*	0.13 *(0.87)*	0.21 *(0.80)*	0.35 *(0.64)*
Experiments	-0.31	0	0.78 *(0.60)*	0	0.51 *(0.59)*	0.29 *(0.69)*	0.44 *(0.59)*	0.41 *(0.52)*
Simulations	-0.32	0	0.86 *(0.74)*	0	0.27 *(0.77)*	0.27 *(0.74)*	0.33 *(0.69)*	0.22 *(0.70)*
Projects	-0.27	0	0.73 *(0.53)*	0	0.39 *(0.62)*	0.57 *(0.49)*	0.34 *(0.61)*	0.48 *(0.44)*
References	-0.30	0	0	0.57 *(0.32)*	0.68 *(0.38)*	0.69 *(0.34)*	0.66 *(0.27)*	0.66 *(0.28)*
Reflection	-0.28	0	0	0.66 *(0.44)*	0.41 *(0.57)*	0.47 *(0.50)*	0.79 *(029)*	0.71 *(0.32)*
Problem-solving	-0.30	0	0	0.63 *(0.40)*	0.74 *(0.41)*	0.56 *(0.43)*	0.56 *(0.34)*	0.57 *(0.35)*

was statistically significant, its communality estimate was low (0.19) making it a relatively poor marker for its respective factor.

Error Variances, Latent Variable Variance and Covariances

As shown in Table 7, neither the error variance nor the squared multiple correlation[6] (SMC) of measured variables were constant across treatment groups. The SMC in CFA studies serves to assess both the reliability and the proportion of variance of a measured variable accounted for by a given factor. SMC is similar to the communality estimate in traditional factor analytic approaches. For the most part, the reciprocal relationship predicted by classical test theory—as error variance decreases, reliability increases and the converse—was observed. In one case however, the problem-solving measured variable loading on the Investigative Culture factor, reliability decreased as error variance decreased. In the majority of instances (89%) the reliability for measured variables decreased as professional development exposure increased. The one exception was with the textbook item where reliability increased in association with increased exposure to professional development.

The WLS estimates completely standardized to a common metric for factor variances and covariances across the four treatment groups are shown in Table 8. These data indicated that as exposure to professional development increased the variance and covariance of each of the three hypothesized teaching practice factors decreased. The variance decrease associated with increased exposure to reform-minded professional development observed for the Traditional Practice factor was approximately 0.2 standard deviations--a relatively small effect. The variance the two reform-minded factors, Investigative Practice and Culture, was observed to decrease 0.5 standard deviations--a relatively large effect associated with increased exposure to reform-minded professional development.

---

[6] $SMC_{\text{var}i} = \lambda_i^2 / \lambda_i^2 + \Theta_i$

Table 8

WLS Estimates of Variance/Covariance for Latent Variables From Model 3

Factor	Factor		
	Trad. Practice	Invest. Practice	Invest. Culture
GR1 (0 hours)			
Traditional Practice	**1.03**		
Investigative Practice	0.42	**1.21**	
Investigative Culture	0.55	1.10	**1.27**
GR2 (1-19 HOURS)			
Traditional Practice	**1.10**		
Investigative Practice	0.51	**1.04**	
Investigative Culture	0.67	0.87	**1.08**
GR3 (20-39 HOURS)			
Traditional Practice	**1.02**		
Investigative Practice	0.37	**1.02**	
Investigative Culture	0.45	0.73	**0.73**
GR4 (40+ HOURS)			
Traditional Practice	**0.80**		
Investigative Practice	0.03*	**0.72**	
Investigative Culture	0.20*	0.66	**0.77**

Note. **Bold** diagonal values represent factor variance. Plain text off-diagonal values represent factor covariance. * $p<0.05$, covariance not statistically significant. Results reported are completely standardized to a common metric to facilitate comparison across groups.

The decrease in covariance between the Traditional Practice and the Investigative Practice and Culture factors associated with increased exposure to professional development observed was 0.4 standard deviations (0 hours compared to 40+ hours)— a relatively large effect. The relationship between the Traditional factor and both reform-minded factors supported by the baseline model for the unexposed control group was not found to be

statistically significant at the highest level of exposure to professional development. Similarly, the decrease in covariance between the Investigative Practice and Culture factors associated with increased exposure to reform-minded professional development was observed to be 0.4 standard deviations (0 hours compared to 40+ hours)—a relatively large effect.

Factor Means

The factor, or latent variable mean scores, for the four treatment groups are compared in Table 9. The factor mean scores were fixed at zero for the control group (0 hours professional development) to permit the estimation of this parameter in the three remaining treatment groups. These estimates showed relatively little change, less than 0.1 standard deviation, in the mean score for the Traditional Practice factor associated with increased exposure to reform-minded professional development. Alternatively, these estimates indicated a moderate increase, 0.3 standard deviation, in the mean score for the Investigative Culture factor and a relatively large increase, 0.5 standard deviation, in the mean score for the Investigative Practice factor associated with increased exposure to reform-minded professional development. In addition, these estimates revealed that some degree of increase in the two reformed practice factors was evident with as few as 1-19 hours of exposure and that change was more gradual for the Investigative Culture factor.

In summary, the strong factorial invariance held. Factor structures were found to be invariant across treatment groups to a degree sufficient to allow comparison of factor variances, covariances, and means. Variance and covariance for all three teaching practice factors were found to decrease in association with increasing exposure to reform-minded professional development. The factor mean for the Traditional Practice factor was found to be relatively stable across treatment exposure groups; whereas, the Investigative Practices and Culture factor means were found to increase in association with as little as 1-19 hours of exposure to reform-minded professional development.

Table 9

WLS Estimates of Means for Latent Variables From Model 3

Treatment Group	Factor		
	Trad. Practice	Invest. Practice	Invest. Culture
GR1 (0 hours)	0.00	0.00	0.00
GR2 (1-19 hours)	0.09	0.22	0.17
GR3 (20-39 hours)	0.04	0.52	0.23
GR4 (40+ hours)	0.08	0.51	0.33

Note. Results reported are completely standardized to a common metric to facilitate comparison across groups.

CHAPTER 5

DISCUSSION AND CONCLUSION

Findings

The findings from this study point to two key aspects of change of which evaluators faced with the challenges presented by the assessment of change should now be aware: (a) conceptual change which results in alterations in the measurement structure for a given construct, and (b) mean level change which results in changes in the magnitude of a factor score for a given construct. The strategies deployed for the detection and measurement of change in an evaluation setting should, in light of this study, be appropriate for the type of change assessed. This chapter provides discussion of the findings with respect to these areas of critical interest to the field of evaluation, presents the implications of the findings within the context of systemic educational reform, points to limitations of the study, and offers suggestions for furthering this line of inquiry.

The Detection and Disentanglement of Conceptual and Mean Level Change

This study was designed to enable the researcher to explore the extent to which CFA could be used to address some of the measurement challenges faced by evaluators engaged in the interpretation of self-report survey data collected under quasi-experimental conditions. In particular, this study presents the case for strong factorial invariance as the central prerequisite to the valid and reliable use of linear composite scores to gauge the relative influence of treatment across groups. Quite simply, regardless of the conditions under which the evaluation was conducted or the groups surveyed, the evaluator needs evidence that not only speaks to the magnitude of treatment effects but also to the substantive coherence and

measurement equivalence of pertinent constructs.

Factor structures, or measurement models as they are referred to in CFA, were used in this study as a way to examine the substantive and measurement equivalence of a multi-dimensional construct across control and treatment groups. The methods research presented here demonstrates the extent to which factorial invariance holds across control and treatment conditions. In addition, these methods enable the researcher to not only to compare treatment effect size across groups from a more robust measurement position—but also to make limited inferences about the relationship between treatment exposure and the emergence of differences (i.e., conceptual change) in the mental models held by respondents for the construct(s) of interest. Factorial invariance provides a useful methodological lens to detect and distinguish between tangled evidence of conceptual and mean level change that occur across comparison groups.

Given that these data were found to support the condition of strong factorial invariance across groups, it was reasonable to assert that exposure to reform-minded professional development in and of itself was not associated with wholesale alterations in the mental model or measurement structure for the multi-dimensional teaching practice construct. Sufficient evidence of factorial invariance was found to support the comparison of factor means and other parameters across groups. However, these data do provide evidence of substantive conceptual change in the relationships among factors and of moderate to large effect sizes for changes in the mean level of investigative aspects of science instruction (practice and culture) associated with exposure to treatment. In addition interpretation of these results presents clear implications and suggestions to inform evaluation practice where the assessment of change is concerned. The reporting of factor scores without addressing issues of factorial invariance in effect can bury significant information about the undercurrents of change and limit the valid use of change scores.

Evidence of Conceptual Change

First, acceptance of the configural invariance hypothesis established that an equivalent factor pattern, in this case simple structure, describes teaching practice across groups irrespective of the amount of treatment exposure. There was no evidence to support that conceptual change of the most serious gamma type had occurred with respect to the relationship among items and their factors. Thus, the meaning or the way in which respondents "see" teaching practice items relate to teaching practice factors was found to be equivalent across groups. Given the complex nature of the actual factor structure representative for the entire data set rather than the abbreviated slice examined here, these findings affirm the importance of simple structure. This does present somewhat of a paradox however. In particular, for many developmental processes such as those studied here, the dimensionality of constructs is known to increase over time (Maruyama, 1998). Forcing items to achieve simple structure, although the model fits, precludes the possibility that items slide on and off factors as the mental models for developmental constructs, such as teaching practice, evolve. The use of change scores without evidence of configural invariance seriously compromises ability of the evaluator to establish that the constructs being assessed are the same across groups.

Second, acceptance of the weak and strong factorial invariance hypotheses established that factor loadings and item intercepts are equivalent across groups. There was no evidence to support that conceptual change of the beta type had occurred among the regression weights or intercepts of items with respect to their specific factors. Thus the scale or the way in which respondents "weight" teaching practice items relative to each other and to each specific factor was found to be equivalent across groups. Respondents perceive and calibrate the three factor scales that comprise teaching practice in the same way across groups irrespective of the amount of treatment exposure. The relative emphasis given each

item is a valuable part of the information contained in a factor score in that it is a weighted linear composite and not just the simple average of summed item scores. The reporting of factor scores without evidence of strict factorial invariance compromises the interpretation of differences and similarities across groups.

Third, rejection of the strict factorial invariance hypothesis established that the error variance for items was not equivalent across groups. There was evidence to suggest that respondents may not be equally well able to understand and respond to the items. In addition this evidence of the beta type of conceptual change suggested that item reliability, as measured by squared multiple correlation, may depend to some extent on group situation. Although tests of weak and strong factorial invariance were accepted, this really only meant that the differences in item loading and intercepts present across groups were not sufficiently large, given the error, to detect statistically significant differences—not that there was no difference. Even though strict factorial invariance usually is neither a precondition for the comparison of factor scores nor expected, it would have profound implications for studies where large numbers of items with complex loading are involved or where bias was a particular concern.

Fourth, rejection of the variance/covariance factorial invariance hypothesis established that factor variance and the relationships among factors were not equivalent across groups. There was evidence to suggest that there was an association between exposure to treatment and the extent to which respondents saw greater coherence within a given factor but also greater distinction among factors. This evidence of conceptual change in terms of boundary shift of scale score range within (variance-beta type) and distinctiveness among factors (covariance-gamma type) presents the most intriguing finding from the study. Evidence that the variance for each and covariance among the three teaching practice factors decreased in association with increased exposure to treatment suggested that professional development may influence the conceptual clarity with which respondents report the

traditional, investigative practice and culture aspects of their teaching. The reporting of change scores in isolation, without assessing variance/covariance factorial invariance, would not capture evidence of changes in the relationships among constructs.

In summary, relative to the work of Smithson & Porter (1994) and others on the relationship between training and the accuracy of self-report behavioral data, it was not surprising that as teachers became more exposed to the tenets and tenor of the science educational reform agenda, the clarity with which they perceived science teaching practice as a multi-dimensional construct increased. Teachers not yet exposed to reform-minded professional development were found to have a fairly fuzzy and diffuse mental model of teaching practice. Evidence here suggested that they see the investigative aspects of their science teaching (practice and culture) as nearly indistinguishable from one another and only somewhat dissimilar from the traditional aspects of teaching practice. On the other hand, teachers exposed to the highest levels of professional development appear to have come to express a more tightly focused mental model for teaching practice where the reform factors are substantially more distinct and divergent. There was evidence to support the assertion that the investigative aspects of science teaching (practice and culture) are seen to be related to each other conceptually to a much lesser degree by exposed groups than by the unexposed group. In addition, there was evidence to support that neither the investigative practice nor culture aspects of science teaching were perceived by exposed groups to be as similar conceptually to traditional aspects of practice as for the unexposed group. This in and of itself presents compelling evidence that conceptual change can and should be captured as a routine part of evaluating the efficacy of treatments thought to influence knowledge and behavioral structures.

Evidence of Mean Level Change

In addition, given that these data support the condition of strong factorial invariance

across groups, it was reasonable to assert that there was an association found between exposure to reform-minded professional development and increases in the mean level of factor scores. Rejection of the null hypothesis for factor mean level difference was interpreted to illustrate the strength and direction of the association of reform-minded professional development with the self-reported frequency of traditional and investigative science teaching practices.

The frequency with which teachers report engaging in traditional science teaching practices was found to be quite stable relative to treatment exposure. This finding was consistent with observational and interview data reported in the literature that suggests teachers add reformed practices to an already full portfolio of teaching practices that continues to rely on a set amount of lecture, textbooks, and worksheets (Spillane & Zeuli, 1999). There was no evidence to suggest that there was any reduction in the self-reported frequency of traditional teaching practices[7].

In addition, these data suggested that teachers are more likely to report increases in the frequency with which their students engage in episodic science experiments, simulations, and projects than the frequency with which students delve into the more substantive scientific habits of mind like reliance on problem-solving, reference, and reflection tools. As suggested by Spillane & Zeuli (1999) some aspects of multi-dimensional teaching practice are more responsive and amenable to reform than others. As reported here, evidence of changes in science teaching practice is slower to emerge for exposing students to the scientific habits of mind than it is for providing opportunity for students to engage in scientific activities.

---

[7] LSC evaluation did not include findings related to the Traditional Practices factor; however, moderate effect sizes (0.66-0.69 respectively) were reported across treatment groups for the Investigative Culture and Practices factors (Weiss et al., 1998, p. 80). It is important to note that the construction and scoring of the factors/composites were approached differently in the HRI evaluation and this study.

Application of Findings to Systemic Reform as a Change Venue

> Because different teachers bring different knowledge, beliefs, and experience to reformers proposals they often construct different ideas about what the reforms mean for their teaching and pursue different courses of action. An issue here concerns what if any patterns exist in teachers' diverse responses to reform (Spillane & Zeuli, 1999, p. 2).

The reform community has long held that teachers possess the key to the initiation and sustenance of lasting change (Darling-Hammond, 1990; Fullan, 1993; Holmes Group, 1986, 1990; Shields, Anderson, Bamburg, Hawkins, Knapp, Ruskin, & Wilson, 1995). Systemic reform requires teachers to revise their roles and responsibilities in order to acquire the knowledge and skill needed to implement the mandated changes in curriculum and instruction that serve as the necessary antecedents to improvements in student achievement. "The success of efforts to increase and reach high standards depends largely on the success of teachers and their ability to acquire the content knowledge and instructional practices necessary to teach to high academic standards" (Improving America's Schools Association, 1996, p. 5). The current national and state systemic reform policy agenda requires much of educators, and the institutions they serve, in order to improve the educational environment in ways that will positively impact student achievement (Cohen & Spillane, 1991; Darling-Hammond, 1997a, 1997b). These findings suggest that exposure to reform-minded professional development can be strongly associated with at least the initiation of the kinds of conceptual and behavioral changes sought by the reform movement.

Systemic educational reform takes a wide-angle view of school change that regards all parts of the system as a whole and recognizes that to achieve enduring change, every component of the system must be "irreversibly and permanently altered" (National Science Foundation, 1997, p. 2). This reform perspective is patterned after modern models for change that transcend the mechanistic, reductionist models that once dominated contemporary thought (Watzlawick, Weakland, & Fisch, 1974). Systemic, or systems thinking

(Senge, 1990) requires reformers to acknowledge that the way that educational systems are put together has effects on the way people perform, what they acquire, and how they master what they learn (Schlechty, 1997). Systemic reform emphasizes standards-based coherence among state and local policies, in the hope that coordinated policies will influence progress toward classroom practices that are aligned with state and emerging national curricular goals (Fuhrman, 1993). The implicit overarching assumption here is that reform strategies impact the educational system–organizational structure and culture–in such a way as to encourage changes in the teaching and learning environment at the classroom level thought to support improvement in student learning and achievement (Goertz, Floden, & O'Day, 1995; Schlechty, 1997). The logic that drives much of the current, systemic reform agenda, which focuses mainly on science and mathematics, is based on three assumptions:

1. The first assumption is that standards-based reform strategies directed at organizations–for example, the alignment of curriculum and assessment policies at the state level or instructional leadership and professional development at the district and building level–promote concomitant changes at the individual level–for instance, the will and capacity of teachers to embrace and enact the reform agenda (Cohen & Spillane, 1993; Corcoran, 1995; Darling-Hammond, 1990; Darling-Hammond & McLaughlin, 1995).

2. The second assumption is that changes in the will and capacity of teachers to engage in reform are associated with the modification of teaching practice toward the reform ideal–depth over breadth, student-centered instructional strategies, and the creation of an investigative learning environment (Cohen, McLaughlin, & Talbert, 1993; NSF, 1994).

3. The third assumption is that changes in teaching practice are required to improve student achievement (Brophy & Good, 1986; Center for Policy Research in Education, 1995).

Consequently, these assumptions contribute to the formulation of an implicit theory of systemic reform. This implicit theory predicts the impact of current policies on student achievement, either at the state or local level, will be limited by the extent to which teachers

and the organizations within which they work possess the will and capacity to contribute toward the evolution of classroom practice (Knapp, 1997; Spillane, 1994; Spillane & Thompson, 1997). Given that one of the major thrusts of the systemic educational reform movement is to transform classroom practice from traditional teacher-centered methods toward more student-centered constructivist methods these data provide evidence in support of movement along this intended trajectory. Evidence of conceptual change supports aspects of teacher will and evidence of mean level change supports aspects of teacher capacity to engage in reform to the extent that the constraints and conditions inherent to the evaluation allow.

Limitations of the Study and Suggestions for Further Research

There are at least four methodological limitations to both the execution and interpretation of the secondary data analysis presented here. These methodological limitations that were identified *a priori* to the study also point to valuable avenues along which further research in this area could be pursued. They include within and/or between project variability, within individual variability, alternative CFA models not under consideration, and level of measurement.

Project and Individual Variability

Since the evaluation of differences within or between projects was not the focus of this study, the clustered nature of the data set was not considered here and these data were analyzed in aggregate. All of the projects adhere to a set of LSC principles that align implementation strategies with current systemic reform theory; therefore little variation was expected to be contributed by specific differences in project strategies. Although not considered here, the determination of factorial invariance across projects from an hierarchical CFA perspective would be an excellent additional line of investigation (Hox, 1994; Muthén,

1991). In addition, variability in teaching practice outcomes across projects due to the influence of systemic reform strategies, such as, state standards, or policy coherence and restructuring at the local level, was neither examined nor utilized to assess the plausibility of the systemic reform assumptions; however, this as well may serve as an additional avenue for future research.

Although the costs of survey designs that include repeated measures are often prohibitive given the limited budget allocated for the purposes of project or initiative evaluation, they may well be the best approach for the study of change over time (Collins, 1991). The practice of opting not to follow the same respondents over time complicates any longitudinal approach to evaluation of the effectiveness of an on-going treatment. The interpretation of any results from a non-experimental approach such as this must take the oft-cited threats to validity into account (Trochim, 1998). In particular, the evaluator must assume that the control and treatment groups were comparable prior to the initiation of the LSC initiative and that social threats such as diffusion of treatment influence was minimal during the first year. In that it was not possible to consider individual rates of change, newer methods such as latent growth curve modeling that would have been applicable with a more dynamic, longitudinally sensitive design, were not appropriate. As with the previous limitation this also presents an interesting and possibly fruitful line of further research.

Alternative CFA Models and Level of Measurement

As noted by Jöreskog (1993), there are many additional alternative models that could have been envisioned and tested to fit the data equally well or better than those assessed by this or any given study. The model that ultimately best represented these data from among those tested was only one of the many models that could have been specified for these data. Jöreskog (1993) cautions users of CFA to avoid the pitfalls of failing to recognize that they have not tested the universe of models, just a select few. It is important to note the subjective

nature of this study and to alert the reader to the possibility that other configurations and analyses may have provided alternative scenarios and explanations of the phenomenon at hand.

Although CFA relies on hypotheses, it is not an exact science. In fact, like its cousin EFA it is truly more of an art. In addition, although it was conceived probable, at the time this study was proposed, that the more restrictive forms of factorial invariance would be rejected, this study did not include any pursuit of evaluating CFA results using the partial metric invariance methods proposed by Byrne, Shavelson, and Muthén (1989). This too, may prove to provide future opportunity to deepen the knowledge base for the study of change.

As is common practice for most evaluations today, the scores used in these analyses were obtained from 5-point Likert scales. CFA and the related structural equation modeling techniques are based on stringent assumptions of linearity—Jöreskog (1993) asserts that rating scale data provide only weak support for such assumptions. The use of ordinal data in analyses, such as those presented here, are quite common. Many authors make convincing arguments as to why the practice of assuming an underlying interval scale is acceptable without the modifications suggested by Jöreskog (1993) and implemented to execute this research. Another interesting follow-up study would be to determine the impact of the methodological decision to use the fixed threshold method, as suggested to convert these data to an interval scale, rather than rely on the assumption of underlying continuous variables. Clearly, somewhat different conclusions would have probably been drawn if these analyses were based on unconverted ordinal scale data.

Conclusion

This study illuminates the measurement challenges that arise under the clear design constraints present in the at best quasi-experimental conditions frequently faced by evaluators. This study affirms the importance of determining and reporting the extent to

which comparison groups share the same mental model for the construct under investigation. CFA techniques present a valuable new lens with which to address many of the measurement challenges inherent in the real world practice of evaluation. CFA can be used to examine the mental models held across groups as well as to detect and disentangle the conceptual from mean level changes that occur in association with categorical group membership, treatment exposure, and/or the passage of time. Group LISREL thus presents a theoretically appealing and now empirically tested channel to explore the extent to which survey respondents relate observed measures to latent constructs in the same way across groups.

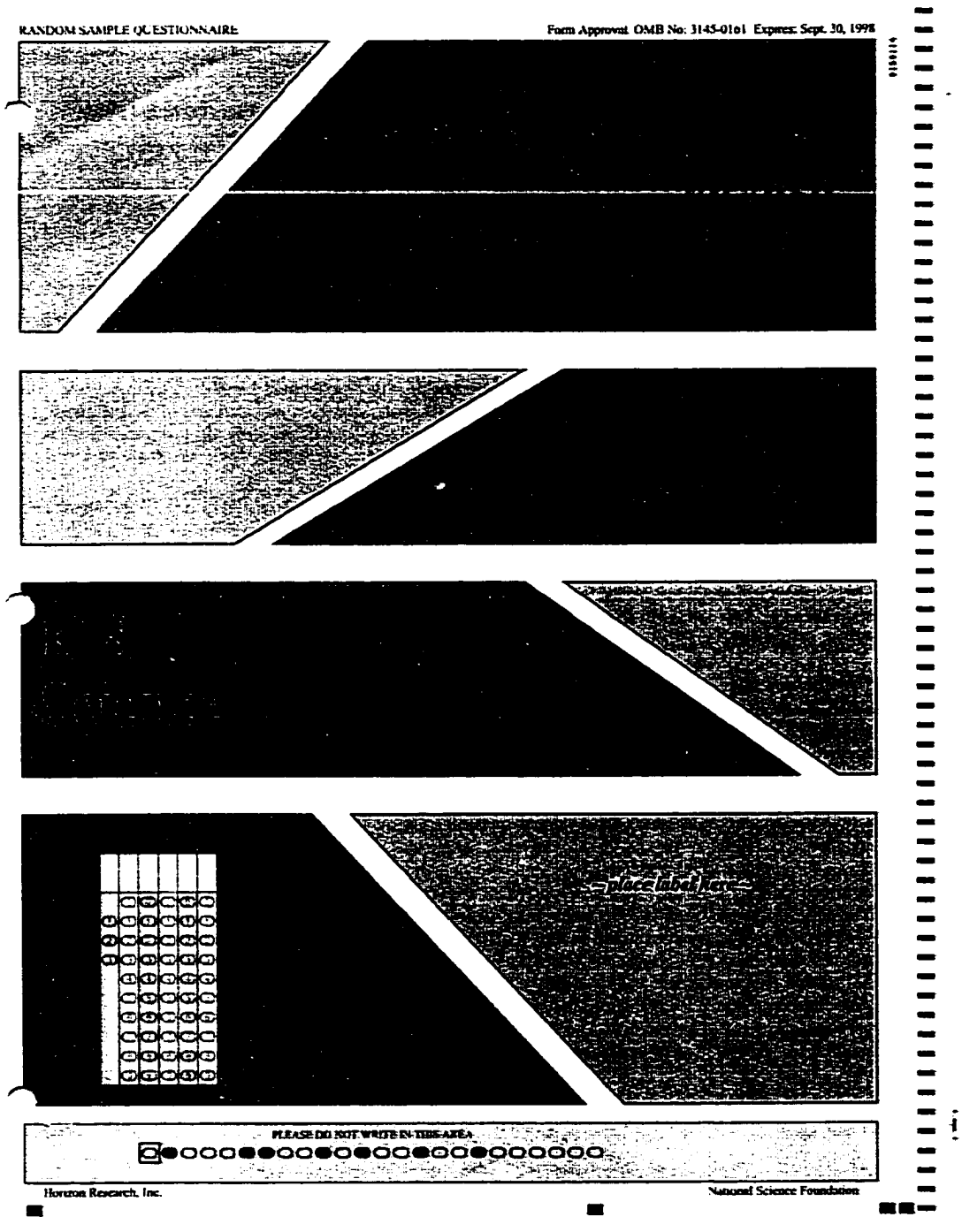
Under those situations common to evaluation where groups are identified rather than assigned as control and treatment groups the use of composite scores to compare the efficacy of an intervention or to compare group differences is severely compromised. At best the strongest assertions the majority of evaluations are able to make are in relation to the contribution of a treatment or other program influences to observed group differences and only under the most stringent of conditions can the case for attribution be made. Adequate evidence to support the equivalence of measurement structures across comparison groups under marginal conditions would serve to strengthen the measurement position that belies findings of conceptual or mean level change. However, should the groups of interest be shown not to share the same mental model for the construct(s) under investigation (i.e., fail to demonstrate strong factorial invariance) then the use of composite scores to compare the level and or relationship among constructs across treatment groups should not be supported. It is important to note that under some conditions analyses such as those performed here could be expected to fail to support configural or weak factorial invariance and thus, provide evidence of large-scale conceptual change. The CFA techniques explored here can and should be used more frequently to better explicate and understand the complex nature of the change process. In so doing evaluators will be better able to describe and track the changes, both conceptual and mean level, that occur across groups in association with exposure to the

treatments and interventions currently targeted by program evaluation practice.

These CFA methods, in particular as applied here to the venue of the systemic reform of science education, speak as well to increasing our understanding of the evolving relationship between reform intent and teaching practice. The work of Spillane & Jennings (1997) indicated that teacher beliefs about subject matter as well as attitudes toward teaching and learning all contribute to the manner in which teachers interpret instructional policies and construct mental models related to the reform of teaching practice. The methods explored here identify the very patterns of conceptual reconstitution related to instructional practice theorized to occur in the wake of reform (Spillane & Zeuli, 1999).

APPENDIX A

Local Systemic Change Through Teacher Enhancement--1997 Teacher Questionnaire



The National Science Foundation's Local Systemic Change (LSC) through Teacher Enhancement Program's Core Evaluation

You have been selected to participate in the nationwide evaluation of the federally-funded Local Systemic Change (LSC) program. LSC is a National Science Foundation Teacher Enhancement program that is currently funding about 50 local projects that offer science and mathematics professional development to teachers in 23 states around the country. The cover letter accompanying this questionnaire identifies the LSC project in your area.

A variety of strategies
The general purpose of LSC projects is to offer teachers high-quality professional development in content and pedagogy. These activities are based on the national standards for reforming science and mathematics education. LSC projects are reaching teachers in grades K-12, although most local projects focus on either elementary or secondary teachers. LSC initiatives are helping teachers around the country to implement quality science and mathematics curriculum materials. The size, strategies, and activities of the individual LSC projects vary widely based on local needs.

The national evaluation
The National Science Foundation is accountable to Congress for the programs it funds, and the purpose of the LSC core evaluation is to provide both the leadership at NSF, and ultimately Congress, with information about the quality and impact of the Local Systemic Change program. This national evaluation is a system for collecting similar information from all LSC projects through various means, including teacher and principal questionnaires. A small number of randomly-selected teachers in each project is asked to provide additional information in interviews, sometimes in conjunction with a classroom visit. In order to continue receiving federal funding, each LSC project must participate in this national evaluation.

This questionnaire
Each LSC project will administer questionnaires each spring to a randomly-selected sample of teachers who are targeted to participate in the local project's professional development activities. (A different group of teachers will be selected each year, but there is a chance over the course of several years that you could be selected to participate again in the future. For statistical reasons, some smaller LSC projects must administer this questionnaire to each participating teacher annually.) Note that you may be asked to complete this questionnaire even if you have not yet participated in the project's professional development; your response is important, regardless of whether you have already participated.

Confidentiality
Data collection procedures have been developed to ensure high quality data and protect teacher confidentiality. Your responses will be kept strictly confidential; they will be combined with the responses of the other teachers in your project and used only for the LSC evaluation. The name label and numbering on this questionnaire are used to help local projects deliver questionnaires to the proper teachers and follow up with teachers that have not responded; no information identifying individual teachers will be reported under any circumstances. After you complete the questionnaire, you should remove the name label and return the questionnaire as specified by your local LSC project.

Thank you very much for participating in this survey!

DesignExpert™ by NCS Printed in U.S.A. Mark Reflex® EVR-210895-1-854321 HR36

Instructions: Please use a #2 pencil to complete this questionnaire. Darken ovals completely, but do not stray into adjacent ovals. Be sure to erase completely any stray marks.

Teacher Opinions and Preparedness

1. Please provide your opinion about each of the following statements. (Darken one oval on each line.)

	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
a. Students generally learn science best in classes with students of similar abilities.	○	○	○	○	○
b. I feel supported by colleagues to try out new ideas in teaching science.	○	○	○	○	○
c. Teachers in this school have a shared vision of effective science instruction.	○	○	○	○	○
d. Teachers in this school regularly share ideas and materials related to science.	○	○	○	○	○
e. Teachers in this school are well-supplied with materials for investigative science instruction.	○	○	○	○	○
f. I have time during the regular school week to work with my peers on science curriculum and instruction.	○	○	○	○	○
g. I have adequate access to computers for teaching science.	○	○	○	○	○
h. I enjoy teaching science.	○	○	○	○	○
i. I am well-informed about the NRC *National Science Education Standards* for the grades I teach.	○	○	○	○	○
j. The science program in this school is strongly supported by local organizations, institutions and/or businesses.	○	○	○	○	○

2. In the left section, please rate each of the following in terms of its **importance** for effective science instruction in the grades you teach. In the right section, please indicate how **prepared** you feel to do each one. (Darken one oval in each section on each line.)

	Importance				Preparation			
	Not Important	Somewhat Important	Fairly Important	Very Important	Not Adequately Prepared	Somewhat Prepared	Fairly Well Prepared	Very Well Prepared
a. Provide concrete experience before abstract concepts.	○	○	○	○	○	○	○	○
b. Develop students' conceptual understanding of science.	○	○	○	○	○	○	○	○
c. Take students' prior understanding into account when planning curriculum and instruction.	○	○	○	○	○	○	○	○
d. Make connections between science and other disciplines.	○	○	○	○	○	○	○	○
e. Have students work in cooperative learning groups.	○	○	○	○	○	○	○	○
f. Have students participate in appropriate hands-on activities.	○	○	○	○	○	○	○	○
g. Engage students in inquiry-oriented activities.	○	○	○	○	○	○	○	○
h. Use computers.	○	○	○	○	○	○	○	○
i. Engage students in applications of science in a variety of contexts.	○	○	○	○	○	○	○	○
j. Use performance-based assessment.	○	○	○	○	○	○	○	○
k. Use portfolios.	○	○	○	○	○	○	○	○
l. Use informal questioning to assess student understanding.	○	○	○	○	○	○	○	○

3. My school principal: (Darken one oval on each line.)

	Strongly Disagree	Disagree	No Opinion	Agree	Strongly Agree
a. Encourages me to select science content and instructional strategies that address individual students' learning.	○	○	○	○	○
b. Accepts the noise that comes with an active classroom.	○	○	○	○	○
c. Encourages the implementation of current national standards in science education.	○	○	○	○	○
d. Encourages innovative instructional practices.	○	○	○	○	○
e. Enhances the science program by providing me with needed materials and equipment.	○	○	○	○	○
f. Provides time for teachers to meet and share ideas with one another.	○	○	○	○	○
g. Encourages me to observe exemplary science teachers.	○	○	○	○	○
h. Encourages teachers to make connections across disciplines.	○	○	○	○	○
i. Acts as a buffer between teachers and external pressures (e.g., parents).	○	○	○	○	○

4. Many teachers feel better prepared to teach some subject areas than others. How well prepared do you feel to teach each of the following subjects at the grade levels you teach, whether or not they are currently included in your curriculum? (Darken one oval on each line.)

	Not Adequately Prepared	Somewhat Prepared	Fairly Well Prepared	Very Well Prepared
a. Science	○	○	○	○
b. Mathematics	○	○	○	○
c. Reading/Language Arts	○	○	○	○
d. Social Studies	○	○	○	○

5. Within science, many teachers feel better prepared to teach some topics than others. How well prepared do you feel to teach each of the following topics at the grade levels you teach, whether or not they are currently included in your curriculum? (Darken one oval on each line.)

	Not Adequately Prepared	Somewhat Prepared	Fairly Well Prepared	Very Well Prepared
a. The human body	○	○	○	○
b. Ecology	○	○	○	○
c. Rocks and soils	○	○	○	○
d. Astronomy	○	○	○	○
e. Processes of change over time (e.g., evolution)	○	○	○	○
f. Mixtures and solutions	○	○	○	○
g. Electricity	○	○	○	○
h. Sound	○	○	○	○
i. Forces and motion	○	○	○	○
j. Machines	○	○	○	○
k. Engineering and design principles (e.g., structures, models)	○	○	○	○

6. Please indicate how well prepared you feel to do each of the following. (Darken one oval on each line.)

	Not Adequately Prepared		Somewhat Prepared		Fairly Well Prepared		Very Well Prepared
a. Lead a class of students using investigative strategies.							
b. Manage a class of students engaged in hands-on/project-based work.							
c. Help students take responsibility for their own learning.							
d. Recognize and respond to student diversity.							
e. Encourage students' interest in science.							
f. Use strategies that specifically encourage participation of females and minorities in science.							
g. Involve parents in the science education of their students.							

7. Please rate the effect of each of the following on your science instruction. (Darken one oval on each line.)

	Inhibits effective instruction		Neutral or mixed		Encourages effective instruction	NA/ Don't Know
a. State and/or district curriculum frameworks.						
b. State and/or district testing policies and practices.						
c. State, district, and/or school grading policies and practices.						
d. District/school structures for recognizing and rewarding teachers.						
e. Quality of available instructional materials.						
f. Access to computers for science instruction.						
g. Funds for purchasing equipment and supplies for science.						
h. System of managing instructional resources at the district or school level.						
i. Time available for teachers to plan and prepare lessons.						
j. Opportunities for teachers to work with other teachers.						
k. Opportunities for teacher professional development.						
l. Importance that the school places on science.						
m. Consistency of science reform efforts with other school/district reforms.						
n. Public attitudes toward reform.						

8. How many of your students' parents do each of the following? (Darken one oval on each line.)

	Few or none		About 1/2		Almost All
a. Volunteer to assist with class activities.					
b. Donate money or materials for classroom instruction.					
c. Attend parent-teacher conferences.					
d. Attend school activities such as PTA meetings and Family Science nights.					
e. Voice support for the use of an investigative approach to science instruction.					
f. Voice support for traditional approaches to science instruction.					

B. Your Science Teaching

9. What grade level(s) are you currently teaching? (Darken all ovals that apply.)

Questions 10-14 ask about your science teaching. If you teach more than one science class, please answer for your first class of the day.

10. About how often do you do each of the following in your science instruction? (Darken one oval on each line.)	Never	Rarely (e.g., a few times a year)	Sometimes (e.g., once or twice a month)	Often (e.g., once or twice a week)	All or almost all science lessons
a. Introduce content through formal presentations.	○	○	○	○	○
b. Demonstrate a science-related principle or phenomenon.	○	○	○	○	○
c. Arrange seating to facilitate student discussion.	○	○	○	○	○
d. Use open-ended questions.	○	○	○	○	○
e. Require students to supply evidence to support their claims.	○	○	○	○	○
f. Encourage students to explain concepts to one another.	○	○	○	○	○
g. Encourage students to consider alternative explanations.	○	○	○	○	○
h. Allow students to work at their own pace.	○	○	○	○	○
i. Help students see connections between science and other disciplines.	○	○	○	○	○
j. Use assessment to find out what students know before or during a unit.	○	○	○	○	○
k. Embed assessment in regular class activities.	○	○	○	○	○
l. Assign science homework.	○	○	○	○	○
m. Read and comment on the reflections students have written in their notebooks or journals.	○	○	○	○	○

11. About how often do students in this class take part in each of the following types of activities as part of their science instruction? (Darken one oval on each line.)	Never	Rarely (e.g., a few times a year)	Sometimes (e.g., once or twice a month)	Often (e.g., once or twice a week)	All or almost all science lessons
a. Participate in student-led discussions.	○	○	○	○	○
b. Participate in discussions with the teacher to further science understanding.	○	○	○	○	○
c. Work in cooperative learning groups.	○	○	○	○	○
d. Make formal presentations to the class.	○	○	○	○	○
e. Read from a science textbook in class.	○	○	○	○	○
f. Read other (non-textbook) science-related materials in class.	○	○	○	○	○
g. Answer textbook/worksheet questions.	○	○	○	○	○
h. Review homework/worksheet assignments.	○	○	○	○	○
i. Work on solving a real-world problem.	○	○	○	○	○
j. Share ideas or solve problems with each other in small groups.	○	○	○	○	○
k. Engage in hands-on science activities.	○	○	○	○	○
l. Follow prescribed steps in an activity or investigation.	○	○	○	○	○
m. Design or implement their *own* investigation.	○	○	○	○	○
n. Design objects within constraints (e.g., egg drop, toothpick bridge, aluminum boats).	○	○	○	○	○
o. Work on models or simulations.	○	○	○	○	○
p. Work on extended science investigations or projects (a week or more in duration).	○	○	○	○	○

11. (continued)

	Never	Rarely (e.g., a few times a year)	Sometimes (e.g., once or twice a month)	Often (e.g., once or twice a week)	All or almost all science lessons
q. Participate in field work.					
r. Record, represent, and/or analyze data.					
s. Write reflections in a notebook or journal.					
t. Prepare written science reports.					
u. Use mathematics as a tool in problem-solving.					
v. Use computers.					
w. Work on portfolios.					
x. Take short answer tests (e.g., multiple choice, true/false, fill-in-the-blank).					
y. Take tests requiring open-ended responses (e.g., descriptions, explanations).					
z. Engage in performance tasks for assessment purposes.					

12. In how many of the last five school days did you teach each of the following in this class? (Darken one oval on each line.)

	Number of Days					
	none	one	two	three	four	five
a. Science						
b. Mathematics						
c. Reading/Language Arts						
d. Social Studies						

13. Which of the following activities were included in your **most recent science lesson** in this class? (Darken all ovals that apply.)

- a. Formal presentation by teacher
- b. Small group work
- c. Hands-on/investigative/research/field activities
- d. Reading about science
- e. Work on solving real-world or abstract problems
- f. Use of computers
- g. Answering textbook/worksheet questions
- h. Writing reflections in a notebook or journal
- i. Informal assessment (e.g., questioning for understanding)
- j. Short-answer tests
- k. Tests requiring open-ended responses
- l. Performance-based assessments
- m. Work on portfolios

14. How much time was spent on each of the following in that lesson? (Darken all that apply.)

	None	10 minutes or less	11-20 minutes	21-30 minutes	More than 30 minutes
a. life science					
b. physical science					
c. earth/space science					
d. engineering and design principles					

C. LSC Professional Development

Questions 15-18 refer to the NSF-supported Local Systemic Change (LSC) program. Please refer to the letter accompanying this questionnaire for information about the LSC project activities in your district.

15. To what extent is each of the following true of LSC science-related professional development in your district? (Darken one oval on each line.)

	Not at all				To a great extent
a. Adequate opportunities are available to me for science-related professional development.	①	②	③	④	⑤
b. I am involved in planning my science-related professional development.	①	②	③	④	⑤
c. I am encouraged to develop an individual professional development plan to address my needs and interests related to science education.	①	②	③	④	⑤
d. I am given time to work with other teachers as part of my professional development.	①	②	③	④	⑤
e. I am given time to reflect on what I've learned and how to apply it to the classroom.	①	②	③	④	⑤
f. I receive support as I try to implement what I've learned.	①	②	③	④	⑤

16. Approximately how many ***hours*** have you spent on professional development in science/science education as part of the LSC project? (Darken one oval.)

○ 0	○ 20-39	○ 80-99	○ 160-199
○ 1-9	○ 40-59	○ 100-129	○ 200 or greater
○ 10-19	○ 60-79	○ 130-159	

17. How would you rate the overall quality of the LSC professional development? (Darken one oval.)

Very Poor	Poor	Fair	Good	Very Good	Excellent
○	○	○	○	○	○

18. Have you been identified as a lead teacher for your district's NSF-supported LSC project?

○ Yes
○ No

D. Teacher Demographic Information

19. Are you: ○ Male ○ Female

20. Are you:

○ African-American (not of Hispanic origin)
○ American Indian or Alaskan Native
○ Asian or Pacific Islander
○ Hispanic
○ White (not of Hispanic origin)
○ Other

21. Did your college science coursework include the equivalent of at least one semester of: (Darken one oval on each line.)

	Yes	No
a. life science	○	○
b. earth and space science	○	○
c. physical science	○	○

22. How many college science courses have you completed? (Darken one oval.)

○ none
○ 1 semester
○ 2 semesters
○ 3 semesters
○ 4 semesters
○ 5 or more semesters

23. How many years have you taught prior to this school year? (Darken one oval.)

○ 0-2
○ 3-5
○ 6-10
○ 11-20
○ 21+

PLEASE DO NOT WRITE IN THIS AREA

APPENDIX B

Instrument Duplication Permission From Horizon Research, Inc

horizon
RESEARCH, INC.

October 1, 1999

To Whom It May Concern:

Cynthia Phillips has permission to: (1) use the partial data set requested; (2) reproduce the 1997 K–8 Science Teacher Questionnaire to attach to HSIRB application; and (3) reproduce the 1997 K–8 Science Teacher Questionnaire as an appendix, with the understanding that Horizon Research, Inc. and the National Science Foundation (RED-9255369) be appropriately referenced.

Sincerely,

Iris R. Weiss
President

IRW/sbh

111 Cloister Court • Suite 220 • Chapel Hill, NC 27514-2296
(919) 489-1725 • Fax (919) 493-7589 • hri@horizon-research.com

APPENDIX C

Protocol Clearance From the Human Subjects Institutional Review Board

Human Subjects Institutional Review Board

Kalamazoo, Michigan 49008-3899

WESTERN MICHIGAN UNIVERSITY

Date: 2 November 1999

To: MaryAnne Bunda, Principal Investigator
Cynthia Phillips, Student Investigator for dissertation

From: Sylvia Culp, Chair

Re: HSIRB Project Number 99-09-24

This letter will serve as confirmation that your research project entitled "The Structural Dynamics of Conceptual Change: Teachers Evolving Perceptions of Classroom Practice" has been **approved** under the **expedited** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note that you may **only** conduct this research exactly in the form it was approved. You must seek specific board approval for any changes in this project. You must also seek reapproval if the project extends beyond the termination date noted below. In addition if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: 2 November 2000

BIBLIOGRAPHY

Ahmavaara, Y. (1954). The mathematical theory of factorial invariance under selection. Psychometrika, 19, 27-38.

Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. Evaluation Review, 14, 374-390.

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. Journal of Consulting and Clinical Psychology, 50, 488-499.

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson & E. F. Borgatta (Eds.), Factor analysis and measurement in sociological research (pp. 249-279). Beverly Hills, CA: Sage Publications, Inc.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103, 411-423.

Argyris, C., & Schön, D. A. (1974). Theory in practice: Increasing professional effectiveness. San Francisco: Jossey-Bass.

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. Structural Equation Modeling, 1, 35-67.

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

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bollen, K. A. (1989). Structural equations with latent variables. New York: John Wiley.

Bollen, K. A., & Long, J. S. (Eds.). (1993). Testing structural equation models. Newbury Park, CA: Sage Publications.

Braverman, M. T. (1996). Sources of survey error: Implications for evaluation studies. In M. T. Braverman & J. K. Slater (Eds.), New directions for evaluation, 70, 17-28. San Francisco: Jossey-Bass Publishers.

Brooks, J. G., & Brooks, M. G. (1993). The case for constructivist classrooms. Alexandria, VA: Association for Supervision and Curriculum Development.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 328-376). New York: Macmillan.

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 Publications

Burt, C. (1939). The relations of educational abilities. British Journal of Educational Psychology, 9, 45-71.

Byrne, B. M. (1998). Structural equation modeling with Lisrel, Prelis, and Simplis: Basic concepts, applications, and programming (Multivariate Applications Book Series). Hillsdale, NJ: Lawrence Erlbaum Associates.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariances and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105, 456-466.

Campbell, D. T., and Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Cattell, R. B. (1947). r_p and other coefficients of pattern similarity. Psychometrika, 14: 279-298.

Center for Policy Research in Education. (1995). Reforming science, mathematics, and technology education: NSF's Statewide Systemic Initiatives (CPRE Policy Brief RB-15). New Brunswick, NJ: Rutgers University, Author.

Chou, C., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications (pp. 37-55). Newbury Park, CA: Sage Publications.

Cohen, D. K., McLaughlin, M. W., & Talbert, J. E. (1993). Teaching for understanding: Challenges for policy and practice. San Francisco: Jossey-Bass.

Cohen, D. K., & Spillane, J. P. (1991). Policy and practice: The relationship between governance and instruction. (ERIC Document Reproduction Service No. ED 337 865).

Collins, L. M. (1991). Measurement in longitudinal research. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions. (pp. 137-148). Washington, DC: American Psychological Association.

Comrey, A. L., & Lee, H. B. (1992). A first course in factor analysis. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Corcoran, T. (1995). Helping teachers teach well: Transforming professional development. Consortium for Policy Research in Education (CPRE) Policy Brief RB-16.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich, Inc.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change—"or should we? Psychological Bulletin, 74, 68-80.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cudeck, R. & Browne, M. W. (1983). Cross-validation of covariance structures. Multivariate Behavioral Research, 18, 147-157.

Cunningham, W. R. (1991). Issues in factorial invariance. In L. M. Collins & J. L Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions. (pp. 106-113). Washington, DC: American Psychological Association.

Darling-Hammond, L. (1990). Instructional policy into practice: "The power of the bottom over the top." Educational Evaluation and Policy Analysis, 12,339-347.

Darling-Hammond, L. (1997a). Quality teaching: The critical key to learning. Principal, 77(1), 5-6.

Darling-Hammond, L. (1997b). School reform at the crossroads: Confronting the central issues of teaching. Educational Policy, 11(2), 151-166.

Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. Phi Delta Kappan, April, 597-604.

Drascow, F. (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19-29.

Drascow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology, 70, 662-680.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. Educational Researcher, 32(7), 5-12.

Evans, R. (1996). The human side of school change: Reform, resistance, and the real-life problems of innovation. San Francisco: Jossey-Bass.

Elmore, R., Peterson, P., & McCarthy, S. J. (1996). Restructuring in the classroom: Teaching, learning, and school organization. San Francisco: Jossey-Bass Publishers.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp.105-146). New York: Macmillan.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the Self-Description Questionnaire II. Journal of Research in Personality, 29, 168-188.

Flora, D. B., & Panter, A. T. (1998). Technical report: Analysis of the psychometric structure of the LSC surveys. Chapel Hill, NC: University of North Carolina, Thurstone Psychometric Laboratory.

Fosnot, C. T. (1993). Rethinking science education: A defense of Piagetian constructivism. Journal for Research on Science Teaching, 30, 1189-1201.

Frederiksen, N. (1987). How do you tell if a test measures the same thing in different cultures? In Y. H. Pootinga (Ed.), Basic problems in cross cultural psychology (pp. 14-18). Amsterdam: Swets & Zeitlinger.

Fuhrman, S. H. (1993). The politics of coherence. In S. H. Fuhrman (Ed.), Designing coherent educational policy: Improving the system (pp. 1-34). San Francisco: Jossey-Bass.

Fullan, M. (1993). Change forces: Probing the depths of educational reform. New York: The Falmer Press.

Fullan, M. (1995). The school as a learning organization: Distant dreams. Theory into Practice, 34(4), 230-235.

Gabella, M. S. (1995). Unlearning certainty: Toward a culture of student inquiry. Theory into Practice, 34(4), 236-242.

Gerbing, D. W., & Anderson, J. C. (1993). Monte carlo evaluation of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 40-65). Newbury Park, CA: Sage Publications.

Goertz, M. E., Floden, R. E., & O' Day, J. (1995). Studies of education reform: Systemic reform: Vol. I. Findings and conclusions. Newark, NJ: Rutgers, the State University of New Jersey, Center of Policy Research in Education.

Golembiewski, R., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. Journal of Applied Behavioral Science, 12, 133-157.

Halford, G. S. (1995). Learning processes in cognitive development: A reassessment with some unexpected implications. Human Development, 38(6), 295-301.

Heck, R. H., & Marcoulides, G. A. (1989). Examining the generalizability of administrative personnel allocation decisions. The Urban Review, 21, 51-62.

Hirsch, E. D. (1996). Reality's revenge: Research and ideology. American Educator, Fall: 4-6, 31-46.

Holmes Group (1986). Tomorrow's teachers: A report of the Holmes Group. East Lansing, MI: Author.

Holmes Group (1990). Tomorrow's schools: Principles for the design of professional development: A report. . East Lansing, MI: Author.

Horn, J. L. (1991). Comments on "Issues in factorial invariance." In L. M. Collins & J. L Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions. (pp. 114-125). Washington, DC: American Psychological Association

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18, 117-144.

Horn, McArdle, & Mason (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. Southern Psychologist, 4, 179-188.

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. Journal of Applied Psychology, 64, 144-150.

Howard, G. S., Schmeck, R. R., & Bray, J. H. (1979). Internal validity in studies employing self-report instruments: A suggested remedy. Journal of Educational Measurement, 16(2), 129-135.

Hox, J. J. (1994). Factor analysis of multilevel data: Gauging the Muthén model. In J. H. L. Oud & R. A. W. van Blokland-Vogelesang (Eds.), Advances in longitudinal and multivariate analysis in the behavioral sciences (pp. 141-156). Nijmegen, Netherlands: ITS.

Improving America's Schools Association (1996). Rethinking professional development. Newsletter on Issues in School Reform. Washington, DC: US Department of Education.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), Structural equation models in the social sciences (pp. 85-112). New York: Seminar Press.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. Quality and Quantity, 24, 387-404.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. Psychometrika, 59(3), 381-389.

Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Hillsdale, NJ: Lawrence Erlbaum Associates, Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996). PRELIS 2: User's reference guide. Hillsdale, NJ: Lawrence Erlbaum Associates, Scientific Software International.

Jöreskog, K. G., Sörbom, D., du Toit, S., & du Toit, M. (1999). Lisrel 8: New statistical features. Chicago: Scientific Software International.

Kelloway, E. K. (1998). Using Lisrel for structural equation modeling: A researcher's guide. Thousand Oaks, CA: Sage Publications.

Khattri, N., & Miles, M B. (1995). Mapping basic beliefs about learner centered schools. Theory into Practice, 34(4), 279-287.

Knapp, M. W. (1997). Between systemic reforms and the mathematics and science classroom: The dynamics of innovation, implementation, and professional learning. Review of Educational Research, 67(2), 227-266.

LaBouvie, E. & Ruetsch, C. (1995). Wholes or parts? Multivariate Behavioral Research, 30(1), 121-123.

Lindell, M. K., & Drexler, J. A. (1979). Issues in using survey methods for measuring organizational change. Academy of Management Review, 4, 13-19.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Louis, M. R. (1980). Surprise and sense-making: What newcomers experience in entering unfamiliar organizational settings. Administrative Science Quarterly, 25, 226-251.

Marcoulides, G. A., & Heck, R, H. (1993). Organizational culture and performance: Proposing and testing model. Organization Science, 4, 209-225.

Maruyama, G. M. (1998). Basics of structural equation modeling. Thousand Oaks, CA: Sage Publications.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? Educational Evaluation and Policy Analysis, 21(1), 29-45.

McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. Multivariate Behavioral Research, 29, 61-101.

McArdle, J.J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (eds.), Life-span developmental psychology: Methodological contributions (pp. 223-267). Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Non-centrality and goodness of fit. Psychological Bulletin, 107, 247-255.

McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. British Journal of Mathematical and Statistical Psychology, 24, 154-168.

Meredith, W. (1964a). Notes on factorial invariance. Psychometrika, 29, 177-185.

Meredith, W. (1964b). Rotation to achieve factorial invariance. Psychometrika, 29, 187-206.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. Psychometrika, 58, 525-543.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp.13-104). New York: Macmillan.

Millsap,R. E., and Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: a structural equation approach. Journal of Applied Psychology, 73, 574-584.

Mulaik, S. A. (1972). The foundations of factor analysis. New York: McGraw-Hill.

Murnane, R. J., Singer, J. D., & Willett, J. B. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research. Educational Researcher, 17(6), 22-30.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. Journal of Educational Measurement, 28, 338-354.

National Science Foundation. (1994). Foundation for the future: The systemic cornerstone. Washington, DC: Author (ERIC Document Reproduction Service No. ED 370 808).

National Science Foundation. (1996). The learning curve: What we are discovering about U. S. science and mathematics education (NSF 96-53). Washington, DC: National Science Foundation.

National Science Foundation. (1997). Foundations: The challenge and promise of K-8 science education reform. [On-line]. Available: http://www.her.nsf.gov/HER/ESIE/FOUNDTNS.htm.

National Science Resources Center (1997). Science for all children: A guide to improving elementary science education in your school district. Washington, DC: National Academy Press.

National Science Teachers Association (1997). NSTA pathways to the science standards: Guidelines for moving the vision into practice. Arlington, VA: NSTA.

Nunnally, J. C., & Bernstein, I. H. (1993). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Pinneau, S., & Newhouse, A. (1964). Measures of invariance and compatibility in factor analysis for fixed variables. Psychometrika, 29, 271-181.

Pitts, S. C., West, S. G., & J. Y. Tein (1996). Longitudinal measurement models in evaluation research: Examining stability and change. Evaluation and Program Planning, 19, 333-350.

Porras, J. I., & Berg, P. O. (1978). Evaluation methodology in organizational development: An analysis and critique. Journal of Applied Behavioral Science, 14, 151-173.

Rahim, M. A., & Magner, N. R. (1996). Confirmatory factor analysis of the bases of leader power: First-order factor model and its invariance across groups. Multivariate Behavioral Research, 31, 495-516.

Rallis, S. (1995). Creating learner centered schools: Dreams and practices. Theory into Practice, 34(4), 224-229.

Regional Educational Laboratories. (1995). Facilitating systemic change in science and mathematics education: A toolkit for professional developers. Andover, MA: The Regional Laboratory for Educational Improvement of the Northeast and Islands.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114, 552-566.

Reynolds, C. R., & Harding, R. E. (1983). Outcome in two large sample studies of factorial similarity under six methods of comparison. Educational and Psychological Measurement, 43, 723-728.

Rhoton, J., & Bowers, P. (Eds.). (1996). Issues in science education. Arlington, VA: National Science Teachers Association.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. Journal of Applied Psychology, 74, 892-900.

Schlechty, P. C. (1997). Inventing better schools: An action plan for education reform. San Francisco: Jossey-Bass.

Shields, P. M., Anderson, L., Bamburg, J., Hawkins, R. T, Knapp, M. S., Ruskus, J. R., & Wilson, C. L. (1995). Improving schools from the bottom up: From effective schools to restructuring. Washington, DC: U.S. Department of Education.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. Multivariate Behavioral Research, 17, 343-358.

Schmitt, N., Pulakos, E. D., & Lieblein, A. (1984). Comparison of three techniques to assess group-level beta and gamma change. Applied Psychological Measurement, 8, 249-260.

Schorr, L. B. (1997). Common purpose: Strengthening families and neighborhoods to rebuild America. New York: Anchor Books, Doubleday.

Senge, P. M. (1990). The fifth discipline: The art and practice of the learning organization. New York: Currency/Doubleday.

Sörbom, D., & Jöreskog, K. G. (1981). The use of LISREL in sociological model building. In D. J. Jackson & E. F. Borgatta (Eds.), Factor analysis and measurement in sociological research (pp. 201-220). Beverly Hills, CA: Sage Publications, Inc.

Smithson, J. L., & Porter, A. C. (1994). Measuring classroom practice: Lessons learned from the efforts to describe the enacted curriculum--The reform up-close study. Madison, WI: Consortium for Policy Research in Education.

Spillane, J. (1994). How districts mediate between state policy and teachers' practice. In R. F. Elmore & S. H. Fuhrman (Eds.), The governance of curriculum (pp.167-185). Alexandria, VA: Association for Supervision and Curriculum Development.

Spillane, J. P., & Jennings, N. E. (1997). Aligned instructional policy and ambitious pedagogy: Exploring instructional reform from the classroom perspective. Teachers College Record, 98(3), 449-481.

Spillane, J., & Thompson, C. L. (1997). Reconstruction conceptions of local capacity: The local education agency's capacity for ambitious instructional reform. Educational Evaluation and Policy Analysis, 19(2), 185-203.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. Educational Evaluation and Policy Analysis, 21(1), 1-27.

St. John, M., Century, J., Tibbetts, F., & Heenan, B. (1995). Reforming elementary science education in urban districts. Inverness, CA: Inverness Research Associates.

Taris,T. W., Bok, I. A., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric propoerties of multi-item concepts across different situations: A general approach. The Journal of Psychology, 132, 301-317.

Taylor, P. C. S. (1990). The influence of teacher beliefs on constructivist teaching practice. Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 17-20, 1990).

Terborg, J. R., Howard, G. S., & Maxwell, S. E. (1980). Evaluating planned organizational change: A method for assessing alpha, beta, and gamma change. Academy of Management Review, 5(1), 109-121.

Thompson, R. C., & Hunt, J. G. (1996). Inside the black box of alpha, beta, and gamma change: Using a cognitive-processing model to assess attitude structure. Academy of Management Review, 21(3), 655-690.

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

Thurstone, L. L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

Tisak, J., & Meredith, W. (1991). Longitudinal factor analysis. In A. von Eye (Ed.), Statistical methods in longitudinal research: Vol. 1. Principles and structuring change (pp. 125-150). San Diego: Academic Press.

Trochim, William M. The Research Methods Knowledge Base, (2nd ed.). [On-line]. Available: http://trochim.human.cornell.edu/kb/index.htm.

Tucker, L. (1951). A method for synthesis of factor analysis studies. Personnel Research Section Report No. 984. Washington, DC: Department of the Army.

Van de Vliert, E., Huismans, S. E., & Stok, J. L. L. (1985). The criterion approach to unraveling beta and alpha change. Academy of Management Review, 10(2), 269-274.

Watzlawick, P., Weakland, J. H., & Fisch, R. (1974). Change: Principles of problem formation and problem resolution. New York: Norton.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). An introduction to survey research, polling, and data analysis (3rd Ed.), Thousand Oaks, CA: Sage Publications.

Weiss, I. R., Montgomery, D. L., Ridgway, C. J., & Bond, S. L. (1998). Local systemic change through teacher enhancement: Year three cross-site report. Chapel Hill, NC: Horizon Research, Inc.

West, S. G., & Aiken, L. S. (1997). Toward understanding individual effects in multicomponent prevention programs: Design and analysis strategies. In K. J. Bryant, M. Windle, & S. G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research. (pp. 167-210). Washington, DC: American Psychological Association.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications (pp. 56-75). Newbury Park, CA: Sage Publications.

Widaman, K. F. (1991). Qualitative transitions amid quantitative development: A challenge for measuring and representing change. In L. M. Collins & J. L Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions. (pp. 204-217). Washington, DC: American Psychological Association.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance abuse domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), The science of prevention: Methodological advances

from alcohol and substance abuse research. (pp. 281-324). Washington, DC: American Psychological Association.

Windle, M., Iwawaki, S., & Lerner, R. M. (1988). Cross-cultural comparability of temperament among Japanese and American preschool children. International Journal of Psychology, 23, 547-567.

Zmud, R. W., & Armenakis, A. A. (1978). Understanding the measurement of change. Academy of Management Review, 3, 661-669.