1-1-2014

Validation of the Scores of the Instructional Pedagogical and Instructional Student Engagement Components of Fidelity of Implementation

Sandra F. Naoom
University of South Florida, sandranaoom@gmail.com

Validation of the Scores of the Instructional Pedagogical and Instructional Student Engagement Components of Fidelity of Implementation

by

Sandra F. Naoom

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Measurement, Research, and Evaluation
College of Education
University of South Florida

Major Professor: Robert F. Dedrick, Ph.D.
Lilliana Rodriguez-Campos, Ph.D.
John Ferron, Ph.D.
Sara Kiefer, Ph.D.

Date of Approval:
August 28, 2014

Keywords: fidelity, participant responsiveness, implementation, multi-level modeling

DEDICATION

First and foremost, I dedicate this dissertation to my Heavenly Father without whom I could not have gotten this far and who makes the impossible possible.

I give thanks to the beneficent and merciful God, the Father of our Lord, God and Savior Jesus Christ for He has covered me, helped me, guarded me, accepted me unto Him, spared me, supported me and brought me to this hour.

-Prayer of Thanksgiving, The Coptic Book of Hours

I also dedicate this dissertation to my family, many friends and church family who have encouraged and supported me throughout the process. There are no words I can say to express my gratitude for your prayers, support, understanding, and most of all, your love. I am especially grateful to my parents, Dina and Farouk Naoom and to my sisters Kirstie, Leslie, Sofia and my brother Samuel.

ACKNOWLEDGMENTS

This dissertation would never have come to fruition without the invaluable support and assistance of many individuals, and it is with great pleasure and gratitude that I acknowledge their efforts. I would like to express my deepest gratitude to Dr. Robert Dedrick, my program advisor, for the countless hours of guidance, support and feedback he provided me throughout this dissertation process. Dr. Rodriguez-Campos for her encouragement and support, Dr. Ferron for the gentle feedback he provides (that always comes by way of guided questions with answers), and to Dr. Keifer for her time and for chairing my committee. I would also like to thank Dean Fixsen and Karen Blase who not only encouraged me to pursue my doctoral degree, but also provided me the opportunity to focus on fidelity as I built my career in Implementation Science. I'd also like to thank my NIRN colleagues for their encouragement and support throughout. I am grateful to Amy Cassata, Dae Kim and Jeanne Century at CEMSE. They provided me with the data to complete this study and welcomed me as part of their team. Also, this work would not have been possible without the work of many others (cited in this dissertation) in the fields of implementation science, fidelity and evaluation. Without their seminal work on measurement of fidelity this study would not have been possible. Last but certainly not least, a special thanks goes out to a bright college student, Peter Mikhail, who helped me when no one else could, and spent an entire Sunday into the wee hours develop the models presented in this dissertation. Your kindness, time and support will not be forgotten. To all those who listened, supported and encouraged: Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Students cannot benefit from what they do not experience. Multiple reasons exist for why an intervention may not be delivered as it was designed. In this era of educational accountability and limited dollars to go around, understanding how an intervention is delivered in the classroom is key to understanding program outcomes. In order to assess whether a program has been implemented as intended, an assessment of fidelity is needed. However assessing fidelity is complex given varying conceptual interpretations, which then fosters inconsistent application of methods to measure the construct. Additionally the methods for validating fidelity measures are still unclear. The current study evaluated the reliability and validity of the student Instructional Pedagogical (10 items) and Instructional Student Engagement (15 items) scores for use in assessing teachers' fidelity of implementation on the participant responsiveness component of fidelity. The sample consisted of over 5,000 responses from students and 242 teachers in Mathematics and Science across three school districts and 41 schools to an online fidelity of implementation questionnaire. Given that students were nested within teachers, the data structure was multilevel, which warranted that the psychometric analyses be conducted using a multilevel framework. Instructional Pedagogy is represented by 10 items that measure three factors. Multilevel confirmatory factor analysis was used to test a two-level model that had three factors at the student-level and three factors at the teacher-level. Instructional Student Engagement is represented by 15 items that measure four factors.

Multilevel confirmatory factor analysis was used to test a two-level model that had four factors at the student-level and four factors at the teacher-level. The psychometric results of the student questionnaire assessing the student engagement components of fidelity were mixed. Support for the factorial validity of the multilevel student models was mixed, with model fit indicating that some of the measured variables did not load strongly on their respective factors and some of the factors lacked discriminant validity. Lastly, the correlations between students' and teachers' scores for both the observed and latent variables (ranging from -.15 to .72 in math; -.07 to .41 in science) displayed limited convergent validity

CHAPTER 1
THE PROBLEM AND ITS CLARIFYING COMPONENTS

"The bridge between a promising idea and the impact [on students] is implementation, but innovations are seldom implemented as intended" (Berman & McLaughlin, 1976, p. 349). In their 1976 report to Rand on the Implementation of Educational Innovations, Berman and McLaughlin analyzed the implementation of nationally disseminated educational innovations and found that there was a consistent lack of fidelity in the implementation of school programs. In order to produce behavior change, a program must be implemented as intended (Sanetti & Kratochwill, 2008). Programs consist of essential features that must be measured to determine whether a program is present or not (Century, Rudnick, & Freeman, 2010). Program fidelity refers to "the degree with which a particular program follows a program model…a well-defined set of prescribed interventions and procedures…types and amounts of services persons should receive, the manner in which services should be provided, and the administrative arrangements necessary to support service delivery" (Bond et al., 2000, p.1).

The failure to demonstrate fidelity is a methodological problem that has significant implications for internal and external validity, construct validity, and power. For internal validity, interpreting treatment outcomes is dependent in part on the strength of the evidence for fidelity. If the outcomes are positive, but fidelity was not assessed, the positive outcomes could be due to the intervention or possibly a range of other factors. In the same respect if the results are not significant and we had no information on fidelity it would be difficult to conclude if the intervention was ineffective or inadequately administered. The failure to implement the program

as planned or designed and to erroneously conclude that the observed findings are attributed to the intervention is referred to in the literature as a Type III error.

When interventions are adopted, fidelity measures can assist implementation and be used to monitor quality and performance, to ensure that the replications demonstrate fidelity to the model's critical components and are thereby likely to produce the intended outcomes (i.e., outcomes achieved in the original efficacy and effectiveness studies) (Bond et al., 2001). Fidelity measures can also promote external validity by providing adequate documentation and guidelines for replication. In order to replicate an intervention in a new setting, descriptions of the core components of the intervention and its implementation with fidelity are imperative.

To evaluate fidelity, the underlying core of the treatment intervention must be understood. Fidelity can be compromised by a deliverer's interpretation of the treatment protocol/intervention, as well as by confounding the intervention with other variables associated with the treatment. For example, if a deliverer does not understand the underlying theory of change for the intervention being put in place, the program deliverer may unknowingly omit key components of the intervention. Given that adaptation and program drift is common in non-research settings, fidelity measures provide methods to document deviations from an intended model and differences among the variations of a model (Mowbray, Holter, Teague, & Bybee, 2003).

Conceptualizing and operationalizing fidelity can be challenging. There is no singular agreement on how fidelity should be conceptualized or operationalized. Uniformity is lacking in the construct and definition of fidelity (Gearing et al., 2011). Some researchers view fidelity as unidimensional, while others see it as a multidimensional construct. Definitional inconsistency and varying conceptual interpretations undermine what constitutes the core components of

fidelity, and foster inconsistent application of methods to measure the construct (Gearing et al., 2011). Five aspects have been cited multiple times in the literature on the components that comprise fidelity (Dane & Schneider, 1998; Durlak & DuPre, 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003). These five components include the following:

- Adherence – program components are delivered as prescribed;
- Exposure – amount of program content received by participants;
- Quality of the delivery – theory-based ideal in terms of processes and content;
- Participant responsiveness – engagement of the participants; and
- Program differentiation – unique features of the intervention are distinguishable from other programs (including the counterfactual).

Even when fidelity has been conceptualized as a multidimensional construct, few studies assess more than a single dimension (Berkel, Mauricio, Schoenfelder, & Sandler, 2011). Typically the two dimensions measured most frequently are dosage and adherence (referred to as structural dimensions of fidelity), as they are more easily assessed than the interactional dimensions of quality and participant responsiveness. When fidelity is discussed in the literature it is not uncommon to hear the terms structural fidelity and procedural fidelity. Structural fidelity refers to the framework for service delivery and involves an objective look at whether important pieces of the intervention were delivered (e.g., program adherence; dosage as represented by time allocation and/or intervention completion). Procedural or process fidelity refers to the ways in which services are delivered. Process dimensions of fidelity are focused on assessing the quality of intervention delivery and/or the nature and quality of teacher-student interactions during intervention. Objectively establishing measurement reliability is more of a concern for procedural fidelity than for structural measures of fidelity. Rather than simply

determining if the intervention occurred or a component was delivered, procedural fidelity assessments must capture how well or to what degree the intervention or component was delivered (Harn, Parisi, & Stoolmiller, 2013). Harn et al. (2013) noted in their paper on fidelity, that it has been suggested by other researchers in the field (Gersten et al., 2005; Mowbray et al., 2003) that the process dimension is more directly relevant to student outcomes, even though it is more subjective and difficult to reliably measure. Given this, it is not sufficient to assess only the structural dimensions of fidelity and to leave the process and interactional dimensions of fidelity that examine the relationship between the deliverer and the recipient unmeasured. According to Zvoch (2012), 'in recent years ''treatment fidelity'' has developed as a multidimensional construct that reflects not only the degree to which providers deliver an intended treatment, program, or service, but also the extent to which targets receive and interact with treatment components' (p.548).

Similar to any measurement instrument, before it can be used successfully, the fidelity measure must be validated (Mowbray, Bybee, Holter, & Lewandowski, 2006). Over the last several years many researchers have identified critical steps in fidelity development and measurement (Bond et al., 2001; Century, et al, 2010; Mowbray et al., 2003; O'Donnell, 2008). Validation and the methods for validating fidelity measures are still unclear. For validation purposes, most studies have reported on inter-rater agreement (e.g., participant reports are compared to program developer reports) or the internal consistency reliability of scales, which are only pre-requisites to establishing validity and not validity itself (Mowbray et al., 2006). Factor analysis, another method used in the validation process in some areas of instrument development, is underutilized when assessing fidelity. The typical ways in which validity is assessed are: content or face validity, predictive validity, construct validity, and discriminant

validity. Most measures have content validity in that experts are typically used to develop the measures (nominate and select items, etc.). Predictive validity, the extent to which participants in high fidelity programs achieve significantly better outcomes than those in low fidelity outcomes, is also a common validation strategy. Calsyn (2000) identified content validity and predictive validity as methods that could be used to validate fidelity measures. Each of these methods can be problematic. When validating fidelity using a predictive validity method, consumer outcome measures are used, but fidelity can play a key role as a mediator or moderator variable in testing the effectiveness of a program model (Mowbray et al., 2006). Fidelity can play the role of mediator or moderator when using a predictive validity method because just by virtue of attending to fidelity, through the use of fidelity assessment where key components are highlighted and attended to, deliverers may implement with higher fidelity. This presents a confound, as using fidelity ratings as moderators or mediators assumes that we have a true and valid measure of adherence to a given program to the agreed on treatment practices. Discriminant validity is the ability to discriminate between those receiving the intervention and those receiving treatment as usual (by examining and comparing the fidelity scores of each). Mowbray et al. (2003) describe two promising methods for validating fidelity measures, noting:

> It seems desirable for validation purposes to examine fidelity measures for model replicas compared to other treatment programs serving the same populations and to test for significant differences [discriminant validity], or to examine convergent validity (information about a single program, but obtained from differing sources, such as records, client or key informant reports, site visits for certification purposes). (p. 332)

Discriminant validity can be limited though if the comparison programs adopt components of the intervention (contamination). In order to ensure that fidelity measures of effective interventions

are assessing the activities or components that they are intended to evaluate, fidelity measures must be validated and the methods used to validate the measures should be free from bias and confounding. Convergent validity is a validation method that may limit bias and confounding. Convergent validity involves examining the agreement between two different sources of information about the program and its operations (e.g., compare records and documents with on-site observations) and/or comparing the same measures of fidelity across diverse information sources (teachers, students, observers). The existing research literature on fidelity lacks studies of convergent validity and the feasibility of using consumers as an information source (Mook, 2010).

According to Mowbray et al. (2003), the most common methods to assess fidelity are: (1) ratings by experts, based on project documentation and/or client records, site observations, interviews, and/or videotaped sessions; and (2) surveys or interviews completed by individuals delivering the services or receiving them. When measuring a construct, validity is increased when multiple sources are used. There are many different sources that can be used to measure fidelity. Direct observation is the gold standard when it comes to methods to assess fidelity. Direct observation requires an operational definition of the intervention components, a record of the occurrence or nonoccurrence of each component, and a calculation of the percentage of treatment components (Sanetti & Kratochwill, 2008). Self-report, a more commonly used method for assessing fidelity requires the deliverer to record the level of fidelity subsequent to intervention implementation. Relying on the deliverer to accurately report activity (or lack thereof) may limit actual or perceived validity, through a social desirability bias, especially if staff suspect that the ratings may be a reflection of their performance.

Observation is thought to be more objective, valid and reliable than self-report (Rohrbach, Dent, Skara, Sun, & Sussman, 2007) but observation is costly and not always feasible, as observers need to be identified and trained. It has been suggested in the literature on fidelity that alternatives to observation and deliverer self-reports for assessing fidelity that are valid and feasible are needed (Berkel et al., 2011). Although deliverer self-reports are more feasible and less costly than observation, this method introduces self-report bias. An alternative method for assessing fidelity that has shown promise in the child and adult mental health field is using consumers of the intervention to assess fidelity (Lucca, 2000; Mook, 2010; Mowbray, et al., 2006). According to Mook (2010), who studied consumers' roles in rating the fidelity of a supported employment program, the four advantages to using a consumer (recipient) measure of fidelity is that the measure: (a) increases the consumers' role in research and program evaluation; (b) increases the validity of current methods for assessing fidelity; (c) expands fidelity measurement to include individual measures of fidelity, and (d) decreases the burden of current methods for assessing fidelity. Additionally, another advantage to using consumer self-reports of fidelity is that some information may not be attainable from anywhere else besides directly from the consumer (Baldwin, 2000) and that other sources of similar information may lack validity or add bias. Consumers are not going to know about all the activities going on in a program (Mowbray et al., 2006), in the same way that observers or delivers would, but when examining the process or interactional piece of fidelity --participant responsiveness and engagement-- consumers are likely to be the best source. Assessing participant responsiveness from the perspective of the participant may provide a more feasible, more objective, and less biased method of assessing fidelity when studying participant responsiveness, compared to observation and teacher self-report. When compared to other dimensions of fidelity, fewer studies have

assessed participant responsiveness, especially outside the confines of a research study. Given its limited use as a measure of fidelity, the need to attend to procedural fidelity, and the potential benefits (greater objectivity and feasibility), there is an emerging interest in assessing participant responsiveness from the consumer's perspective. Interest in including participant responsiveness in the assessment of fidelity is emerging; CEMSE's interest in developing and studying participant responsiveness measures in math and science education is an example of this emerging interest and the reason for the research.

Purpose of the Study

The University of Chicago's Center for Elementary Math and Science (CEMSE) team (with funding from the National Science Foundation) developed, piloted and field-tested eight instruments aimed at measuring the FOI of reform based K-8 science and mathematics instructional materials programs. This was done in recognition of the practical need for valid and reliable measures of fidelity of implementation of reform based STEM instructional materials and the theoretical need in the field for a shared conceptual framework for Fidelity of Implementation (FOI). The instruments, which provide a variety of data collection approaches, focus on clearly and specifically describing the nature of program implementation using constructs representing the essential elements of reform-based mathematics and science instructional materials programs organized into a conceptual framework. The conceptual framework supporting their instrument development efforts, the Fidelity of Implementation (FOI) Framework (Century, Freeman, & Rudnick, 2008), organizes program elements into two broad categories: Structural Critical Components and Instructional Critical Components. Then, each main category has subcategories that further classify the critical components.

In the Structural category, *procedural critical components* are the specific organizing structural elements of the program that focus on what the teacher needs to do; *educative critical components* represent the developer's expectations for what content and pedagogical knowledge the teacher needs to know and provide at a basic level to implement the program with fidelity.

In the Instructional category, *pedagogical critical components* reflect the developer's expectations about the behavior and interactions with students the teacher needs to enact in order to use the program as intended. Similarly, there are *student engagement critical components* that reflect the developers' expectations for student behaviors and interactions during instruction (e.g., teacher-student interactions, student-student interactions). Items for each of category by component can be found in Tables 8 and 9 in Chapter 3.

The larger University of Chicago's Center for Elementary Math and Science Education (CEMSE) study is looking at several dimensions of fidelity, but for the purpose of this dissertation study, the focus will be on participant responsiveness (i.e., student engagement). Participant responsiveness, an aspect or component of fidelity is defined as 'levels of participation and enthusiasm' (Dane & Schneider, 1998, p. 45). This is frequently measured or assessed by capturing the number of sessions attended by participants. Other measures include participant reports of satisfaction, and facilitator reports of participants' participation (Berkel et al., 2011). The justification for assessing participant responsiveness (i.e., student engagement) is that it is a component that is frequently not assessed when assessing fidelity, and when assessed, the way in which it is assessed does not capture the interactional relationship between the deliverer and the recipient. For the school-based study that provides the context for the assessment of fidelity, the treatment is a mathematics or science instructional intervention, delivered by teachers, with students as the recipients.

The student engagement measure of fidelity at the center of this dissertation was developed and refined by the University of Chicago. As a collaborator with the University of Chicago, I supported the development and refinement of the student fidelity instrument (details on the development procedures are provided in Chapter 3). CEMSE administered the student surveys and collected data in the fall of 2012, as part of their project scope. As part of this project, CEMSE also collected teacher-report data using the Teacher Instructional Questionnaire, Teacher Instructional Log, and Teacher Observation Protocol (these measures are described in Chapter 3). To accomplish the goal of validating the scores from the student fidelity questionnaire, the CEMSE team agreed to share their student and related teacher data with me for the purposes of this dissertation study. This was a secondary data analysis study using a quantitative research design to assess the reliability and validity of scores from the Fidelity of Implementation student questionnaire. This student questionnaire was administered to 3rd, 4th and 5th grade students and their teachers across 41 schools in three districts each located in different states (CO, CT, and IL). This assessment was conducted within the context of specific reform-based mathematics and science programs using four elementary-level curricula: Full Option Science System (FOSS), Science and Technology for Children (STC), Science Companion, and Everyday Mathematics (EM).

The objective of this analysis was to evaluate the reliability and validity of the scores from these instruments as indicators of fidelity of implementation (by testing the a priori models). The focus of this study will be on the Instructional Pedagogical (IP; e.g. teacher facilitation of student discussion, teacher facilitation of student interest) and Instructional Student Engagement (ISE; e.g. students engage in discussion, students demonstrate autonomy) components of Fidelity of Implementation (FOI) that are specific to the participant

responsiveness aspects of assessing fidelity. The convergent validity method will be used to examine the relationship between two different sources of information about a program and its operations (i.e., teacher and student reports). A visual of the models to be fitted and validated can be found in Chapter 4.

Research Questions

Building upon the work of University of Chicago's Center for Elementary Math and Science Education (CEMSE), and using the data collected by CEMSE in their administration of the teacher and student instruments, this study will focus on answering the following research questions:

1. What is the internal consistency reliability of the scores for the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) components?
2. Do individual items provide valid measures for the two FOI subcategories being examined in the Student Questionnaire, Instructional Pedagogical (IP), and Instructional Student Engagement (ISE)?
3. What is the convergent validity of the scores from the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) scales in mathematics and in science when measured by teacher- and student-reports?

Instructional Pedagogy:

- Teacher facilitation of student discussion (IP2)
- Teacher facilitation of student interest (IP7)
- Teacher use of differentiation (IP10)

Instructional Student Engagement:

- Students Contribute to Small Group Work (ISE1)
- Students Engage in Discussion (ISE2)
- Students Engage in Cog Demanding Work (ISE3)
- Students Take Risks (ISE4)

Significance of the Study

Students cannot benefit from what they do not experience. There are multiple reasons why an intervention may not be delivered in its entirety. For example, it would be impossible to determine if an intervention designed to improve student outcomes in math failed because it was ill conceived and based on a faulty model, or if it failed because the theory was sound but the intervention was implemented poorly. In order to assess whether a program has been implemented as intended, an assessment of fidelity is needed. As with all measures, an evaluation of their psychometric quality is also necessary. The current proposal extends prior research on fidelity assessment by studying the participant responsiveness dimension of fidelity from the perspective of consumers (students) and further by validating this student fidelity measure using a convergent validity approach.

Delimitations and Limitations

This study is delimited to elementary science and mathematics education. The student sample consisted of 3rd through 5th graders enrolled in the participating schools as of the fall of 2012, who had parental permission, and who themselves assented to participate in the research project. Each student completed a science questionnaire and a mathematics questionnaire, with the timing of questionnaire completion corresponding to the teacher's completion of each teacher

implementation questionnaire (TIQ). Students in K-2 grades were not included in the student sample because pre-literate children may not possess the linguistic or information processing skills to articulate differentiated self-beliefs within a particular domain such as a science (Mantzicopoulos, Patrick, & Samarapungavan, 2008). Also, graded response formats may tax the cognitive processing abilities of young children, who tend to respond at the extreme points of rating scales, particularly with items referring to social situations and psychological states (Chambers & Johnston, 2002).

This is a secondary data analysis study. There are limitations to secondary data analysis in that the researcher has no control over the purpose, research design, choice of methods of data collection, sampling methods and populations studied, and variables included in the study. It should be noted, however, that as part of this dissertation project, I participated in the development and refinement of the student questionnaire.

Another limitation of the study is that the data to be reviewed and analyzed for this study are student self-report data and teacher self-report data collected using a common method (online questionnaire). The use of use of self-reported data assumes honest reporting but there may be some error introduced into the data as a result of method effects (e.g., social desirability). As is common with most self-report questionnaires there was some missing data and participants who did not complete the questionnaires. These issues will be addressed in Chapter 3.

The data collected and analyzed for this dissertation project are cross-sectional. There are limitations to using cross sectional data in that these data cannot be used to infer causation, and since the data are collected at one moment in time, cannot capture change like in a longitudinal study.

Definitions of Terms

Adaptation: "Substantive deletions and enhancements, as well as changes to the manner or intensity with which a program is delivered" (Ringwalt et al., 2003, p. 376).

Adherence: Program components are delivered as prescribed (Dusenbury et al., 2003).

Program Differentiation: Identification of the unique components of a program so that the components of the program can be differentiated from one another (Dusenbury et al., 2003).

Program Fidelity: Also referred to as program integrity and treatment integrity. Program fidelity refers to "the degree with which a particular program follows a program model…a well-defined set of prescribed interventions and procedures…types and amounts of services persons should receive, the manner in which services should be provided, and the administrative arrangements necessary to support service delivery" (Bond et al., 2000, p.1).

Dosage: Also referred to as exposure. The amount of program content received by participants (Dusenbury et al., 2003).

Structure: Structural fidelity refers to the framework for service delivery and involves an objective look at whether important pieces of the intervention were delivered (Mowbray et al., 2003).

Participant Responsiveness: Engagement of the participants (Dusenbury et al., 2003).

Procedural or Process Fidelity: Refers to the ways in which services are delivered (Mowbray et al., 2003).

Quality of the Delivery: Theory-based ideal in terms of processes and content (Dusenbury et al., 2003).

Summary of Chapter

This chapter provided the background to the study including a definition of fidelity and its importance and issues related to measuring and validating fidelity. Chapter 2 presents a review of the literature on fidelity and validation. Chapter 3 describes the context for this study, study participants, data collection procedures, and proposed analyses for this study. Chapter 4 will highlight the results found in this study and Chapter 5 will cover conclusions and contributions to the field.

CHAPTER 2
REVIEW OF THE LITERATURE

Today, in an era of accountability, the call for measuring fidelity of K-12 interventions during efficacy and effectiveness trials is receiving increased attention (O'Donnell, 2008). "Educators trying to make choices to help students and schools meet high standards can become overwhelmed by the amount of education research. It can also be hard to identify research with credible and reliable evidence to use in making informed decisions. As an initiative of the U.S. Department of Education's Institute of Education Sciences (IES), the What Works Clearinghouse (WWC) was created in 2002 to be a central and trusted source of scientific evidence for what works in education" (Institute of Education Sciences, 2014, What Works Clearinghouse: About Us). Registries like education's What Works Clearinghouse exist in multiple fields (Blueprints for Prevention, National Registry for Effective Programs and Practices, etc.) and all were developed with the intention of identifying and getting evidence-based practices into the hands of implementers. With this focus on implementing evidence based practices, the technical and methodological demands on researchers have increased, in that the emphasis on evidence-based practice has made it clear that evidence of effectiveness must be accompanied with clear evidence of what produced the effects. The development and use of valid fidelity measures is now an expected component of quality evaluation practice (Vartuli & Rohs, 2009). Only by understanding and measuring whether an intervention has been implemented with fidelity can researchers and practitioners gain a better understanding of how and why an intervention works, and the extent to which outcomes can be improved. Unless such an evaluation is made, it cannot

be determined whether a lack of impact is due to poor implementation or inadequacies inherent in the program itself (Carroll et al., 2007). This chapter has been divided into five sections aimed at moving the reader from understanding what fidelity is and why it is necessary to assess fidelity, to how it has been operationalized, assessed, and validated in the field.

Fidelity Defined

Programs consist of essential features that must be measured to determine whether a program is present or not (Century, Rudnick, & Freeman, 2010). Program fidelity refers to "the degree with which a particular program follows a program model…a well-defined set of prescribed interventions and procedures…types and amounts of services persons should receive, the manner in which services should be provided, and the administrative arrangements necessary to support service delivery" (Bond et al., 2000, p. 1). The concept of fidelity has been around for some time now. Blakely et al., in their 1987 publication on the topic of fidelity, cited unpublished pioneering work by Hall, which described social programs as consisting of a finite number of components and fidelity as the proportion of program components that were implemented (O'Donnell, 2008). Interest though in measuring fidelity began to increase in the 1970s. According to Dusenbury et al. (2003), "in the 1960s and 1970s the Research, Development and Diffusion (RD & D) model, inspired by the space program, emphasized the importance of rigorous evaluation and validation in demonstration projects" (p. 238). An assumption of this model was that consumers would value the results of these evaluation studies and base their program adoption decisions on the results of these evaluation studies. As consumers were viewed as passive actors in this model, the expectation was that consumers would implement programs as intended by the program developers (Rogers, 1995). The assumptions of this model were called into question beginning in the mid-to-late 1970s with a

study conducted by the Rand Institute. A Rand report on the Implementation of Educational Innovations analyzed the implementation of nationally disseminated educational innovations and found that there was a consistent lack of fidelity in the implementation of school programs.

This report and other studies underscored the importance of assessing intervention (program) fidelity, arguing that to produce behavior change, an intervention must be implemented as intended (with fidelity); without a formative assessment of fidelity there is no way to determine whether unsuccessful outcomes reflect a failure of the model or failure to implement the model as intended (Type III error). "Fidelity is important because we typically do not know which components of a program may be responsible for the positive outcomes. Therefore, the belief that some intervention is better than none may be erroneous" (Mihalic, 2004, p. 83).

Why Assess Fidelity?

Without specific criteria governing program implementation, an innovation or evidence-based program can revert back to the status-quo in replications, and thus fidelity measures can be used as a guide to implement an intervention as intended or for monitoring programs for quality (Mowbray et al., 2003). Researchers have highlighted several important purposes for collecting fidelity data (Backer, 2001; Dane & Schneider, 1998; Domitrovich & Greenberg, 2000; Mowbray et al., 2003; Pankratz et al., 2006) among them are understanding program implementation, examining theoretical assumptions, interpreting outcome findings, providing feedback for continuous quality improvement, and providing feedback to program developers about the program (James Bell Associates, 2009).

In addition, fidelity affects all the major threats to validity described by Cook and Campbell in 1979. The failure to demonstrate fidelity is a methodological problem that has

significant implications for internal and external validity, construct validity, and power. Fidelity measures are the tools used to assess the adequacy with which a program was delivered and implemented. When interventions are adopted, fidelity measures can assist implementation and be used to monitor quality and performance, to ensure that the replications demonstrate fidelity to the model's critical components and are thereby likely to produce the intended outcomes (i.e., outcomes achieved in the original efficacy and effectiveness studies) (Bond et al., 2001). Fidelity measures can also promote external validity by providing adequate documentation and guidelines for replication. In order to replicate an intervention in a new setting, descriptions of the core components of the intervention and its implementation with fidelity are imperative. In multi-site studies, fidelity measures are critical to ensuring that the services studied across sites are the same, or, if there are differences, those differences are documented and measured (Paulson, Post, Henricks, & Risser, 2002). Program developers have noted that when key elements are left out of replications, intended outcomes are not achieved (Bond et al., 2000). In terms of construct validity, because fidelity measures are derived from theory, by definition they are relevant to construct validity (Calsyn, 2000). To evaluate fidelity the underlying core of an intervention must be understood. Fidelity can be compromised by the practitioner's understanding of the treatment protocol, as well as by confounding of the independent variable with other variables associated with treatment. The use of fidelity measures to identify the core components of an intervention is an example of how fidelity studies approach construct validity (Bond et al., 2000). In a research setting, well-developed and valid measures can increase statistical power in treatment outcome studies by acting as moderating variables to help explain variance in outcomes (Teague, Drake & Ackerson, 1995).

Fidelity plays an important role in outcome effectiveness (Blakely, 1987; Dane & Schneider, 1998; Dusenbury et al., 2003; Hansen, Graham, Wolkenstein, & Rohrbach, 1991; Lipsey, 1995; Pentz et al., 1990). Studies across numerous fields and disciplines have demonstrated that the fidelity with which an intervention is implemented affects how well it succeeds (Abbott et al., 1998; Burke et al., 2011; Dane & Schneider, 1998; Durlak & DuPre, 2008; Dusenbury et al., 2003; Elliot & Mihalic, 2004; Mihalic, Fagan, Irwin, Ballard & Elliott, 2002; Mihalic, 2004) while poor fidelity is associated with reduced program effects (Pentz et al., 1990). There is strong evidence that fidelity levels are significantly related to the amount of positive change achieved by a program. For example, in a review of over 500 studies, Durlak and DuPre (2008) found that mean effect sizes were at least two to three times higher when programs were implemented with high levels of fidelity, especially in terms of adherence and exposure. In addition, studies that incorporate implementation data into outcome analyses often find stronger effect sizes than analyses conducted without these data (Dane & Schneider, 1998; Dusenbury et al., 2003; Harachi, Abbott, Catalano, Haggerty, & Fleming, 1999; Lillehoj, Griffin & Spoth, 2004). For instance, in a study of a parent training program, it was found that when the program was implemented with high fidelity, as assessed by the Fidelity of Implementation Rating System (FIMP), an observation-based measure assessing competent adherence to the Oregon model of Parent Management Training, parenting practices improved significantly, but the effect was much less when implementation fidelity was low (Forgatch, Patterson, & DeGarmo, 2005). In another study focused on assessing fidelity of multi-component family support programs for improving educational outcomes for at risk youth, strong positive relationships were found between overall program fidelity and program-level outcomes achieved by student participants (Kalafat, Illback, & Sanders, 2007). In two studies examining programs

to help people with mental health issues obtain employment, it was found that employment outcomes were weakest for those in programs where fidelity was poor (McGrew & Griss, 2005; Resnick, Neale, & Rosenheck, 2003).

Conceptualization/Operationalization

Conceptualizing and operationalizing (i.e., measuring) fidelity can be challenging. There is no singular agreement on how fidelity should be conceptualized or operationalized. Uniformity is lacking in the construct and definition of fidelity (Gearing et al., 2011). Some researchers view fidelity as unidimensional, while others see it as a multidimensional construct. Definitional inconsistency and varying conceptual interpretations undermine what constitutes the core components of fidelity, and foster inconsistent application of methods to measure the construct (Gearing et al., 2011). Five aspects have been cited multiple times in the literature on the components that comprise fidelity (Berkel et al., 2011, Dane & Schneider 1998; Durlak & DuPre, 2008; Dusenbury et al., 2003; Giles et al., 2008; Hill et al., 2013). These five components are:

- Adherence – program components are delivered as prescribed;
- Exposure – amount of program content received by participants;
- Quality of the delivery – theory-based ideal in terms of processes and content;
- Participant responsiveness – engagement of the participants; and
- Program differentiation – unique features of the intervention are distinguishable from other programs (including the counterfactual).

In modern conceptualizations of fidelity, variation in intervention receipt and intervention delivery matters, as an intervention can be delivered with a high degree of skill and integrity but

participants still may not receive or interact with the intervention as intended. Receipt and delivery breakdowns occur when participants are not engaged during treatment delivery, fail to comprehend or follow through on intervention protocols, and/or intermittently attend sessions (Zvoch, 2012).

Adherence can be defined as the "extent to which implementation of particular activities and methods is consistent with the way the program is written" (Dusenbury, 2003, p. 241). Fidelity is sometimes defined as adherence (Blakely et al., 1987). Programs and intervention can consist of essential and non-essential elements; a critical first step to assessing fidelity is to identify those elements that are critical to the program (McGrew, Bond, Dietzen, & Salyers, 1994). These critical elements can then be used to measure adherence to the intervention. For example, Heck, Stieglebauer, Hall, and Loucks (1981) conceptualized social programs as consisting of a number or proportion of program components to be implemented (Hall and Louck also developed a method for identifying and classifying program components). Fidelity or adherence could then be assessed at implementing sites by determining the number or proportion of program components implemented. In a study by Dusenbury et al. (2005) adherence to a school based prevention program was assessed through classroom observations. Six items measured adherence and observers coded the number of objectives and, separately, major points completed by teachers. Full points were awarded when objectives or major points were met and half points were awarded when these were partially met. Observers also provided a summary judgment about the proportion of objectives and major points that were covered. In a study by Skara, Rohrbach, Sun and Sussman (2005), adherence to program objectives was measured by teacher self-report. Teachers were asked how much they adhered to the lesson plan (1 = not at

all, 7 = great deal) and how difficult was it to teach the program (1 = not at all difficult, 7 = very difficult).

Dosage or exposure is defined as the amount of a program delivered to a target audience or the amount of program material received by participants. Dosage may provide important information about fidelity when a program is delivered in the real world, such as in a classroom where dosage may vary based on length of classroom sessions, competing demands on students and teachers, etc., as opposed to a controlled research setting where dosage is more likely to be high (Dusenbury et al., 2003). According to Botvin, Baker, Dusenbury, Botvin, and Diaz (1990) dosage or how much of the material is delivered (Botvin calls this 'completeness of program implementation') is an important aspect of fidelity, as students who received more of the prevention program showed greater change in behavioral outcomes, demonstrating the importance of measuring implementation fidelity. In a study by Pentz et al. (1990), quality of prevention program implementation, as measured by the amount of implementation or program exposure, was shown to prevent increases in drug use and had a significant effect on changing adolescent drug use behavior. In a study on case management, three hypotheses were tested, one of which was higher fidelity of case management implementation predicts a lower probability of dropping out of substance abuse treatment. It was found that as fidelity increased, the risk of dropping out of substance abuse treatment decreased.

> Only fidelity of case management implementation and proportion of total case management time spent on case management core functions (i.e., outreach, assessment, service planning and resource identification, linking clients to services, service coordination, monitoring service delivery, and advocacy) had a statistically significant impact on attrition. With each unit increase in the case management fidelity score, the

risk for dropping out of substance abuse treatment decreased by 21%. (Noel, 2006, p. 322)

Dosage or exposure is commonly assessed through the use of trained observers using observation monitoring forms to determine the proportion of objectives covered out of the total objectives per session. Additionally, provider self-report, as well as attendance data can measure dosage for participants (Botvin, Griffin, Diaz, & Ifill-Williams, 2001; Botvin et al., 1990; Dusenbury et al., 2003).

Hansen et al. (1991) hypothesized that program integrity, the quality of program delivery, is a variable that may moderate program effectiveness. According to the authors, the two components that contribute to program integrity are the variability of quality of program delivery and the reception of the program by its target audience. Program integrity is an important construct, as one would expect a poorly implemented and poorly received program to be less effective than the same program when it is implemented with fidelity. The authors argue that without evaluating program integrity it is difficult to know whether or why a program has succeeded or failed. In a study by Dusenbury et al. (2005), quality of program delivery of a school based prevention program was assessed through classroom observations. Observers rated how well lessons were delivered and received. Ratings were obtained for: (a) teacher-student interactivity, (b) teacher enthusiasm, (c) teachers' communication of goals and objectives, (d) student engagement, (e) student attentiveness, and (f) students expressing their opinions.

Adaptation could be defined simply as the opposite of fidelity. For a more technical definition, adaptation refers to "substantive deletions and enhancements, as well as changes to the manner or intensity with which a program is delivered" (Ringwalt et al., 2003, p. 376). There is great debate in the literature about the appropriateness of adaptation (also called re-invention

in the literature). Everett Rogers (1995) in his work on diffusion of innovations posited that significant adaptation or reinvention of programs is necessary to preserve program effectiveness. Rogers argued that implementation problems individuals or organizations may face are unpredictable, so changes to the innovation should often occur, and that adaptation (re-invention) of an innovation may instead reduce mistakes and encourage customization of the innovation to fit with local or changing conditions. Others have argued that any departure from exact replication dilutes the effects of the program and will not produce the promised outcomes. In a study by Kelly et al. (2000) on the transfer of HIV prevention interventions to community service providers the authors suggest that although adaptation may be necessary to better meet the needs of consumers, communities, or organizations "the core elements of the intervention cannot be changed without fundamentally changing the intervention" (p. 1087). Regardless of which camp researchers side with, measuring adaptation as a part of fidelity is necessary. Identifying whether practitioners have made additions or modifications to the content or the delivery of a program helps us to understand the degree to which the program or innovation has been implemented with fidelity, as well as whether outcomes, whether they be positive or negative, are associated with the program under study.

In a study by Dusenbury et al. (2005), adaptation of a school based prevention program was assessed through observation and through interviews with teachers. Observers noted how content and activities were altered from those outlined in the manual, and then rated whether these were consistent with or detracted from the program's objectives. The scale ranged from -2 to +2 with negative scores representing detracting adaptations and positive scores representing enhancing adaptations. The number of and average valence of adaptations were calculated. In

interviews, teachers were asked whether and how they had altered the program when they taught it. Researchers counted how many adaptations teachers reported.

Participant responsiveness can be defined as the extent to which participants are engaged by and involved in the content and activities of the program delivered (Dusenbury et al., 2003). In a study by Hawkins et al. (1991), students were asked about intervention components that should have been delivered by their teachers. To evaluate student responsiveness to a drug-abuse prevention program, students were asked to rate how much they liked each program session. Additionally, students were instructed to take a minute to think about the drug prevention program, about the topic and activities completed each day, to form a general opinion about the program overall, considering all 12 sessions, and then rate the program on 12 adjectives (Skara et al., 2005).

Program differentiation is the identification of the unique components of a program so that the components of the program can be differentiated from one another (Dusenbury et al., 2003). "The measurement of program differentiation is essential in assessing aspects of fidelity that are related to immediate outcomes. Program differentiation helps to evaluate the essential elements of effective programs (i.e., component analysis) because it allows for determination of whether each component of the program changed its respective targeted immediate outcomes" (Skara et al., 2005, p. 308). There are few measures of program differentiation in the literature on fidelity. In a study of a school-based, drug-abuse prevention program, differentiation was evaluated by assessing student knowledge of curriculum content (Skara et al., 2005).

Even when fidelity has been conceptualized as a multidimensional construct, few studies assess more than a single dimension (Berkel et al., 2011). Typically the two dimensions measured most frequently are dosage and adherence. In a review of the implementation of 34

programs determined to be effective in a review conducted by the Prevention Research Center for the Center for Mental Health Services (Domitrovich & Greenberg, 2000), the authors looked at the presence or absence of five of the factors described above. The authors found that adherence and dosage were the two aspects of implementation that were monitored most often. Twenty programs (59%) included some rating of adherence in their implementation data; of these, the majority tracked the program's essential components with ratings made by independent observers or the program implementers. Dosage was reported in 33% of the studies. Four programs (12%) assessed participant responsiveness, and two programs (6%) assessed program differentiation. Interestingly, only 11 of the 34 studies (32%) utilized implementation information as a source of data for outcome analyses. In some cases, descriptive statistics were conducted on the implementation information but the data were not related to program outcomes. Four studies examined dosage-response relationships and results indicated that higher quantities of the intervention were related to better outcomes. Seven studies used adherence ratings to examine whether quality of implementation was related to outcomes. When significant results were found, higher fidelity was related to stronger program outcomes.

Measuring Fidelity

Intervention fidelity is a central issue across many fields and disciplines, but in view of the need for accountability it has taken on added importance in the field of education, particularly in the reform efforts in the science, technology, engineering, and mathematics (STEM) areas. Careful description and measurement of fidelity are necessary to understanding which components of reform based mathematics and science programs bolster or hinder student performance, or to determine the differential effects of incomplete or incorrect implementation of instructional materials (Fullan, 2001; Lynch & O'Donnell, 2005; Ruiz-Primo, 2005).

Historically though, very few studies have published results of fidelity, not just in education, but also across various fields and disciplines. Across fields and disciplines, it is not uncommon to find that less than one-third of treatment effectiveness studies report evidence of intervention fidelity. Durlak reported that out of 1200 studies, only 5% addressed fidelity, and of 181 studies in special education, 14% addressed fidelity (Durlak & DuPre, 2008) and Dane and Schneider reported 17% in the 1980s, but 31% in the 1990s. Even within these studies, the models of fidelity and methods used to assess or assure fidelity differed greatly. In a review of K-12 core curriculum, it was found that there were insufficient studies to guide researchers on how fidelity to core curriculum intervention should be measured and that very few early childhood studies provided assurances of fidelity, which makes implications for practice questionable (O'Donnell, 2008). Most of the studies of fidelity have been in the areas of mental health programs, public health programs, and supplements to K-12 education such as prevention programs, but there have been few studies about core elementary and secondary school subjects published (Bond et al., 2000; Resnick et al., 2005; Vartuli & Rohs, 2009). Fidelity measures are less developed and under theorized in effectiveness studies of curriculum and instruction where results are based on student learning of discipline context. Although understanding and measuring fidelity is of importance to education researchers, it is of increasing importance to practitioners in school systems as they try to understand the effectiveness of interventions initiated in response to federal pressures to improve student performance (Lynch, 2007).

Prior to measuring fidelity, the critical components of an intervention or program must be specified, operationalized, and validated. The determination of whether instructional materials have been adequately and faithfully implemented necessitates reliable and valid indicators of the extent, quality, and type of the implementation of the materials (NRC, 2004). Mowbray et al.

(2006) related that there are three steps to establishing fidelity criteria: identify possible indicators or critical components of a given model, collect data to measure indicators, and examine the indicators in terms of their reliability and validity.

In the first step, fidelity criteria must be identified. Fidelity criteria should reflect the core components of a program, sometimes referred to as the active or essential ingredients. Structural fidelity criteria refer to service delivery and involve an objective look at whether important pieces of the intervention were delivered (e.g., program adherence, dosage represented by time allocation and/or intervention completion). Procedural or process fidelity refers to the ways in which services are delivered. Process dimensions of fidelity are focused on assessing the quality of intervention delivery and/or the nature and quality of teacher-student interactions during intervention, as well as the values, principles, and climate of an implementing organization (Mowbray et al., 2003). Fidelity criteria can be developed by multiple methods. Some examples are conducting a components analysis, an analysis done to determine which program components are essential; gathering expert opinion through surveys of experts or literature reviews; reviewing program materials, such as curricula and training guides; and drawing from a program's logic model to understand the theory of change.

Following the identification of fidelity criteria, measurement tools must be identified and developed. In terms of measuring fidelity, detailed descriptions for how fidelity was measured are often included in the literature. The two most common methods to assess fidelity are: (1) ratings by experts (based on documentation and/or client records, site observations, interviews, and/or videotaped sessions); and (2) surveys or interviews completed by individuals delivering the services or receiving them (Mowbray et al., 2003). When measuring a construct, validity is

increased when multiple sources are used. There are many different sources that can be used to measure fidelity.

One of the most common methods for assessing fidelity involves the completion of an implementation checklist, log, or survey by program service providers James Bell Associates, 2009). For example, in a study by Mills and Ragan (2000), teachers who used the intervention under study completed checklists (based on fidelity criteria). Self-report requires the deliverer to record the level of fidelity subsequent to intervention implementation. Relying on the deliverer to accurately report activity (or lack thereof) may limit actual or perceived validity, through a social desirability bias, especially if staff suspect that the ratings may be a reflection of their performance. There is a significant potential for positivity bias among teachers (Lillehoj et al., 2004), which may be related to concerns that fidelity data might be used to evaluate performance (Donaldson and Grant-Vallone, 2002).

Direct observation is the gold standard when it comes to methods to assess fidelity. Direct observation requires an operational definition of the intervention components, a record of the occurrence or nonoccurrence of each component, and a calculation of the percentage of treatment components (Sanetti & Kratochwill, 2008). Observation can be made either in person or by watching videotapes of program activities being implemented James Bell Associates, 2009). For example, in a study by Clarke (1995), the rating of fidelity to a Family-Focused Treatment (FFT) model involved three experts trained in FFT. The experts utilized the Therapist Competence/Adherence Scale to evaluate the videotape of the first family session in each segment of treatment. Observation is thought to be more objective, valid, and reliable than self-report (Rohrbach et al., 2007) and observational data are more strongly correlated to program outcomes than self-report data (e.g., Dane & Schneider, 1998). Observation is the gold standard

for assessing fidelity but observation is costly and not always feasible, as observers need to be identified and trained. Although deliverer self-reports are more feasible and less costly than observation, this method introduces self-report bias. It has been suggested in the literature on fidelity that alternatives to observation and deliverer self-reports for assessing fidelity that are valid and feasible are needed (Berkel et al., 2011).

Review of archival and administrative data, such as attendance records, case records, and training manuals are also useful for assessing fidelity (James Bell Associates, 2009). Data review is typically done to complement another method of fidelity assessment. For example, in a study by Hernandez et al. (2001) on the implementation of Systems of Care, as part of the fidelity assessment record-keeping instruments, reviews of treatment plans and individualized educational plans from case records were reviewed, in addition to provider and participant interviews.

An alternative method for assessing fidelity that has shown promise in the child and adult mental health field is using consumers of the intervention to assess fidelity (Lucca, 2000; Mook, 2010; Mowbray et al., 2006). For example, in the Henggeler, Schoenwald, Liao, Letourneau, and Edwards (2002) study of fidelity, a Therapist Adherence Measure was administered to families receiving multi-systemic therapy (MST) services, after the start of treatment and monthly thereafter, through phone interviews by an MST employee other than the family therapist. According to Mook (2010), who studied consumers' roles in rating the fidelity of a supported employment program, the four advantages to using a consumer (recipient) measure of fidelity is that the measure: (a) increases the consumers' role in research and program evaluation; (b) increases the validity of current methods for assessing fidelity; (c) expands fidelity measurement to include individual measures of fidelity, and (d) decreases the burden of current

methods for assessing fidelity. Additionally, another advantage to using consumer self-reports of fidelity is that some information may not be attainable from anywhere else besides directly from the consumer (Baldwin, 2000) and that other sources of similar information may lack validity or add bias. Consumers are not going to know about all the activities going on in a program (Mowbray et al., 2006), in the same way that observers or delivers would, but when examining the process or interactional piece of fidelity -- participant responsiveness and engagement-- consumers are likely to be the best source. Assessing participant responsiveness from the perspective of the participant may provide a more feasible, more objective, and less biased method of assessing fidelity when studying participant responsiveness, compared to observation and teacher self-report. When compared to other dimensions of fidelity, fewer studies have assessed participant responsiveness, especially outside the confines of a research study. Given its limited use as a measure of fidelity, the need to attend to procedural fidelity, and the potential benefits (greater objectivity and feasibility), there is an emerging interest in assessing participant responsiveness from the consumer's perspective.

There exists a diversity of methods and sources for assessing fidelity, and using multiple methods and multiple sources to establish fidelity is a recommended practice. The methods presented, however, do present several issues that are not unlike those found in other fields of research. Relying on practitioners and staff to accurately report their activity or lack thereof may limit actual or perceived validity through a social desirability bias. Many agree, "direct observation is the most accurate assessment and that self-monitoring reports often produce inflated estimates of levels of performance relative to direct observation" (Vartuli & Rohs, 2009, p. 505). The use of experts or supervisors to assess fidelity through observation or other methods may also pose validity issues because they are not blind to the program they are rating. With

consumer ratings you have to consider whether those who elected to participate differ from those who did not and research has shown that volunteer participants may be overly positive or overly negative. If consumers or program users are more active stakeholders in assessing fidelity, indicators of critical processes may more effectively complement the indicators of structural features that consumers can more expertly assess (e.g., asking consumers to report on services they receive, such as, asking a student whether or not a teacher delivered a program component). All these issue can be lessened though when the fidelity scale uses objective, behaviorally anchored criteria for each scale point, involving little rater inference. The use of multiple sources and methods can also serve to increase reliability and validity in fidelity ratings (Emshoff et al., 1987; Ruiz-Primo, 2005; Summerfelt, 2003; Vartuli & Rohs, 2009; Zvoch, Letourneau, & Parker, 2007).

In the third and final step of establishing fidelity criteria, reliability and validity must be assessed. Similar to any measurement instrument, before it can be used successfully, the fidelity measure must be validated (Mowbray et al., 2006) and the methods used to validate the measures should be free from bias and confounding.

Measurement Quality

According to the *Standards for Educational and Psychological Testing* (1999) 'validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed use of the tests' (p.9). Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validating involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretation of scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated to the

proposed use. Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the construct or concepts the test is intended to measure. Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use.

Cronbach (1971) described validation as the process by which a test developer or test user collects evidence to support the types of inferences that are to be drawn from test scores. To plan a validation study, the desired inference must be clearly identified. Then an empirical study is designed to gather evidence of the usefulness of the scores for such inferences. Three major types of validation studies are:

- Content validation for situations where the test user desires to draw an inference from the examinee's test score to a larger domain of items similar to those on the test itself;
- Criterion-related validation for situations where the test user desires to draw an inference from the examinee's test score to performance on some real behavioral variable of practical importance; and
- Construct validation for situations where no criterion is accepted as entirely adequate to define the quality to be measured (Cronbach & Meehl, 1955), but the test user desires to draw an inference from the test score to performances that can be grouped under the label of a particular psychological construct.

Validating Fidelity Measures

Over the last several years many researchers have identified critical steps in fidelity development and measurement (Bond et al., 2000; Century, Rudnick, & Freeman, 2010; Mowbray et al., 2003; O'Donnell, 2008), but validation and the methods for validating fidelity

measures are still unclear. According to Mowbray et al. (2003) there are five different approaches that have been used to assess reliability and validity in the literature on fidelity. In terms of assessing reliability, reliability has been assessed across respondents, calculating the extent of inter-rater agreement thru coefficient kappa, intra-class correlations (ICC), percent agreement, or Pearson correlations (Clarke, 1995; Henggeler et al., 2002; Weisman et al., 1998). Reliability has also been assessed using measures of internal consistency reliability (e.g., Cronbach's alpha). The second approach, which focuses more on validity, has involved examining the internal structure of the data empirically and in relationship to expected results, such as through exploratory and confirmatory factor analysis (Henggeler et al., 2002), or cluster analysis (Mills & Ragan, 2000). The third approach is the method of known groups where one examines differences in fidelity scores across programs that are expected to be different (Bond et al., 2000; Hernandez et al., 2001; Lucca, 2000; Teague et al., 1995). Typically this involves a comparison of the new intervention compared to traditional or treatment as usual. Convergent validity is the fourth approach to validation. In convergent validity the focus is on examining the extent of agreement between two different sources of information about the program and its operations. For example, Blakely et al. (1987) compared records and documents with on-site observations and Macias, Propst, Rodican, and Boyd (2001) examined self-ratings of compliance with clubhouse standards on the Clubhouse Research and Evaluation Screening Survey (CRESS) to the results from on-site, extensive certification procedures, comparing CRESS scores of certified to non-certified agencies. McGrew, Pescosolido, and Wright (2003) sought additional validation of Assertive Community Treatment (ACT) criteria by surveying ACT team members as to the extent to which they considered the critical activities involved to be of benefit. Lucca (2000) examined the correlation between Clubhouse fidelity index scores and scores on a

Principles of Psychosocial Rehabilitation scale, to address convergent validity (Mowbray et al., 2003). The final approach, examining the relationship between fidelity measures and participant outcomes, is probably the most commonly used validation approach in research on interventions, as researchers are interested in understanding whether the intervention was implemented as intended to achieve the desired outcomes.

For validation purposes, most studies have reported on inter-rater agreement or the internal consistency reliability of scales, which are only pre-requisites to establishing validity and not validity itself (Mowbray et al., 2006). Calsyn (2000) identified content validity and predictive validity as methods that could be used to validate fidelity measures. Most measures have content validity in that experts were used to develop the measures (nominate and select items, etc.). Content validity refers to how adequately the fidelity items cover fidelity. Predictive validity, the extent to which participants in high fidelity programs achieve significantly better outcomes than those in low fidelity programs, is also a common validation strategy. Each of these methods can be problematic. When validating fidelity using a predictive validity method, consumer outcome measures are used, but fidelity can play a key role as a mediator or moderator variable in testing the effectiveness of a program model (Mowbray et al., 2006). This presents a confounding as using fidelity ratings as moderators or mediators assumes that we have a true and valid measure of adherence to a given program to the agreed on treatment practices. With respect to construct validity, fidelity can be compromised by the practitioner's interpretation of the intervention, as well as by confounding of the independent variable with other variables associated with the intervention. Discriminant validity, which is recommended in the literature for further use, is the ability to discriminate between those receiving the intervention and those receiving treatment as usual (by examining and comparing the fidelity

scores of each). Discriminant validity can be limited though if the comparison programs adopt components of the intervention (contamination). Another, less often-used method of validation cited in the literature is concurrent validity. With concurrent validity, replications of a model are compared with programs that serve the same population but use distinctively different methods.

Convergent validity occurs when the fidelity measure being studied correlates highly with other fidelity measures of the same construct (Calsyn, 2000). This involves examining the agreement of fidelity scores between two different sources of information about the program and its operations (e.g., compare records and documents with on-site observations), as well as comparisons across diverse information sources (staff, records, observations). The use of multiple sources and methods serves to increase confidence in fidelity ratings. Multiple studies within mental health have used convergent validity methods (Blakely et al., 1987; Lucca, 2000; Macias et al., 2001; McGrew, Pescosolido, & Wright, 2003). For example, in a study by Lucca (2000), on a vocational program for adults with psychiatric disabilities, the fidelity measure used was a 15-item checklist derived from the literature of program components essential to the clubhouse model. Convergent validity was established by examining the relationship of the fidelity score assigned to each program by non-staff evaluators with staff members' responses on a scale measuring how consistently their programs followed psychosocial rehabilitation principles. There was a significant correlation between the number of model components a clubhouse had in place and the programs adherence to rehabilitation principles, as reported by staff members.

Issues with Validating Fidelity Measures

According to Mowbray et al. (2003), most analyses to validate fidelity criteria are aggregations of individual data within programs, in which analysis is conducted at the program

level, ignoring within-program variability. While others have conducted their analysis at the individual level, coding program-level variables as attributes of all units associated with the program, ignoring the fact that the individual units are not independent. Aggregating individual ratings to form group level variables may seem appealing, but the validity of inferences based on these aggregated variables must mean that the aggregates refer to the same constructs as the individual responses (Schweig, 2014). For example, an evaluator may be interested in understanding whether a teacher has delivered a mathematics intervention with fidelity. This evaluator will be surveying teachers and students about the delivery of the intervention. Given that students are nested within teachers/classrooms, a multilevel approach to assessing fidelity is more appropriate when assessing students within classes along a measure of interest. The total variance in responses will be made up of both between (variance that exists between classes) and within (variance that exists within classes) group variance (Zyphur, Kaplan, & Christian, 2008).

When data is aggregated or analyses are conducted at the individual level, like in a single level model, than the variation within students is represented (and the assumption is students within the same class are independent), but the variation between classes is ignored. "Substantively, it is assumed that the aggregates refer to the same constructs as the individual responses. Statistically, it is assumed that there is cross level invariance in the measurement model; that is, there is invariance in the measurement structure across the individual (within-group) level and the between-group level" (Schweig, 2014, p259). The problem with that is that when individuals are associated with groups (i.e. students with classrooms), this independence assumption is likely to be violated. We can understand measurement invariance by applying the generalizability theory principle of measurement equivalence is that an assessment is relatively consistent across a variety of relevant situations, similarly, if a group member's true score

depends in part on the group which a member belongs, than measurement cannot be said to be equivalent across groups (Bonito, Ruppel, & Keyton, 2012). Dyer, Hanges, and Hall (2005) describe three issues to consider when trying to establish factorial validity for data that is multilevel:

1. When observations are correlated rather than independent, the fundamental independence assumption underlying many commonly used statistical techniques is violated;
2. Assumptions of invariance- relationships among constructs may vary at different levels and have different meanings or factor structures at different levels of analyses (compilation variables which occur only at the group level, composition variables are constructs that emerge from responses to individuals within groups, and fuzzy composition variables that are partially identical in that they operate at multiple levels, but their factor structure can vary across levels); and
3. The lack of empirical studies of construct validation of aggregate measures means that we don't know whether a given construct is structurally identical across different levels of analysis, or whether its structure is fuzzy or varies across levels.

Data collection designs that involve the use of multiple informants within a group, across multiple groups produce a multilevel structure that needs to be taken into account in the psychometric analyses of the data (Dedrick and Greenbaum, 2011). Ignoring the multilevel structure can lead to overestimates of the standard errors for factor loadings, inflation of Type I errors, and lead to inferences that are not consistent with either the within or between level analysis (Chen, Mathieu, & Bliese, 2004; Cronbach, 1971; Dedrick & Greenbaum, 2011; Dyer,

Hanges, & Hall, 2005; Raudenbusch & Bryk, 2002; Zyphur et al., 2008).

Single level psychometric analyses such as CFA of nested data of measures of group variables are problematic as they assume incorrectly that the data are independent and single level CFA operates using a single covariance matrix that does not take into account the multiple levels and ignores the fact that the factor structure of a group measure and it's psychometric properties (e.g. reliability) may not be the same at each level of analysis (Dedrick & Greenbaum, 2011).

Multilevel confirmatory factor analysis (MCFA) should be used considered when subjects are meaningfully nested within groups and the evaluation of the factor structure of a set of indicators is desired (Muthen, 1994). Multilevel modeling is the solution to the measurement of non-independent data; multilevel modeling is used to estimate variances at item, individual, and group level of analysis (Bonito et al., 2012; Raudenbush et al., 1991). When analyzing nested data, fitting a multilevel CFA (rather than aggregating data for a single level model or only using individual level data) leads to an analysis that involves both the within latent factors and between latent factors and within and between loadings are used to assess validity for students as well as classes.

Nesting and multilevel analyses are also to be considered when assessing reliability. For example, student scores on fidelity in a given classroom might be more alike than those of students in another classroom. Estimating reliability from data collected at multiple levels (e.g., students nested within teachers) can confound the within-group variance and between-group variance and lead to biased reliability estimates when the assumption of independent residuals is violated. As a consequence, single level reliability estimates may not reflect the true scale reliability at any single level of the analysis as it assumes a single level factor structure (Geldhof,

Preacher, & Zyphur, 2013). Multilevel confirmatory factor analysis can be used to estimate reliability within and between clusters in a multilevel model. The strength of the multilevel latent variable approach is that by partitioning the variance in the scores into within- and between teacher/class components, the reliability of the teacher/class for each factor can be obtained at each level (Dedrick & Greenbaum, 2011). Reporting Cronbach's alpha as evidence of acceptable reliability for multilevel data is not appropriate given that it assumes a single level factor structure. Therefore it is important to estimate multilevel reliability when analyzing multilevel data.

Summary

In an era of educational accountability and the need for transparency understanding how an intervention is delivered in the classroom is key to understanding why a program succeeds or fails. As discussed in detail in this chapter assessing fidelity is the key to examining the extent to which a program was implemented as intended. Only by understanding and measuring whether an intervention has been implemented with fidelity can researchers and practitioners gain a better understanding of how and why an intervention works, and the extent to which outcomes can be improved. Unless such an evaluation is made, it cannot be determined whether a lack of impact is due to poor implementation or inadequacies inherent in the program itself. The consequences of not assessing fidelity are not only methodological issues, as noted earlier, but also have substantive implications for student performance, if students do not receive the intended benefits of an intervention due to issues in intervention delivery. In recent efforts to conceptualize and measure the multilevel, multi-dimensional fidelity construct, greater awareness of the role of delivery and receipt of an intervention has been identified as playing a role in the evaluation of program effects (Zvoch, 2012).

The field will grow when fidelity measures are developed that extend beyond assessing adherence or dosage and move towards incorporating other key constructs of fidelity (such as participant responsiveness). Following that movement, evaluators and researchers need to take steps to establish the reliability and validity of these fidelity instruments. Finally, for contexts in which there is nesting, multilevel psychometric analyses should be conducted. This study takes these steps towards developing and validating measures of fidelity. In the following chapter on methods, the context for the creation and validation of the student fidelity measure will be presented.

The purpose of this study is to provide initial validation of student fidelity measures using confirmatory factor analysis to assess factorial validity (by testing the a priori models). In addition, convergent validity will be evaluated by examining the agreement between two different sources of information about a program and its operations (i.e., teacher and student reports) focusing on the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) components of Fidelity of Implementation (FOI). The IP and ISE components are specific to participant responsiveness aspects of fidelity.

CHAPTER 3
METHODOLOGY

In recognition of the practical need for valid and reliable measures of fidelity of implementation of reform based STEM instructional materials and the theoretical need in the field for a shared conceptual framework for Fidelity of Implementation (FOI), the University of Chicago's Center for Elementary Math and Science (CEMSE) team, with funding from the National Science Foundation, developed, pilot and field tested a suite of eight instruments aimed at measuring the FOI of reform based K-8 science and mathematics instructional materials programs. Various aspects of teacher and student interactions in classroom constitute the most important measurement dimensions of the fidelity of implementation (FOI) of instructional materials. The present study used a quantitative research design using data collected from the CEMSE Project to assess the reliability and validity of scores from the Fidelity of Implementation student questionnaire, which was designed to assess the participant engagement aspect of fidelity. This study also examined the extent to which teacher and student reports produce comparable data (i.e., convergent validity) on their interactions during science or mathematics class.

This chapter is organized into five sections. The first section begins with a brief description of the reforms that provide the educational context for the fidelity measures examined in this study. After this description, this chapter presents descriptions of the participants (schools, teachers, and students); measures; procedures; and data analyses used to address each research question.

Context

The Center for Elementary Mathematics and Science Education is a Research and Development Center within the University of Chicago. "The Center for Elementary Mathematics and Science Education continues the University of Chicago's long-standing commitment to improving precollege education and aims to support high quality mathematics and science instruction and learning for all students. Through the sharing of knowledge and the creation of useful products and programs, CEMSE seeks to make a positive difference for mathematics and science instruction throughout the nation" (Center for Elementary Mathematics and Science Education, 2014, About CEMSE). Their work comprises three components: (1) Research and Evaluation, (2) Tool Development, and (3) School Support Services. It is through their Research and Evaluation component (OUTLIER) that the data for this study was collected. Outlier Research & Evaluation received support from the Institute of Education Sciences to validate three teacher-level instruments for measuring innovation implementation (Teacher Questionnaire, Teacher Log, Classroom Observation Protocol) and to develop and validate a student-level questionnaire focused on student-reported engagement in mathematics and science instruction.

The participant engagement aspect of Fidelity of Implementation was assessed within the context of reform-based mathematics and science programs, which included four elementary-level curricula, Full Option Science System (FOSS), Science and Technology for Children (STC), Science Companion, and Everyday Mathematics (EM). Descriptive information about these interventions can be found in Appendix A.

Participants

In the fall of 2012, the teacher questionnaire and revised student questionnaire were administered in three districts: Kirby School District 140 (in Tinley Park, IL, a Chicago suburb), Stamford Public Schools (Stamford, CT), and Denver Public Schools (Denver, CO). These districts were recruited as part of the overall grant. Since students were completing the questionnaire online, the questionnaire administration was staggered over several weeks beginning mid-October and ending late January. This allowed time for all classrooms to access the lab so that students could take both the math and science online questionnaires.

Schools

A total of 41 elementary schools participated in the study. All elementary schools in Stamford and Kirby were invited to participate. The selection process for all schools participating in data collection from Denver involved a purposive, stratified sampling strategy. That is, within the Denver district, elementary schools were selected that best represented the district in terms of school size, student demographics, and/or student achievement. Twenty-four of the schools were located in the Denver, Colorado school district, 12 were located in Stamford, Connecticut and 5 were located in Tinley Park, Illinois. In Stamford and Kirby, only 12 and 5 schools, respectively were selected because that was the total number of schools in their districts. In Denver, the district was large with many schools, so CEMSE worked with the district to select schools that were representative of students in their district.

Teachers

Four hundred and twenty-nine, third, fourth and fifth grade classroom teachers from the sample schools completed the Teacher Instructional Questionnaire (TIQ). Tables 1 and 2 show

the number of teachers who participated in each of the surveys (mathematics and/or science) by grade and district. Of the 429 teachers who participated, only 242 (152 in math, 90 in science) were used in the analyses. In order to be used in the analyses teachers had to have a teacher ID number, so that their data could be connected to their respective students. According to the CEMSE team, who collected the data for this study, the reason that there were teachers without IDs was that some of the teachers of the students who participated in this study did not take the teacher questionnaire, so although those students identified their teachers there was no corresponding teacher survey to match to the student data.

Table 1

Teacher Math Survey

District	**Total N**	**Grade 3**	**Grade 4**	**Grade 5**
Denver	155	47	59	49
Stamford	70	21	26	23
Kirby	37	17	10	10
Total	262	85	95	82

Note. The teachers in this table represent all the teachers who completed the questionnaires, but only a subset of these teachers participated in this study.

Table 2

Teacher Science Survey

District	Total N	Grade 3	Grade 4	Grade 5
Denver	100	36	30	34
Stamford	37	13	14	10
Kirby	30	14	9	7
Total	167	63	53	51

Note. The teachers in this table represent all the teachers who completed the questionnaires, but only a subset of these teachers participated in this study.

Students

The student sample consisted of 10,403, 3rd, 4th and 5th graders who were enrolled in the 41 participating schools in the Fall of 2012, who had parental permission, and who themselves assented to participate in the research project. Each student was to complete a science questionnaire and a mathematics questionnaire, with the timing of questionnaire completion corresponding to the teacher's completion of each TIQ. Tables 3 and 5 show how many students completed the student questionnaire by subject, grade and district. Demographic information describing the students and teachers who participated can be found in Chapter 4 by content area (math and science). It is important to note that although there was a large sample of students who completed the student questionnaire, some of the student data did not have teacher identifiers (teacher ID) attached to their data. So for analyses that required a teacher ID, such as single level confirmatory factor analyses in which the standard errors were adjusted for the nested data within teachers, and for the two-level confirmatory factor analyses used to examine the student and teacher level models, students without a related teacher ID were dropped from

the analyses (Tables 4 and 6) for the number of students who participated in each of the surveys (mathematics and/or science) by grade and district.

Table 3

Student Math Survey

District	Total N	Grade 3	Grade 4	Grade 5
Denver	3416	1194	1239	983
Stamford	1777	590	588	599
Kirby	793	270	278	245
Total	5986	2054	2105	1827

Table 4

Student Math Survey for Students with a Teacher ID

District	Total N	Grade 3	Grade 4	Grade 5
Denver	2042	592	768	522
Stamford	461	133	219	144
Kirby	605	268	193	97
Total	3108	993	1180	763

Table 5

Student Science Survey

District	Total N	Grade 3	Grade 4	Grade 5
Denver	2317	815	783	719
Stamford	1356	507	444	405
Kirby	737	269	245	223
Total	4410	1591	1472	1347

Table 6

Student Science Survey for Students with a Teacher ID

District	Total N	Grade 3	Grade 4	Grade 5
Denver	1200	523	376	301
Stamford	262	113	104	45
Kirby	561	245	179	137
Total	2023	881	659	483

Measures

Development of the Student Questionnaire

In order to create a 20- to 25-item student questionnaire, an iterative approach incorporating already validated items as well as newly developed items was used. Selected items that appeared to fit the instructional pedagogical (IP) and instructional student engagement (ISE) critical components were modified and incorporated. In order to find these items, a literature review of instruments in the fields of both student engagement and learning environments was conducted initially by the Center for Elementary Math and Science Education (CEMSE). As part of my participation in this project, I supported CEMSE in the development of the student questionnaire aimed at measuring student engagement and teacher practices. This included searching for items in existing instruments on student engagement (Table 7) for the list of instruments (reviewed), writing new items, and modifying items to correspond with items that measure the same construct in the Teacher Instructional Questionnaire (TIQ). From the student

engagement instruments reviewed (for both the instructional pedagogy and instructional student engagement components) items were modified from WIHIC, ICEQ, CLES, and TROFLEI to better fit the study, as well as to align with what was measured in the TIQ. For critical components CEMSE wanted to measure but for which an inadequate number of appropriate items existed in the literature, items were created to fit the same response scale as the modified items.

Table 7

Student Engagement Instruments Reviewed for Item Development

Instrument	Purpose	Dimensions	Items & Scale	Grade of Respondents
CLES-CS-Constructivist Learning Environment Scale	Extent to which certain psychosocial factors are prevalent in science class taught by teachers who attended ISLE program	Personal relevance Uncertainty of science Critical voice Shared control Student negotiation	5 point scale; almost never to almost always	Secondary school
WIHIC-"What is Happening in this Class?'	Measures students' perceptions of their classroom environment	Student cohesiveness Teacher support Involvement Investigation Task orientation Cooperation Equity	5 point scale; almost never to almost always	Science class; Grades 7-9
TROFLEI-Technology-Rich Outcomes Focused Learning Environment Inventory	Assesses classroom environment	Student cohesiveness Teacher support Involvement Task orientation Investigation Cooperation Equity Differentiation Computer usage Young adult ethos Attitude to subject Attitude to computer use Academic efficacy	80 items; 10 scales; 5 point scale; almost never to almost always *7 of the 10 dimensions come from the WIHIC instrument	Grades 11-12
LEI-Learning Environment Inventory	Descriptive of typical school classes		105 items; Strongly disagree to Strongly agree	Junior and Senior High school
CES- Classroom Environment Scale	Perceptual measures of human environments		90 items; True–False response format	High school
ICEQ-Individualized Classroom Environment Questionnaire	Assesses dimensions which distinguish individualized classrooms from conventional ones		50 items; Almost never to Very often	High school

Table 7 (Continued)

Instrument	Purpose	Dimensions	Items & Scale	Grade of Respondents
QTI- Questionnaire on Teacher-Student Interaction	Developed to assess student perceptions of 8 behavior aspects (relationship between teacher and students)		5 point scale; Never to Always	8th, 9th, 10th grade
SLEI- Science Laboratory Environment Inventory	Assesses environment of science lab		35 items; Almost never to Very often	High School
MSLQ- Motivated Strategies for Learning Questionnaire	Used to measure students' motivational beliefs and self-regulated learning	Intrinsic value Test anxiety Cognitive strategy use Self-regulation Self-efficacy	56 items; 7 point scale; 1= not at all true of me to 7= very true of me	7th graders
MJSES- Morgan-Jinks Student Efficacy Scale	Assesses students' sense of self-efficacy	Self-Efficacy	30 items; 4-point scale; 4=really agree to 1= really disagree	7th and 8th grade

Discussion can occur at any time during a lesson, but must include a back-and-forth exchange (A-B-A) (e.g., it cannot be only a student asks a question and the teacher answers). Examples of strategies include asking students to rephrase, repeat, or respond to others' thoughts; using appropriate wait time; clarifying points students make; and using Think, Pair, Share or a similar strategy. The second section is focused on assessing four Instructional Student Engagement critical components: Students Contribute to Small Group Work (3 items), Students Engage in Discussion (4 items), Students Engage in Cognitively Demanding Work (4 items), and Students Take Risks (4 items). The items are presented in Table 7. Instructional Student Engagement critical components reflect the intended student behaviors and interactions during the enactment of the program. Some of the student engagement critical components are also desired outcomes of these programs, but in this context, they are considered essential elements of program implementation. For example, for Students Take Risks, items are focused on whether students take intellectual or emotional chances. This includes taking risks in trying new things, asking questions, answering questions, and revealing their own uncertainties about their work, and risk taking in other ways.

The items for each of the constructs and their critical components are provided in Tables 8 and 9. These tables also show the parallel teacher and student items by construct and critical component. For the teacher questionnaire, instructions and items were framed in the following way, *"In the next section we want to ask about some of your specific teaching practices. While you may always keep these practices in mind when you are teaching, when answering the following questions, think about how often you intentionally did the following while teaching the most recent complete unit this school year (or the unit you are currently teaching if you have not yet completed a unit this year)"*. Students were instructed as follows, *"Now you will read about some things your teacher may do during science time . Please tell us how much your teacher does each thing during science time: never or hardly ever, sometimes, or a lot."* Screen shots of the instruments can be found in Appendix B.

Included in the third section of the student questionnaire are a series of questions related to intrinsic and extrinsic motivation. Section four includes items assessing student self-efficacy. Finally, the fifth section of the survey requests demographic information from students on their age, grade, gender, and teacher name. Aside from the demographic items, the student questionnaire items utilized a 3-point frequency scale: *Never or Hardly Ever*, *Sometimes*, and, *A Lot*. The use of a 3-point scale is in keeping with other measures of children of like ages and grades (e.g., Achenbach's Child Behavior Checklist- Achenbach, 1991).

Table 8

Teacher and Student Items Measuring Instructional Pedagogical (IP) Critical Components

Construct	Items – Teacher Questionnaire	Items – Student Questionnaire
IP7: Teacher Facilitation of Student Interest	During the module, how often do you explicitly do the following? 7a. Engage student interest by connecting the lesson content with current events and real world phenomena. 7b. Engage student interest by making lesson content relevant to students (e.g., ask about past experiences, apply content to students' daily lives). 7c. Engage student interest through other means (e.g., tell an interesting story, use humor, bring in a guest speaker).	Please tell us how much your teacher does each thing during science time. 7a. My teacher makes science interesting. 7b. My teacher tells us how things we learn in science can be used in the real world. 7c. My teacher does things that make me like science.
IP2: Teacher Facilitation of Student Discussion	During the module, how often do you explicitly do the following? 2a. Ask students to respond to what other students have said. 2b. Clarify points students make during discussion. 2c. Ask questions in order to promote student discussion. 2d. Encourage students to talk and listen to one another.	Please tell us how much your teacher does each thing during science time. 2a. My teacher asks us questions during science time. 2b My teacher wants us to share ideas during science time. 2c. My teacher asks me to talk to my classmates about their science ideas. 2d. My teacher gives me the chance to talk to my classmates about my science schoolwork.
IP10: Teacher Use of Differentiation	During the module, how often do you explicitly do the following? 10a. Scaffold ideas and activities for individual students. 10b. Give students different activities based on ability or learning modality. 10c. Group students based on their ability or learning modality.	Please tell us how much your teacher does each thing during science time. 10a. All students in my science class do the same work at the same time. (R) 10b. During science time, some students do different work than others. 10c. During science time, I do work that is different from what other students are doing.

Table 9

Items Measuring Instructional Student Engagement

Construct	Items – Teacher Questionnaire	Items – Student Questionnaire
ISE2: Students Engage in Discussion	During the module, what proportion of your students regularly did the following? 2a. Shared findings/thoughts with the class. 2b. Conversed with you about the topic. 2c. Responded to your questions in a whole group setting. 2d. Conversed with one another about the topic.	Please tell us how much you do each thing during science time. 2a. I talk to other students about our science work. 2b. Students talk with each other about what we're learning during science time. 2c. During science time, I talk to my teacher about what we are learning. 2d. I am a good listener when my classmates are talking during science time.
ISE3: Students Engage in Cognitively Demanding Work	During the module, what proportion of your students regularly did the following? 3a. Interpreted written text. 3b. Supported conclusions with evidence. 3c. Considered alternative arguments or explanations. 3d. Analyzed (organized, processed, manipulated, and evaluated) data. 3e. Demonstrated reasoning. 3f. Made predictions. content and academic topics. 3g. Considered relationships between lesson content and academic topics. 3h. Considered relationships between lesson content and real world phenomena and current events.	Please tell us how much you do each thing during science time. 3a. During science time, I explain how I get my answer. 3b. When I come up with an answer in science class, I make sure that it makes sense. 3c. I explain why I agree or disagree with things my classmates say in science. understand the lesson. 3d. During science time, I work hard to understand the lesson.
ISE1: Students Contribute to Small Group Work	During the module, what proportion of your students regularly did the following? 1a. Contributed to group work. 1b. Managed time efficiently when in groups. 1c. Worked collaboratively with their peers.	Please tell us how much you do each thing during science time. 1a. When we work in science groups, we work as a team. 1b. During science time, I learn from other students when working in groups. 1c. When we do group work in science, I cooperate with other students.

Table 9 (Continued)

Construct	Items – Teacher Questionnaire	Items – Student Questionnaire
ISE4: Students Take Risks	During the module, what proportion of your students regularly did the following? 5a. Took risks in answering questions. 5b. Took risks in trying new things. 5c. Took other types of risks (expressing alternative viewpoints, asking for help).	Please tell us how much you do each thing during science time. 4a. When working on science problems, I am willing to try something new or different. 4b. I say what I think in science even if it's different from other students in the class. 4c. During science time, I ask questions when I am confused, even when the other students 'get it'. 4d. I am not embarrassed to answer questions during science time.

The Teacher Instructional Questionnaire was comprised of parallel items for the Instructional Pedagogical and Instructional Student Engagement critical components. All teacher questionnaire items used a 5-point frequency scale: *Never*, *A few class sessions*, *About half the class sessions, Many class sessions*, and *Nearly all class sessions*. See Tables 6 and 7 for the teacher items that parallel the student items.

Procedures

Pilot Testing the Student Questionnaire

In order to identify potential problems with new items, cognitive interviews were conducted with a sample of students (Beatty & Willis, 2007; Presser et al., 2004). The Center for Elementary Mathematics and Science Education research team members conducted the cognitive interviews. During this process issues such as difficulties encountered when answering items (addressing issues of comprehension), respondents' interpretations of items, and how respondents arrived at their answers were identified. The goal was to conduct cognitive

interviews with 36 students, representing approximately six students of each gender in each of the three grade levels from third through fifth grade. Cognitive interviews took place within two Chicago metro-areas schools and were conducted only with those students enrolled in grades 3-5, as of the fall of 2012, who had parental permission, and who themselves assented to participate in the research project. Twenty-five cognitive interviews were conducted, representing both genders and the three grade levels. Each student provided feedback to half of the items (items were divided into "Form A" and "Form B" and customized for either mathematics or for science). Items were divided across forms in a "split half" fashion such that each form contained items from each construct. After the interviews were completed, I reviewed and entered all the data provided by the CEMSE Research Team and provided feedback and edits on the Student Questionnaire to the CEMSE Research Team. They then refined the instrument based on feedback from the cognitive interviews. Based on student feedback, the measurement of four critical components were omitted: Enactment of Class Structures, Enactment of Instructional Delivery Formats, Teacher Facilitation of Student Autonomy, and Teacher Facilitation of Students Taking Risks (13 items total). From the remaining 57 items, 28 items were retained, of which 16 items were reworded and 2 were new items.

Field Testing the Student Questionnaire

The revised Student Questionnaire was administered in May of 2012, and 275 students completed the survey as part of the field-testing. Since data from the Student Questionnaire were to be triangulated with the Teacher Instructional Questionnaire data, students completed both science and mathematics questionnaires and administration of Student Questionnaires coincided with Teacher Instructional Questionnaire administration. Thirty-one teachers (of the 102 to whom it was administered) completed the corresponding Teacher Instructional Questionnaire.

These students and their teachers were across eight classrooms (of the 31 classrooms in which it was administered). Participating schools in the field test were recruited from two districts: Champaign and Evanston to minimize the cost of data collection. The investigator of this dissertation study was involved in the analysis of field test data, secondarily, but was not involved in the data collection activities that occurred in Champaign and Evanston.

Student Questionnaire Administration for Validation

Following the field-testing, which occurred in May of 2012, the student questionnaire was revised based on reliability assessments and exploratory factor analysis results. In the Instructional Pedagogical critical component, one item from IP2 and one item from IP10 were omitted. These items were omitted because they had low item to total correlations and in the case of IP10 weak factor loadings. For IP2, the omitted item, "During science time, my teacher talks the whole time and doesn't really give us a chance to ask or answer questions," was replaced with, "My teacher wants us all to share ideas during math [or science] time". For IP10, item 11 was omitted, "My teacher lets me work at my own speed in math [or science] class". In the Instructional Student Engagement critical component, one factor, Students Demonstrate Autonomy, was dropped due to low and negative factor loadings, so only four of the original five factors were retained. One additional item was added, measuring Students Take Risks, "I am not embarrassed to answer questions in math [or science] class" and the wording was revised for two other items in that same scale.

Following these revisions, the student questionnaire was administered online beginning in the fall of 2012. The target number of participants was 4,500 students in mathematics and 4,500 students in science and the Teacher Instructional Questionnaire was administered to approximately 450 math teachers and science teachers across the 41 schools. Approximately

10,403 students completed the survey and 429 teachers (262 math, 167 science). The student response rate for this study was greater than 100% and teacher response rate was also high at 95.3% for all teachers. Students took the questionnaire online (previous administrations of the survey were paper and pencil). Students completed their surveys in the school's computer lab. There is no information available as to whether students were assisted, but the CEMSE researchers worked hard to get the items down to a 2nd grade reading level. CEMSE researchers operated under the assumption that students would be independently completing the surveys. On average it took students 12 minutes to complete the online questionnaire. Teachers also completed their survey online. Teacher questionnaires were lengthier, taking approximately 30 minute to complete, as the instructional pedagogy and instructional student engagement components were just one part of the teacher questionnaire. Teachers were instructed to "participate in completing an online questionnaire about the factors that affect their use of mathematics and/or science instructional materials". Teachers completed one teacher questionnaire for all the math/science classes they taught, so teacher responses were not connected to a specific class. As mentioned earlier some of the student data did not have teacher identifiers (teacher ID) attached to their data. So for analyses that required a teacher ID, such as single level confirmatory factor analyses in which the standard errors were adjusted for the nested data within teachers, and for the two level confirmatory factor analyses used to examine the student and teacher level models, students without a related teacher ID were dropped from the analyses. Additional details about the number of students and teachers in the various analyses are presented in Chapter 4. Also presented in Chapter 4 are analyses looking at whether significant differences exist between students with TIDs and students without TIDs, as well as descriptive information about teachers in both samples. Since students were completing the

questionnaire online, the questionnaire administration was staggered over several weeks beginning mid-October and ending late January. This allowed time for all classrooms to access the lab so that students could take both the mathematics and science online questionnaires.

Data Analysis

The objective of this analysis was to evaluate the reliability and validity of the scores from these instruments as indicators of fidelity of implementation. Prior to conducting the primary analyses addressing validity and reliability, descriptive statistics for the scales (mean, standard deviation, skewness and kurtosis) and items were examined. Intercorrelations of the variables and missing data were also examined. Preliminary analyses for this study were conducted using IBM SPSS Statistics for Macintosh, Version 22.0.

Research Questions

Research Question 1: *What is the internal consistency reliability of the scores for the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) components?*

The questions below were examined for the student data by both mathematics and science. Single-level and multilevel estimates of reliability for the IP and ISE scores were calculated.

1a. What is the internal consistency reliability of the scores for each of the three factors of Instructional Pedagogical (IP)?

1b. What is the internal consistency reliability of the scores for the overall Instructional Pedagogical (IP) component?

1c. What is the internal consistency reliability of the scores for each of the four factors of Instructional Student Engagement (ISE)?

1d. What is the internal consistency reliability of the scores for the overall Instructional Student Engagement (ISE) component?

As part of the preliminary analyses, internal consistency reliability analyses (Cronbach's alpha) were conducted to determine the reliability of the scores from the student questionnaire math measure and the student questionnaire science measure, looking at the IP and ISE critical components separately and in combination. Item-to-total correlations were used as part of the item analyses.

Estimating reliability from data collected at multiple levels (e.g., students nested within teachers) can confound the within-group variance and between-group variance and lead to biased reliability estimates when the assumption of independent residuals is violated. As a consequence, single level reliability estimates may not reflect the true scale reliability at any single level of the analysis as it assumes a single level factor structure (Geldhof et al., 2013). Therefore it is important to estimate multilevel reliability when analyzing multilevel data. Following the single level reliability analyses, multilevel reliability analyses were computed for IP and ISE using theintraclass correlation coefficients (ICCs) with the Spearman-Brown formula for both the mathematics and science data clustered by teacher.

Research Question 2*: Do individual items provide valid measures for the two FOI subcategories being examined in the Student Questionnaire, Instructional Pedagogical (IP) and Instructional Student Engagement (ISE)?*

2a. How well does the three-factor model of Instructional Pedagogical (IP) and the four-factor model of Instructional Student Engagement (ISE) fit the student self-report data in mathematics?

2b. How well does the three-factor model of Instructional Pedagogical (IP) and the four-factor model of Instructional Student Engagement (ISE) fit the student self-report data in science?

2c. How well does the three-factor model of Instructional Pedagogical (IP) and the four-factor model of Instructional Student Engagement (ISE) fit the teacher self-report data in mathematics?

2d. How well does the three-factor model of Instructional Pedagogical (IP) and the four-factor model of Instructional Student Engagement (ISE) fit the teacher self-report data in science?

Prior to this analysis, the factor structure was examined using exploratory factor analysis (principal axis with promax rotation) on the field test data. The results of this analysis were inconclusive, and may have been limited by sample size (n = 252 students). To assess dimensionality, the fit of the models for research questions 2A to 2D were evaluated using confirmatory factor analysis for mathematics instruction and for science instruction, separately. According to Brown (2006), "confirmatory factor analysis requires a strong empirical or conceptual foundation to guide the specification and evaluation of the factor model. CFA is typically used in the later stages of scale development or construct validation after the underlying structure has been tentatively established by prior empirical analyses using EFA, as well as on theoretical grounds" (pp. 40-41). Following Brown's guidance, CFA was selected to examine the fit of the factor models, following the EFA conducted in the field test, and was guided by the CEMSE Team's previous work in assessing factorial validity of the TIQ.

Using the statistical package of SPSS (Version 22.0), the data were screened for outliers, and examined for response distributions and missing data. Normality was not assumed or part of

the data screening procedures as the data was treated as ordered categorical variables (using Weighted Least Squares Means and Variance adjusted estimation method). The first step of CFA was to specify the model. Two models were specified. A three-factor model was posited whereby the 10 observed measures of Instructional Pedagogy were hypothesized to load on Teacher Facilitation of Student Interest, Teacher Facilitation of Student Discussion, and Teacher Use of Differentiation. The 15 items representing Instructional Student Engagement were hypothesized to load on four factors: Students Contribute to Small Group Work, Students Engage in Discussion, Students Engage in Cognitively Demanding Work, and Students Take Risks. Each model was run separately, but identically for both teacher and student data. I began my analyses by conducting single-level CFAs using Type = Complex in Mplus to take into account that the students were nested within teachers. Following that, I looked at multilevel (two-level CFAs). Prior to conducting the MCFA, the variability between and within teachers on each item was examined by computing the intra-class correlations (ICCs) for each of the items in each of the domains. The ICCs for the observed variables provide a measure of the amount of variability between teachers and the degree of non-independence or clustering of the data within teachers. Using a random effects model, the ICC for an item represents the variation between teachers in the intercepts (means) of the item divided by the total variation (sum of the variation between teachers in the intercepts and the variation within teachers). ICCs can range from 0 to 1, with larger values indicating greater clustering effects within teachers. Although there are no firm guidelines for deciding how large an ICC needs to be to warrant multilevel analyses, most of the published MCFAs have reported ICCs greater than .10 (e.g., Dedrick & Greenbaum, 2011; Dyer et al., 2005; Hox, 2002). As a rule of thumb, Hox (2010) considers ICCs of .05, .10, and .15 as small, medium, and large, respectively, for organizational research.

All measurement error was presumed to be unsystematic, implying that there were no correlated measurement errors for any pair of indicators. In addition, for this measurement model the latent factors of Instructional Pedagogical and Instructional Student Engagement were hypothesized to be correlated. Following the specification of the model, the model parameters were estimated. Mplus Version 7 (Muthen & Muthen, 1998-2014) was used, as it takes into account the nested data structure proposed in this study (i.e., students are nested within teachers).

Analyses of the categorical items were based on the polychoric correlations and parameters were obtained using weighted least squares means and variance adjusted estimation method (WLSMV) adjusted chi-square. When WLSMV estimation is used, Mplus uses pairwise deletion for missing data with the assumption that the data are missing completely at random. When variables are measured on an ordinal scale and there are few categories, such as in this case, estimation methods designed for categorical methods are recommended. Also, a categorical approach is less biased when compared with standard ML when the ordinal variable is skewed or kurtotic, as it was in some cases of this study. The acceptability of the fitted CFA solution was evaluated based on overall goodness of fit using multiple goodness of fit indices (e.g., Chi-square and degrees of freedom, Standardized Root Mean Square Residual [SRMR] of $< .08$ when available, Root Mean Square Error of Approximation [RMSEA] $< .06$, and the Comparative Fit Index [CFI] of $> .95$), and interpretability/strength of parameter estimates (Brown, 2006).

Research Question 3: *What is the convergent validity of the scores from the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) scales in mathematics and in science when measured by teacher- and student-reports?*

Finally, the extent to which there is a correlation between teacher and student reports on FOI Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) items was examined. Initial cross-instrument comparisons were conducted by calculating correlations of corresponding factors between the student and teacher scores obtained from the respective questionnaires. Then correlations of corresponding composite indices calculated for the critical components were examined. Individual student questionnaire data were aggregated to the classroom level. Following that, the data were examined in M*plus* (Version 7.2) by estimating the correlation of the latent variables, taking into account the two-level framework. The correlations between teachers' and students' scores on the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) items were examined. By correlating the teacher self-report data to the student self-report data, taking into account the two-level framework the degree of correspondence between the student report and self-reported teacher data can be more rigorously assessed.

Protection of Human Subjects

Institutional Review Board approval from the University of South Florida was not necessary for the scope of this dissertation project, as it was a secondary analysis of the data collected by the CEMSE Research Team and I did not interact with any human subjects. CEMSE obtained parental permission and student assent for students who participated in this study. A waiver of informed consent (parental permission) was used by CEMSE, and students assented to participate in the study. A screen shot of the student assent from the online survey can be found in Appendix B.

CHAPTER 4
ANALYSIS AND RESULTS

The purpose of this study was to evaluate the reliability and validity of the scores from both the student and teacher fidelity of implementation questionnaires. The focus of this study was on the Instructional Pedagogical (IP; e.g., teacher facilitation of student discussion, teacher facilitation of student interest) and Instructional Student Engagement (ISE; e.g., students engage in discussion, students demonstrate autonomy) components of Fidelity of Implementation (FOI) that are specific to the participant responsiveness aspects of assessing fidelity. Convergent validity was evaluated by examining the relationship between two different sources of information about a program and its operations (i.e., teacher and student reports). This chapter presents the results of this study organized by component (i.e., IP, ISE) and content area (i.e., mathematics, science). Within each description of the results of the component and content area, each of the three research questions is addressed. All of the questions are answered using data from a sample of teachers and students in 41 schools across three school districts. To answer the questions addressed in this research, different samples of varying sizes were used. For preliminary single level analysis (not taking into account the nested data structure), such as demographics and item analyses, Cronbach's alpha for reliabilities, and correlations between instruments, as well as confirmatory factor analyses, the entire sample of students was used (N= 5,986 for mathematics, N=4,410 for science). In order to attend to the multilevel nature of the data in the psychometric analyses involving the multilevel confirmatory factor analyses (MCFAs) and convergent validity, a subset of students who had teacher IDs associated with their

responses was used (*N*= 3,103 for Mathematics IP, *N*= 3,096 for Mathematics ISE, *N*=2,023 for Science IP, *N*=2,021 for Science ISE).

The questions addressed by this study include:

1. What is the internal consistency reliability of the scores for the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) components?
2. To what extent does the hypothesized factor structure fit the student and teacher data for the two FOI subcategories being examined in the Student and Teacher Questionnaire: Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) in mathematics and in science?
3. What is the convergent validity of the scores from the Instructional Pedagogical (IP) and Instructional Student Engagement (ISE) scales in mathematics and in science when measured by teacher- and student-reports?

Mathematics Student and Teacher Demographics

For mathematics, there were 5,986 students in the sample. Of those students, 49.4% were boys. The sample was ethnically diverse, in that students came from a range of ethnicities. Whites were the largest ethnicity at 26.2%, followed by 23.6% of the students who identified themselves as Other, Hispanics at 22.4%, 11.1% of students who identified themselves as Mixed, and 7.9% who were African American/Black.

Students participating in this study were in grades 3-5, with 34.3% of students in the 3rd grade, 35.1% in 4th grade, and 30.5% in 5th grade. The mean age for students in this sample was 9 years of age (ranging from 7-12 years). Students came from 41 schools across the three districts in the sample. Mathematics students' predominately came from the Denver district (57.0%), followed by the Stamford district (29.7%), and then the Kirby district (13.3%).

For the 152 mathematics teachers analyzed in this sample, gender, age and ethnicity were not requested demographics. The majority of mathematics teachers held a bachelor's degree (73.5%), followed by a master's degree (25.0%), and few had a doctoral degree (0 .7%). Only 8.6% of these teachers had a degree in Mathematics and 2.2% were mathematics specialists/coaches. In terms of years of teaching experience, mathematics teachers' experience ranged from 6% for one year of experience to 11.3% for teachers who had 25 or more years of experience. Mathematics teachers primarily taught 4th grade (36.8%), followed by 3rd grade (32.9%), and then 5th grade (24.3%).

For the 110 teachers who were not analyzed in this study, the majority of these teachers had a master's degree (74.3%), followed by a bachelor's degree (24.9%, and few had a doctoral degree (0.8%). Similar to the sample of teachers that were analyzed, 8.3% had a degree in Mathematics and 2.8% were mathematics specialists/coaches. In terms of years of teaching experience, these teachers experience ranged from 5.1% for one year of experience to 11.8% for teachers who had 25 or more years of experience. Mathematics teachers primarily taught 4th grade (37.9%), followed by 3rd grade (31.9%), and then 5th grade (24.5%).

Instructional Pedagogical Component in Mathematics

Instrument, Item Descriptives, and Reliability Assessment

As described in the Methods in Chapter 3 the student instrument was composed of two domains: Instructional Pedagogical and Instructional Student Engagement. The first section of the student instrument was focused on assessing three Instructional Pedagogical critical factors: Teacher Facilitation of Student Interest (3 items), Teacher Facilitation of Student Discussion (4

items), and Teacher Use of Differentiation (3 items). Instructional Pedagogical critical components reflect the intended teacher and student behaviors and interactions that take place during program use. For example, in Teacher Facilitation of Student Discussion, items are focused on whether the teacher encourages and promotes students' discussions with one another. In this case, discussion is an on-topic, substantive exchange of ideas. Discussion can occur at any time during a lesson, but must include a back-and-forth exchange (A-B-A) (e.g., it cannot be only a student asks a question and the teacher answers). Examples of strategies include asking students to rephrase, repeat, or respond to others' thoughts; using appropriate wait time; clarifying points students make; and using Think, Pair, Share or a similar strategy. The student questionnaire items utilized a 3-point frequency scale: *Never or Hardly Ever*, *Sometimes*, and *A Lot*. Descriptive statistics for the items and scales can be found in Tables 10 and 11.

Item means ranged from 1.66 (*SD* = 0.66) for 'doing work different from other students' (teacher use of differentiation) to 2.68 (*SD*= 0.51) for 'teacher asking questions during math time' (teacher facilitation of student interest), with sample sizes for the items varying from 5,972 for teacher facilitation of student interest, and teacher facilitation of student discussion to 5,976 for teacher use of differentiation. Less than 1% of cases were missing in the Math sample (.40%). Responses were approximately normally distributed, with skewness ranging from -1.29 to 0.50 and kurtosis values ranging from -0.90 to 0.64 (Table 10).

Table 10

Item Descriptives for the Mathematics Student Questionnaire – Instructional Pedagogical

Subscale Item	*N*	Number of Missing Cases	*M*	*SD*	Skewness	Kurtosis	ICC
Teacher Facilitation of Student Discussion (IP2)							
My teacher asks us questions during math time. (2a)	5976	15	2.68	0.51	-1.29	0.64	.06
My teacher wants us all to share ideas during math time. (2b)	5976	15	2.39	0.63	-0.52	-0.64	.18
My teacher asks me to talk to my classmates about their math ideas. (2c)	5976	15	2.06	0.67	-0.07	-0.77	.30
My teacher gives me the chance to talk to my classmates about my math schoolwork. (2d)	5976	15	2.02	0.69	-0.03	-0.90	.22
Teacher Facilitation of Student Interest (IP7)							
My teacher makes math interesting. (7a)	5976	15	2.55	0.57	-0.85	-0.27	.12
My teacher tells us how things we learn in math can be used in the real world. (7b)	5976	15	2.48	0.62	-0.74	-0.43	.12
My teacher does things that make me like math. (7c)	5976	15	2.51	0.61	-0.84	-0.29	.07
Teacher Use of Differentiation (IP10)							
All students in my math class do the same work at the same time. (10a-reverse coded)	5972	19	2.41	0.60	-0.49	-0.65	.11
During math time, some students do different work than others. (10b)	5972	19	1.89	0.64	0.10	-0.57	.14
During math time, I do work that is different from what other students are doing. (10c)	5972	19	1.66	0.66	0.50	-0.72	.11

Note. ICC = Intraclass correlation coefficient. ICCs are reported only for the sample of students who had a teacher ID (*N*= 3103). Response scale ranged from 1 (*Never or Hardly Ever*) to 3 (*A Lot*).

Table 11

Student Responses for the Mathematics Student Fidelity of Implementation Questionnaire Instructional Pedagogical Domain

Subscale **Item**		***Never or Hardly Ever (1)***	***Sometimes (2)***	**A lot (3)**
	N	***%***	***%***	**%**
Teacher Facilitation of Student Discussion (IP2)	5976			
My teacher asks us questions during math time. (2a)		2.4	27.2	70.5
My teacher wants us all to share ideas during math time. (2b)		7.7	45.6	46.6
My teacher asks me to talk to my classmates about their math ideas. (2c)		19.5	54.9	25.6
My teacher gives me the chance to talk to my classmates about my math schoolwork. (2d)		22.7	52.3	24.9
Teacher Facilitation of Student Interest (IP7)	5976			
My teacher makes math interesting. (7a)		4.1	36.6	59.3
My teacher tells us how things we learn in math can be used in the real world. (7b)		6.4	39.5	54.1
My teacher does things that make me like math. (7c)		6.0	36.9	57.1
Teacher Use of Differentiation (IP10)	5972			
All students in my math class do the same work at the same time. (10a-reverse coded)		47.3	46.7	6.0
During math time, some students do different work than others. (10b)		26.5	58.3	15.2
During math time, I do work that is different from what other students are doing. (10c)		44.7	45.0	10.3

Item means ranged from 1.66 (*SD* = 0.66) for 'doing work different from other students' (teacher use of differentiation) to 2.68 (*SD*= 0.51) for 'teacher asking questions during math time' (teacher facilitation of student interest), with sample sizes for the items varying from 5,972 for teacher facilitation of student interest, and teacher facilitation of student discussion to 5,976 for teacher use of differentiation. Less than 1% of cases were missing in the Math sample

(.40%). Responses were approximately normally distributed, with skewness ranging from -1.29 to 0.50 and kurtosis values ranging from -0.90 to 0.64 (Table 10).

Cronbach's alphas for the three scales described in Table 10, not taking into account the multilevel data structure were .62, .56, and .55, respectively (Table 12). Given the multilevel nature of this data, these Cronbach's alphas represent a first look at the reliability of the data. Further below under the section entitled Multilevel Confirmatory Factor Analysis, the reliabilities are computed using the ICCs with the Spearman-Brown formula for the mathematics sample of students nested within teachers.

Table 12

Internal Consistency of Instructional Pedagogical Subscales for Math

Scale	# of Items	Cronbach's α	N	Item-to-Total Correlation Range
Teacher Facilitation of Student Discussion (IP2)	4	.62	5976	.21 to .51
Teacher Facilitation of Student Interest (IP7)	3	.56	5976	.24 to .47
Teacher Use of Differentiation (IP10)	3	.65	5972	.24 to.45

In order to assess whether significant differences in the mean IP scores existed between students who had teacher ID's and students without teacher ID's (TIDs) an independent-samples t-test was conducted. For Teacher Facilitation of Student Discussion (IP2), there was a significant difference in scores for students with TIDs (M=2.26, SD=0.43), and students without TIDs (M=2.26, SD=0.42; t[5956.95]=4.82, p=.00). The magnitude of the differences in the means was very small (eta squared = .004). For Teacher Facilitation of Student Interest (IP7), there was a significant difference in scores for students with TIDs (M=2.52, SD=0.43), and

students without TIDs (M=2.50, SD=0.45; t [5895.44]=2.15, p=.03). The magnitude of the differences in the means was very small (eta squared = .001). For Teacher Use of Differentiation (IP10), there was a significant difference in scores for students with TIDs (M=1.72, SD=0.46), and students without TIDs (M=1.70, SD=0.46; t[5970]=2.07, p=.04). The magnitude of the differences in the means was very small (eta squared = .001).

Confirmatory Factor Analysis for the Math Instructional Pedagogical Student Model

Confirmatory Factor Analyses (CFA) and Multilevel Confirmatory Factor analyses (MCFA) were conducted using Mplus Version 7 (Muthen & Muthen, 1998-2014). Analyses were based on the polychoric correlations for the ordinally scaled items, and parameters were obtained using WLSMV estimation that assumes missing completely at random (after missing teacher data were removed from the sample, any remaining missingness was assumed to be completely at random). As was described in Chapter 3 a categorical analysis approach was used. The rationale for the use of categorical instead of continuous can be found there.

Overall goodness of fit for the models was evaluated using the X^2 likelihood ratio statistic, Bentler's (1992) normed comparative fit index (CFI), root mean square error of approximation (RMSEA; Steiger & Lind, 1980) and the standardized root mean square residual (SRMR). For MCFA, the between and within SRMR were also evaluated. Acceptable fit was judged by CFI values greater than .95 and SRMR values less than or equal to .08 and RMSEA values less than or equal to .06 (Hu & Bentler, 1999). Multiple fit statistics were used because each has its own limitations.

Confirmatory factor analysis with corrected standard errors for nested data.

Given the complexity of multilevel confirmatory factor analysis (MCFA) models, simpler models are recommended as a preliminary step in conducting MCFA. A multilevel confirmatory factor analysis of the type of data in this study can sometimes run into convergence problems or improper solutions. Therefore, before running the MCFA, I examined the factor structure using a single-level CFA with robust weighted least squares (WLS) approach (estimator = WLSMV in Mplus) and standard errors adjusted to take into account cluster sampling (i.e., nested data) to examine the three-factor measurement model underlying the Instructional Pedagogical domain. The data were clustered by teacher ID. In order to take into account the nested data structure (i.e., student data nested within teachers), it was necessary for the student to have an associated teacher ID. Students without a teacher ID were eliminated from this analysis and later for the multilevel analyses. The single level CFA does not take into account the two-level structure of the data; it is based on the total polychoric correlation matrix of the observed variables (i.e., the total polychoric correlation matrix is not decomposed into between and within, which is the case for the MCFA).

The chi-square value for the single level, three-factor CFA model, X^2 (32, N=3103) = 485.40, $p< .05$, indicated a statistically significant lack of fit. Alternative measures of fit, which are less sensitive to sample size, also suggested a lack of fit. The RMSEA of .07 was slightly higher than Hu and Bentler's (1999) cutoff of .06 and the CFI of .89 was less than the .95 cutoff value for this index. A single, level three factor CFA for students without TIDs was also run to examine if differences existed. The model fit indices for the Student CFA models with TIDs can be found in Table 13 and the model fit indices for the Student CFA models without TIDs can be

found in Table 14. As can be seen in the tables the models fit pretty similarly for both students with TIDs and students without TIDs.

Table 13

Student (Single Level) Confirmatory Factor Analysis Fit Indices for Responses with TIDs

Model	X^2	*df*	CFI	RMSEA
IP Model for Math (*N*=3103)	485.40	32	.89	.07
ISE Model for Math (*N*=3096)	955.98	84	.89	.06
IP Model for Science (*N*=2023)	352.97	32	.93	.07
ISE Model for Science (*N*=2021)	699.83	84	.91	.06

Note. RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index.

Table 14

Student (Single Level) Confirmatory Factor Analysis Fit Indices for Responses without TIDs

Model	X^2	*df*	CFI	RMSEA
IP Model for Math (*N*=2873)	468.66	32	.93	.07
ISE Model for Math (*N*=2868)	1355.10	84	.87	.07
IP Model for Science (*N*=2387)	665.08	32	.93	.09
ISE Model for Science (*N*=2383)	1187.43	84	.91	.07

Note. RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index.

All factor pattern coefficients (loadings) were significantly different from zero *($p < .05$)*. The standardized loadings for the items within the IP2 factor (teacher facilitation of student discussion) ranged from .35 to .79, from .53 to .76 for IP7 (teacher facilitation of student interest), and from .29 to .69 for IP10 (teacher use of differentiation). The correlations between

the factors were positive and significantly different from zero ($p < .05$) with IP2 and IP7, 1P2 and 1P10, and IP7 and IP10 correlating at .57, .22, and .10, respectively.

An alternative one-factor model was also considered. This model did not fit as well as the three-factor model based on the chi-square value, X^2 (35, N=3103) = 1926.43, $p < .05$, and the other fit indices (RMSEA=.13, and CFI=.54). Standardized item loadings on the one-factor model ranged from -.05 to .72.

Given that students were nested within teachers, thus violating the independence assumption, multilevel confirmatory factor analysis was used to further analyze the data for this study.

Multilevel Confirmatory Factor Analysis for the Mathematics Instructional Pedagogical Student Model

Prior to conducting the MCFA, the variability between and within teachers on each item was examined by computing intra-class correlations (ICCs) for each of the 10 items in the Instructional Pedagogical domain. The ICCs for the observed variables provide a measure of the amount of variability between teachers and the degree of non-independence or clustering of the data within teachers. Using a random effects model, the ICC for an item represents the variation between teachers in the intercepts (means) of the item divided by the total variation (sum of the variation between teachers in the intercepts and the variation within teachers). ICCs can range from 0 to 1, with larger values indicating greater clustering effects within teachers. As mentioned in Chapter 3, there are no firm guidelines for deciding how large an ICC needs to be to warrant multilevel analyses. Table 10 displays the ICCs for the 10 items in the Instructional Pedagogical domain for math. The ICCs for each of the observed items ranged from .06 (for

item IP2a within the IP2 factor) to .30 (for item IP2c also within the IP2 factor). These values indicated that there was sufficient between teacher variability to warrant multilevel analysis.

As shown in Figure 1, a three-factor multilevel model, in which the same number of factors at each level was run (3 within factors and 3 between factors). Results of the three-factor multilevel model with loadings freely estimated across levels indicated mixed results in terms of model fit to the data.

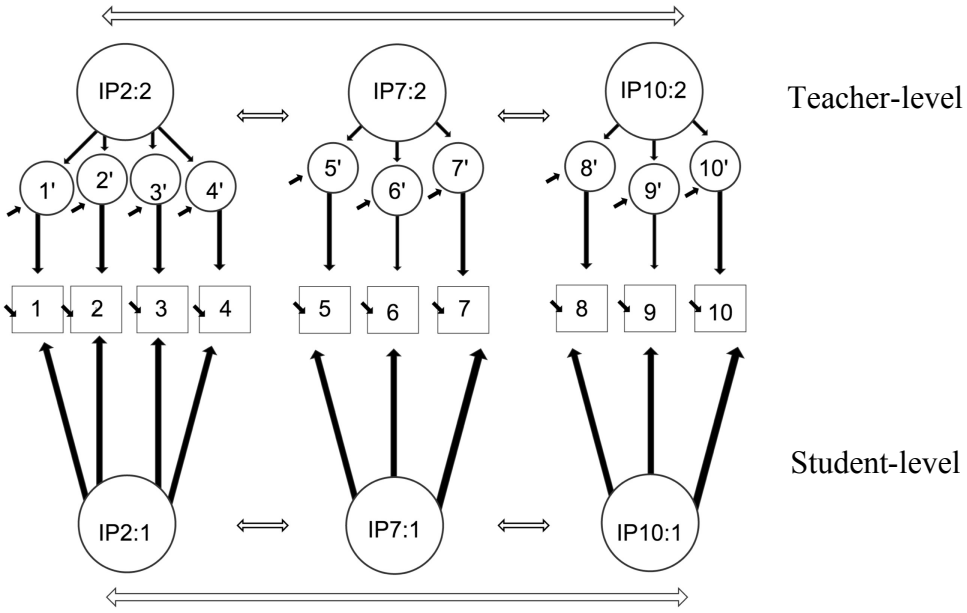


Figure 1. Three-Factor Multilevel Confirmatory Factor Analysis Model for Instructional Pedagogical in Mathematics

The RMSEA of .05 indicated acceptable fit overall but the CFI of .87 indicated less than acceptable fit. The SRMR fit indices at each level indicated that the fit of the Level 1 (within) part of the model was better than at Level 2 (SRMR within= .06 vs. SRMR between= .16; see Table 15 for measures of fit).

Table 15

Student Multilevel Confirmatory Factor Analysis Fit Indices

Model	X^2	*df*	CFI	RMSEA	SRMR
IP Model for Math (*N*=3103)	505.83	65	.87	.05	.06[a]/.16[b]
ISE Model for Math (*N*=3096)	902.68	169	.85	.04	.05[a]/.21[b]
IP Model for Science (*N*=2023)	407.49	66	.91	.05	.07[a]/.21[b]
ISE Model for Science (*N*=2021)	682.20	174	.91	.04	.06[a]/.27[b]

Note. RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual.
[a] Within
[b] Between

At level-1 (student) and level-2 (teacher), all factor pattern coefficients (loadings) were significantly different from zero ($p < .05$). See Table 16 for the unstandardized factor loadings.

In MCFA, fixing residual variances to zero at the between level to zero is often necessary when sample sizes at level-2 (teachers) are small and the true between-group variance is close to zero (Hox , 2002). In the case of IP for mathematics, the residual variances for the level-2 intercepts were fixed to zero for item 10c only.

Inter-factor correlations were .60 ($p < .05$) between IP2 and IP7 at level-1 and .73 ($p < .05$) at level-2; .16 ($p < .05$) between IP2 and IP10 at level-1 and .33 ($p <$. 05) at level-2; and .13 ($p < .05$) between IP7 and IP10 at level-1 and -.05 (not statistically significant) at level-2.

Table 16

Multilevel Confirmatory Factor Analysis: Unstandardized Factor Loadings and Residual Variances for the Three-Factor Model Underlying Student Ratings of Instructional Pedagogy

Item on the Rubric	Students with a TID (*N*=3103) Factor Loading	Teachers (*N*= 152) Factor Loading	Residual Variances
Teacher Facilitation of Student Discussion			
2a	1.00[a] (--)	1.00[a] (--)	0.07 (0.02)
2b	1.73 (0.16)	5.31 (1.70)	0.05 (0.02)
2c	2.63 (0.27)	9.22 (3.25)	0.05 (0.05)
2d	2.31 (0.23)	7.15 (2.47)	0.02 (0.04)
Teacher Facilitation of Student Interest			
7a	1.00[a] (--)	1.00[a] (--)	0.23 (0.05)
7b	0.47 (0.05)	0.90 (0.26)	0.09 (0.03)
7c	0.822 (0.08)	0.77 (0.17)	0.09 (0.03)
Teacher Use of Differentiation			
10a	1.00[a] (--)	1.00[a] (--)	0.10 (0.02)
10b	6.18 (1.51)	3.82 (0.87)	0.02 (0.07)
10c	3.25 (0.42)	2.30 (0.49)	0.00 [b](-)

Note. Numbers in parentheses represent the standard error.
[a] Factor loading fixed to 1.0.
[b] Residual variances were fixed to 0.

Multilevel ICCs and Reliability

Estimating reliability from data collected at multiple levels (e.g., students nested within teachers) can confound the within-group variance and between-group variance and lead to biased reliability estimates when the assumption of independent residuals is violated. As a consequence, single level reliability estimates may not reflect the true scale reliability at any single level of the analysis as it assumes a single level factor structure (Geldhof et al., 2013). Therefore it is important to estimate multilevel reliability when analyzing multilevel data. Using this model, it was possible to calculate the ICCs for the three latent variables and, subsequently, the reliability of each factor when aggregated at the teacher level. The ICC is the variation

between teachers divided by the total variation. Total variation equals the combined within-and between- teacher variation. IP10 had the greatest amount of between teacher variability (ICC= .38), followed by IP7 (ICC= .07), and IP2 (ICC= .06). Using these ICCs with the Spearman-Brown formula, [*k*(ICC)/ [(k-1)(ICC) +1], where k is the average number of students nested within teachers , the estimated reliabilities for the factors in this study, with an average cluster size of 20 respondents (students) per teacher, were .92 for IP10, .60 for IP7, and .56 for IP2. See Tables 36 and 37 at the end of this chapter for summary tables of internal consistency results by level.

Confirmatory Factor Analysis for the Mathematics Instructional Pedagogical Teacher Model

In this section, the model fit based on teachers' self-reported data (rather than students' reports nested within teachers) is presented. The chi-square value for the single level, three-factor CFA model, X^2 (32, N=152) = 64.37, p< .05, indicated a statistically significant lack of fit. However, alternative measures of fit, which are less sensitive to sample size, suggested that the fit was marginally acceptable. The RMSEA of .08 was greater than Hu and Bentler's (1999) cutoff of .06, and the CFI of .98 was greater than the .95 cutoff values for this index.

All factor pattern coefficients (loadings) were significantly different from zero (p< .05). The standardized loadings for the items within the IP2 factor (teacher facilitation of student discussion) ranged from .68 to .88, from .83 to .97 for IP7 (teacher facilitation of student interest), and from .67 to .90 for IP10 (teacher use of differentiation). See Table 17 for the unstandardized factor loadings. The correlations between the factors were positive and significantly different from zero (p< .05) with IP2 and IP7, 1P2 and 1P10, and IP7 and IP10 correlating at .43, .36, and .43, respectively.

Table 17

Confirmatory Factor Analysis: Unstandardized Factor Loadings for the Three Factor Model Underlying Teacher Ratings of Instructional Pedagogy

Item on the Rubric	Teachers with IDs (*N*=152) Factor Loading
Teacher Facilitation of Student Discussion	
2a	1.00[a] (--)
2b	0.89 (0.06)
2c	0.72 (0.08)
2d	0.93 (0.08)
Teacher Facilitation of Student Interest	
7a	1.00[a] (--)
7b	1.17 (0.08)
7c	0.93 (0.06)
Teacher Use of Differentiation	
10a	1.00[a] (--)
10b	0.88 (0.08)
10c	0.74 (0.07)

Convergent Validity

In order to examine convergent validity, meaning the correlation between student and teachers responses on the Instructional Pedagogical domain, the factor scores from the student perspective were correlated with the factor scores from the teacher perspective. Students are informants, relaying information about instructional pedagogy about the teacher, but students also have their own factor model, as do teachers. The dataset consisted of 3,103 students (level-1) nested within 152 teachers of which all students had one teacher (level-2). Each of the 3,103 students provided data on instructional pedagogy from their perspective. These data constituted the lower-level (level-1) unit of analysis in this study. The second-level data included class instructional pedagogy scores for each of the 152 teachers. Data regarding instructional pedagogy were gathered from two sources: from the teachers (self-ascribed instructional

pedagogy) and their students (perceived instructional pedagogy). It should be noted that there is no variability in the teacher data for students in a class, as teacher responses were replicated for each student in that teacher's class. Also, given that in the data set 50 or more students could have been associated with a teacher ID, it is assumed that teachers taught more than one class, but that they only completed the teacher questionnaire once for all the classes they taught.

Preliminary analyses using SPSS were conducted using the observed variables. The student data for a teacher were aggregated to create a teacher mean, as were the teacher data (although given that teacher responses for each student in a class were the same, the mean was the same as the teachers' reported response). The correlations based on the observed variables between teacher and students on instructional pedagogical components Teacher Facilitation of Student Discussion, Teacher Facilitation of Student Interest and Teacher Use of Differentiation were .25, .15, and .42, respectively. Following that, the data were examined in M*plus* by estimating the correlation of the latent variables, taking into account the two-level framework (Figure 2). The data were treated as categorical (ordinal) and the parameters were estimated using robust weighted least squares (estimator WLSMV). This model, as well as the others in this study, was initially run as continuous, but when one model did not converge, it was decided that running these models with the data treated as categorical was more appropriate and in keeping with the analyses of the other single and multilevel models in this study. Also, the correlations between the latent variables for the categorical models were similar to those of the continuous models.

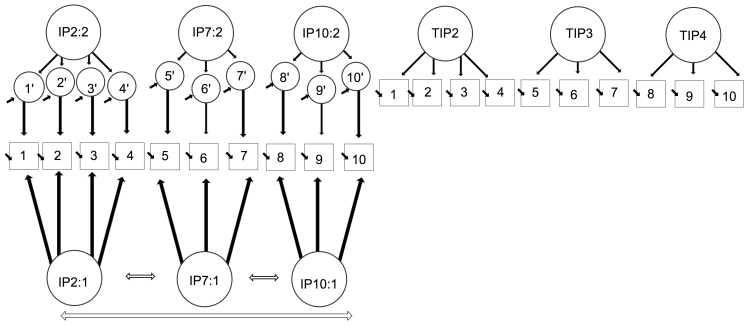


Figure 2. Multilevel Confirmatory Factor Analysis Teacher and Student Model for IP Mathematics Convergent Validity

The correlations between teachers' and students' scores on the instructional pedagogical components of Teacher Facilitation of Student Discussion, Teacher Facilitation of Student Interest, and Teacher Use of Differentiation were .38, .26, and .72, respectively. See Table 18 for Teacher and Student Correlations on the Instructional Pedagogical Domain.

Table 18

Correlations of Instructional Pedagogical Subscales from Student Questionnaire Compared with Teacher Questionnaire Using the Multilevel Confirmatory Factor Analysis Model

		Teacher Questionnaire (*N*= 152)		
	Scale	**IP2**	**IP7**	**IP10**
Student Questionnaire (*N*= 3103)	Teacher Facilitation of Student Discussion (IP2)	.38		
	Teacher Facilitation of Student Interest (IP7)		.26	
	Teacher Use of Differentiation (IP10)			.72

Instructional Student Engagement Component in Mathematics

Instrument, Item Descriptives, and Reliability Assessment

The second section of the student instrument was focused on assessing four Instructional Student Engagement critical factors: Students Contribute to Small Group Work (3 items), Students Engage in Discussion (4 items), Students Engage in Cognitively Demanding Work (4 items), and Students Take Risks (4 items). Instructional Student Engagement critical factors reflect the intended student behaviors and interactions during the enactment of the program. Some of the student engagement critical factors are also desired outcomes of these programs, but in this context, they are considered essential elements of program implementation. For example, for Students Take Risks, items are focused on whether students take intellectual or emotional chances. This includes taking risks in trying new things, asking questions, answering questions, and revealing their own uncertainties about their work, and risk taking in other ways. The student questionnaire items utilized a 3-point frequency scale: *Never or Hardly Ever*, *Sometimes*, and *A Lot*. See Table 19 for student responses to this scale by item.

Item means ranged from 1.96 (*SD* = 0.69) for 'during math time, I talk to my teacher about what we are learning' (students engage in discussion) to 2.74 (*SD*= 0.51) for 'during math time, I work hard to understand a lesson' (students engage in cognitively demanding work), with sample sizes for the items varying from 5,430 for students contribute to small group work to 5,964 for students engage in discussion. A little over nine percent (9.4%) of the data was missing for the factor Students Contribute to Small Group Work (ISE1). This was not random missing data, but rather the result of a screening question (*Do you ever work with a partner or in groups during math time?*) students answered prior to answering the ISE1 items. Responses

were approximately normally distributed, with skewness ranging from -1.55 to 0.50 and kurtosis values ranging from -1.11 to 1.41 (Table 19).

Table 19

Student Responses for the Mathematics Student Fidelity of Implementation Questionnaire Instructional Student Engagement Domain

Subscale **Item**		***Never or Hardly Ever (1)***	***Sometimes (2)***	**A lot (3)**
	N	**%**	**%**	**%**
Students Contribute to Small Group Work (ISE1)	5430			
When we work in math groups, we work as a team. (1a)		3.3	39.9	56.8
During math time, I learn from other students when working in groups. (1b)		6.6	46.7	46.6
When we do group work in math, I cooperate with other students.(1c)		3.8	34.1	62.1
Students Engage in Discussion (ISE2)	5964			
I talk to other students about our math work. (2a)		17.4	60.5	22.1
Students talk with each other about what we're learning during math time. (2b)		19.2	56.0	24.8
During math time, I talk to my teacher about what we are learning. (2c)		25.8	52.2	22.0
I am a good listener when my classmates are talking during math time. (2d)		3.8	32.7	63.4
Students Engage in Cognitively Demanding Work (ISE3)	5955			
During math time, I explain how I get my answer. (3a)		5.2	46.7	48.1
When I come up with an answer in math class, I make sure that it makes sense. (3b)		2.3	29.2	68.6
I explain why I agree or disagree with things my classmates say in math. (3c)		10.4	50.8	38.8
During math time, I work hard to understand a lesson. (3d)		1.7	22.6	75.7
Students Take Risks	5935			
When working on math problems, I am willing to try something new or different. (4a)		4.7	41.9	53.5
I say what I think in math even if it's different from other students. (4b)		9.0	50.7	40.2
During math time, I ask questions when I am confused. (4c)		7.0	43.7	49.3
I am not embarrassed to answer questions during math time. (4d)		18.7	39.2	42.1

Cronbach's alphas for the four scales described in Table 20, not taking into account the multilevel data structure were .46, .55, .57, and .48, respectively (Table 21). Given the multilevel nature of these data, these Cronbach's alphas represent a first look at the reliability of

the data. Further below under the section entitled Multilevel Confirmatory Factor Analysis, the reliabilities are computed using the ICCs with the Spearman-Brown formula for the mathematics sample of students nested within teachers.

In order to assess whether significant differences in the mean ISE scores existed between students who had teacher ID's and students without teacher ID's (TIDs) an independent-samples t-test was conducted. For Students Contribute to Small Group Work (ISE1), there was not a significant difference in scores for students with TIDs (M=2.51, SD=0.40), and students without TIDs (M=2.50, SD=0.40; t[5428]=1.44, p=.15). The magnitude of the differences in the means was very small (eta squared = .000). For Students Engage in Discussion (ISE2), there was a significant difference in scores for students with TIDs (M=2.19, SD=0.42), and students without TIDs (M=2.14, SD=0.41; t[5962]=3.86, p=.00). The magnitude of the differences in the means was very small (eta squared = .002). For Students Engage in Cognitively Demanding Work (ISE3), there was a significant difference in scores for students with TIDs (M=2.55, SD=0.37), and students without TIDs (M=2.51, SD=0.37; t[55953]= 4.28, p=.00). The magnitude of the differences in the means was very small (eta squared = .003). For Students Take Risks (ISE4), there was a significant difference in scores for students with TIDs (M=2.38, SD=0.40), and students without TIDs (M=2.34, SD=0.41; t[5933]=4.09, p=.00). The magnitude of the differences in the means was very small (eta squared = .003).

Table 20

Item Descriptives for the Mathematics Student Fidelity of Implementation Questionnaire Instructional Student Engagement Domain

Subscale Item	*N*	Number of Missing Cases	*M*	*SD*	Skewness	Kurtosis	ICC
Students Contribute to Small Group Work (ISE1)							
When we work in math groups, we work as a team. (1a)	5430	561	2.54	0.56	-0.70	-0.56	.05
During math time, I learn from other students when working in groups. (1b)	5430	561	2.40	0.61	-0.49	-0.64	.10
When we do group work in math, I cooperate with other students. (1c)	5430	561	2.58	0.57	-0.96	-0.09	.08
Students Engage in Discussion (ISE2)							
I talk to other students about our math work. (2a)	5964	27	2.05	0.63	-0.04	-0.46	.20
Students talk with each other about what we're learning during math time. (2b)	5964	27	2.06	0.66	-0.06	-0.72	.15
During math time, I talk to my teacher about what we are learning. (2c)	5964	27	1.96	0.69	0.05	-0.90	.14
I am a good listener when my classmates are talking during math time. (2d)	5964	27	2.60	0.56	-1.02	0.04	.07
Students Engage in Cognitively Demanding Work (ISE3)							
During math time, I explain how I get my answer. (3a)	5955	36	2.43	0.59	-0.48	-0.67	.08
When I come up with an answer in math class, I make sure that it makes sense. (3b)	5955	36	2.66	0.52	0.10	0.30	.05
I explain why I agree or disagree with things my classmates say in math. (3c)	5955	36	2.28	0.64	0.50	-0.71	.09
During math time, I work hard to understand a lesson. (3d)	5955	36	2.74	0.48	-1.55	1.41	.04

Table 20 (continued)

Subscale Item	*N*	Number of Missing Cases	*M*	*SD*	Skewness	Kurtosis	ICC
Students Take Risks							
When working on math problems, I am willing to try something new or different. (4a)	5935	56	2.49	0.59	-0.65	-0.54	.06
I say what I think in math even if it's different from other students. (4b)	5935	56	2.31	0.63	-0.36	-0.68	.05
During math time, I ask questions when I am confused. (4c)	5935	56	2.42	0.62	-0.59	-0.59	.06
I am not embarrassed to answer questions during math time. (4d)	5935	56	2.23	0.74	-0.41	-1.11	.02

Note. ICC = Intraclass correlation coefficient. ICC's are reported only for the sample of students who had a teacher ID (N = 3096). Response scale ranges from 1(*never or hardly ever*) to 3 (*a lot*).

Table 21

Internal Consistency of Instructional Student Engagement Subscales (Cronbach's α) for Mathematics

Scale	**# of Items**	**Cronbach's α**	**N**	**Item-to-Total Correlation Range**
Students Contribute to Small Group Work (ISE1)	3	.46	5430	.24 to .32
Students Engage in Discussion (ISE2)	4	.55	5964	.12 to .44
Students Engage in Cognitively Demanding Work (ISE3)	4	.57	5955	.31 to.39
Students Take Risks (ISE4)	4	.48	5935	.24 to .32

Confirmatory Factor Analysis for the Mathematics Instructional Student Engagement Model

Confirmatory factor analysis with corrected standard errors for nested data.

As noted in the previous section, multilevel confirmatory factor analyses (MCFA) models can be complex, so simpler models are recommended as a preliminary step. Therefore, before running the MCFA, I examined the factor structure using a single-level CFA with robust weighted least squares (WLS) approach (estimator = WLSMV in Mplus) and standard errors adjusted to take into account cluster sampling (i.e., nested data) to examine the four-factor measurement model underlying the Instructional Student Engagement domain. The data were clustered by teacher ID. In order to take into account the nested data structure (i.e., student data nested within teachers), it was necessary for the student to have an associated teacher ID. Students without a teacher ID were eliminated from this analysis and later for the multilevel analyses. The single level CFA does not take into account the two-level structure of the data; it is based on the total polychoric correlation matrix of the observed variables (i.e., the total polychoric correlation matrix is not decomposed into between and within matrices, which is the case for the MCFA).

The chi-square value for the single level, four-factor CFA model, X^2 (84, N=3096) = 955.98, $p< .05$, indicated a statistically significant lack of fit. Alternative measures of fit, which are less sensitive to sample size, were mixed with the RMSEA (.06) indicating acceptable fit, and the CFI of .89 indicating less than acceptable fit.

A single level, four-factor CFA for students without TIDs was also run to examine if differences existed. The model fit indices for the Student CFA models with TIDs can be found in Table 13 and the model fit indices for the Student CFA models without TIDs can be found in

Table 14. As can be seen in the tables the models fit pretty similarly for both students with TIDs and students without TIDs.

All factor pattern coefficients (loadings) were significantly different from zero ($p< .05$). The standardized loadings for the items within the ISE1 factor (students contribute to small group work) ranged from .51 to .60, from .46 to .72 for ISE2 (students engage in discussion), from .60 to .64 for ISE3 (students engage in cognitively demanding work) and from .32 to .58 for ISE4 (students take risks). The correlations between the factors were positive and significantly different from zero ($p< .05$) with ISE1 and ISE2, 1SE1 and 1SE3, and ISE1 and ISE4 correlating at .76, .84, and .84, respectively, and ISE2 and ISE3, ISE2 and ISE4, ISE3 and ISE4 correlating at .72, .68, and .90, respectively.

An alternative one-factor model was also considered. This model did not fit as well as the four-factor model based on the chi-square value, X^2 (90, N=3096) = 1191.52, $p<.05$, and the other fit indices (RMSEA=.06 and CFI=.86). Standardized item loadings on the one-factor model ranged from .29 to .62.

Given that students were nested within teachers, thus violating the independence assumption, multilevel confirmatory factor analysis was used to further analyze the data for this study.

Multilevel Confirmatory Factor Analysis for the Mathematics Instructional Student Engagement Model

Prior to conducting the MCFA, the variability between and within teachers on each item was examined by computing the intra-class correlations (ICCs) for each of the 15 items in the Instructional Student Engagement domain. Table 10 displays the ICCs for these 15 items. The ICCs for each of the observed items ranged from .02 (for item ISE4d within the ISE4 factor) to

.20 (for item ISE2a within the ISE2 factor). These values indicated that there was sufficient between teacher variability to warrant multilevel analysis.

As shown in Figure 3, a four-factor multilevel model, in which the same number of factors at each level (4 within factors and 4 between factors) was run. Results of the four-factor multilevel model with loadings freely estimated across levels indicated mixed results. The RMSEA was .04 and the CFI was .85. The SRMR fit indices at each level indicated that the fit of the level-1 (within or student) part of the model was better than at level-2 (between or teacher; SRMR within= .05 vs. SRMR between= .21; see Table 15 for measures of fit).

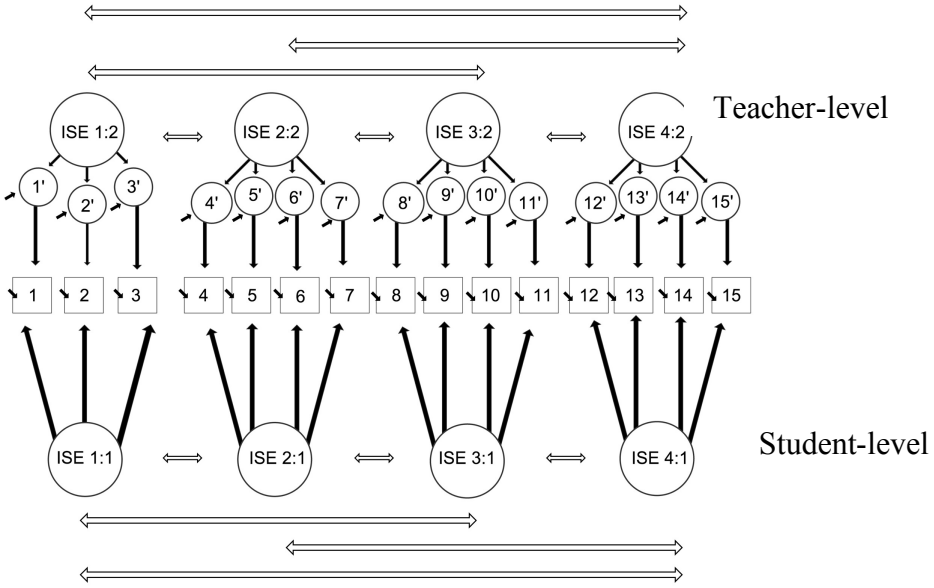


Figure 3. Four-Factor Multilevel Confirmatory Factor Analysis model for Instructional Student Engagement in Mathematics

At level-1 (student) all factor pattern coefficients (loadings) were significantly different from zero ($p < .05$). At level-2 (teacher) all factor pattern coefficients were also significantly

different from zero ($p < .05$) except for item 4d ($p = .50$). See Table 22 for the unstandardized factor loadings.

In MCFA, fixing residual variances to zero at the between level to zero is often necessary when sample sizes at level-2 are small and the true between-group variance is close to zero (Hox, 2002). In the case of ISE for mathematics, the residual variances for the level-2 intercepts were fixed to zero for item 4b only.

Table 22

Multilevel Confirmatory Factor Analysis: Unstandardized Factor Loadings and Residual Variances for the Four-Factor Model Underlying Student Ratings of Instructional Student Engagement

	Students with a TID (*N*=3096)	**Teachers (*N*= 152)**	
Item on the Rubric	**Factor Loading**	**Factor Loading**	**Residual Variances**
Students Contribute to Small Group Work			
1a	1.00[a] (--)	1.00[a] (--)	0.04 (0.02)
1b	1.04 (0.08)	2.18 (0.34)	0.00 (0.03)
1c	1.22(0.10)	0.68 (0.22)	0.12 (0.03)
Student Engage in Discussion			
2a	1.00[a] (--)	1.00[a] (--)	0.05 (0.03)
2b	0.85(0.06)	0.73 (0.07)	0.07 (0.20)
2c	0.95 (0.07)	0.68 (0.08)	0.09 (0.02)
2d	0.69 (0.06)	0.19 (0.07)	0.08 (0.02)
Student Engage in Cognitively Demanding Work			
3a	1.00[a] (--)	1.00[a] (--)	0.03 (0.02)
3b	1.02 (0.07)	0.50 (0.11)	0.06 (0.02)
3c	1.01 (0.07)	1.22 (0.16)	0.00 (0.03)
3d	1.01 (0.07)	0.35 (0.13)	0.05 (0.02)
Students Take Risks			
4a	1.00[a] (--)	1.00[a] (--)	0.05 (0.02)
4b	0.99 (0.07)	1.17 (0.17)	0.00[b] (-)
4c	0.89 (0.06)	0.75 (0.15)	0.05 (0.02)
4d	0.53 (0.05)	0.08 (0.12)	0.00 (0.03)

Note. Numbers in parentheses represent the standard error.
[a]Factor loading fixed to 1.0
[b] Residual variances were fixed to 0.

Inter-factor correlations were .76 ($p< .05$) between ISE1 and ISE2 at level-1 and .92 ($p< .05$) at level-2; .81 ($p< .05$) between ISE1 and ISE3 at level-1 and .83 ($p< .05$) at level-2; .83 ($p<.05$) between ISE1 and ISE4 at level-1 and .87 ($p< .05$) at level-2; .71 ($p< .05$) between ISE2 and ISE4 at level-1 and .78 ($p< .05$) at level-2; 75 ($p< .05$) between ISE2 and ISE3 at level-1 and .67 ($p< .05$) at level-2; and .90 ($p< .05$) between ISE3 and ISE4 at level-1 and .79 ($p< .05$) at level-2.

Multilevel ICCs and Reliability

Using this model, it was possible to calculate the ICCs for the four latent variables and, subsequently, the reliability of each factor when aggregated at the teacher level. The ICC is the variation between teachers divided by the total variation. Total variation equals the combined within-and between- teacher variation. ISE2 had the greatest amount of between teacher variability (ICC= .37), followed by ISE3 (ICC= .16), then ISE4 (ICC=.10) and ISE1 (ICC= .08). Using these ICCs with the Spearman-Brown formula, [k(ICC)/ [(k-1)(ICC) +1], where k is the average number of students nested within teachers , the estimated reliabilities for the factors in this study, with an average cluster size of 20 respondents (students) per teacher, were .92 for ISE2, .79 for ISE3, .69 for ISE4 and .62 for ISE1. See Tables 36 and 37 at the end of this chapter for summary tables of internal consistency results by level.

Confirmatory Factor Analysis for the Mathematics Instructional Student Engagement Teacher Model

In this section, the model fit based on teachers' self-reported data (rather than students' reports nested within teachers) is presented in Figure 4. The chi-square value for the single level,

four-factor CFA model, X^2 (146, N=152) = 295.38, p< .05, indicated a statistically significant lack of fit. However, alternative measures of fit, which are less sensitive to sample size, suggested that the fit was marginally acceptable. The RMSEA of .08 was slightly greater than Hu and Bentler's (1999) cutoff of .06, and the CFI of .94 was just slightly lower than the .95 cutoff values for this index.

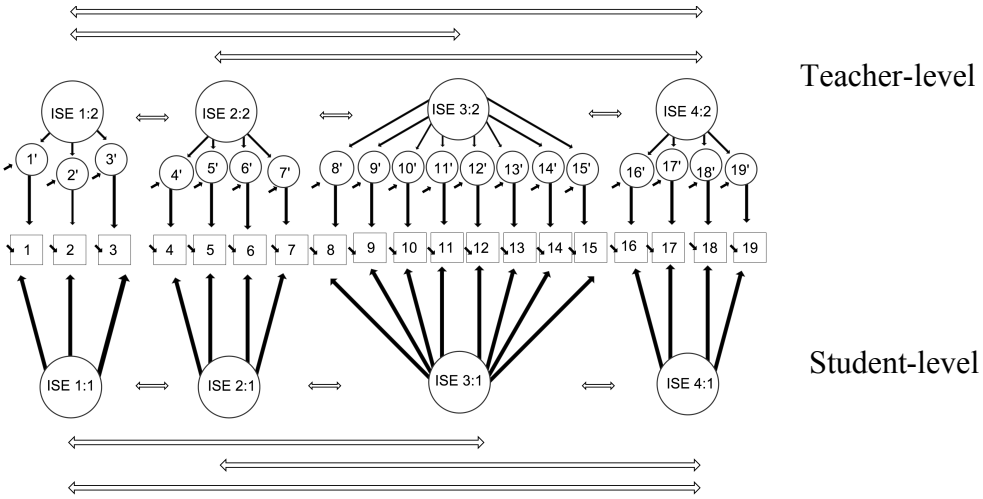


Figure 4. Four-Factor Multilevel Confirmatory Factor Analysis Model for Instructional Student Engagement in Mathematics

All factor pattern coefficients (loadings) were significantly different from zero (p< .05). The standardized loadings for the items within the ISE1 factor (students contribute to small group work) ranged from .60 to .74, from .50 to .82 for ISE2 (students engage in discussion), from .59 to .88 for ISE3 (students engage in cognitively demanding work), and from .61 to .77 for ISE4 (students take risks). See Table 23 for the unstandardized factor loadings. The correlations between the factors were positive and significantly different from zero (p< .05) with ISE1 and ISE2, 1SE1 and 1SE3, and ISE1 and ISE4 correlating at .81, .73, and .60, respectively, and ISE2 and ISE3, ISE2 and ISE4, ISE3 and ISE4 correlating at .72, .63, and .56, respectively.

In order to examine convergent validity, meaning the correlation between students' and teachers' responses on the Instructional Student Engagement domain, the factor scores from the student perspective were correlated with the factor scores from the teacher perspective. The dataset consisted of 3,096 students (level-1) nested within 152 teachers of which all students had one teacher (level-2). Each of the 3,096 students provided data on instructional student engagement from their perspective. These data constituted the lower-level (level-1) unit of analysis in this study. The second-level data included class instructional student engagement scores for each of the 152 teachers.

Preliminary analyses were conducted using the observed variables in SPSS. The student data were aggregated to create a teacher mean, as were the teacher data (although given that teacher responses for each student in a class were the same, the mean was the same as the teachers reported response). The correlations based on the observed variables between teacher and students on the Instructional Student Engagement components of Students Contribute to Small Group Work, Students Engage in Discussion, Students Engage in Cognitively Demanding Work, and Students Take Risks were .03, .23, .07, and .18, respectively. Following that, the data were examined in M*plus* by estimating the correlation of the latent variables, taking into account the two-level framework (Figure 5).

The correlations between the teachers' and students' scores on the Instructional Student Engagement components of Students Contribute to Small Group Work, Students Engage in Discussion, Students Engage in Cognitively Demanding Work, and Students Take Risks were -.07, .28, .20, and .41, respectively. See Table 24 for the teacher and student correlations on the Instructional Pedagogical Domain.

Table 23

Confirmatory Factor Analysis: Unstandardized Factor Loadings for the Four-Factor Model Underlying Teacher Ratings of Instructional Student Engagement

	Teachers with IDs (*N*=152)
Item on the Rubric	**Factor Loading**
Students Contribute to Small Group Work	
1a	1.00[a] (--)
1b	1.24 (0.18)
1c	1.25 (0.17)
Student Engage in Discussion	
2a	1.00[a] (--)
2b	0.83 (0.08)
2c	0.61 (0.09)
2d	0.92 (0.08)
Student Engage in Cognitively Demanding Work	
3a	1.00[a] (--)
3b	1.02 (0.09)
3c	1.04 (0.09)
3d	0.85 (0.08)
3e	1.16 (0.08)
3f	0.91 (0.08)
3g	1.26 (0.08)
3h	1.26 (0.08)
Students Take Risks	
4a	1.00[a] (--)
4b	1.12 (0.15)
4c	1.23 (0.15)
4d	1.17 (0.14)

Note. Numbers in parentheses represent the standard error.
[a]Factor loading fixed to 1.0

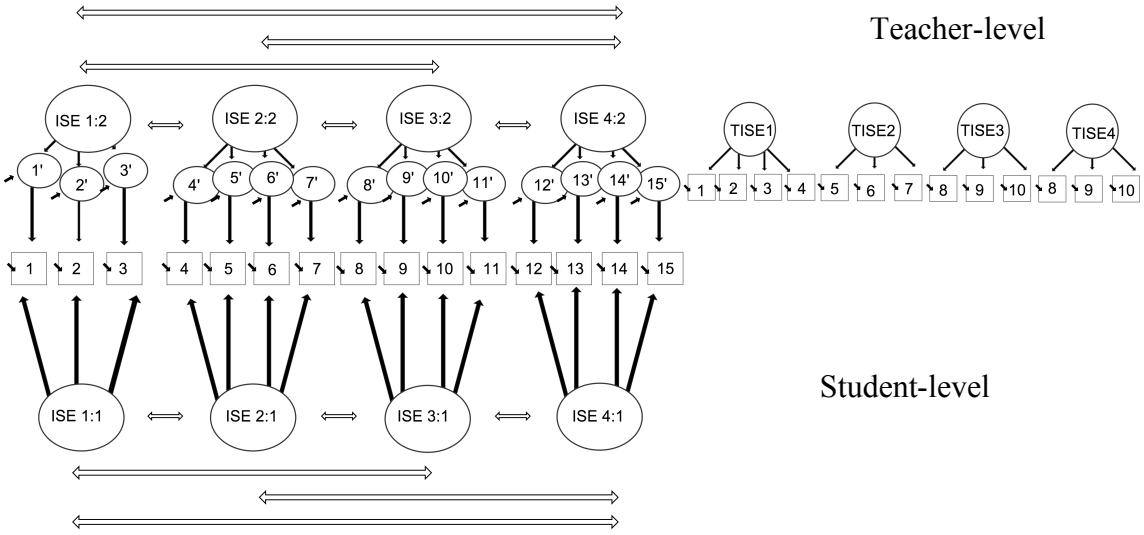


Figure 5. Multilevel Confirmatory Factor Analysis Teacher and Student Model for Instructional Student Engagement Convergent Validity

Table 24

Correlations of Instructional Student Engagement Subscales from Student Questionnaire Compared with Teacher Questionnaire Based on the Multilevel Confirmatory Factor Analysis Model

		Teacher Questionnaire (*N*=152)			
	Scale	**ISE1**	**ISE2**	**ISE3**	**ISE4**
Student Questionnaire (*N*=3096)	Students Contribute to Small Group Work (ISE1)	-.07			
	Students Engage in Discussion (ISE2)		.28		
	Students Engage in Cognitively Demanding Work (ISE3)			.20	
	Students Take Risks (ISE4)				.41

Science Student and Teacher Demographics

For science, there were 4,410 students in the sample. Of those students, 50.5% were boys. The sample was ethnically diverse, in that students came from a range of ethnicities. Whites were the largest ethnicity at 26.1%, followed by Hispanics at 22.9%; 22.3% of students

identified themselves as Other, 10.7% identified themselves as Mixed, and 9.3% identified themselves as African American/Black. Students participating in this study were in grades 3-5, with 36.1% of students in the 3rd grade, 33.4% in 4th grade, and 30.5% in 5th grade. The mean age for students in this sample was 9 years of age (ranging from 7-12 years). Students came from 41 schools across the three districts in the sample. Science students predominately came from the Denver district (52.6%), followed by the Stamford district (30.7%), and then the Kirby district (16.7%).

For the 90 science teachers analyzed in this sample, gender, age and ethnicity were not requested demographics and so are not reported here. The majority of science teachers held a master's degree (70.0%), followed by a bachelor's degree (28.9%), and few had a doctoral degree (1.1%). Only 6.7% of these teachers had a degree in Science or Science Education and only 2.2% were a science specialist/coach. In terms of years of teaching experience, science teachers' experience ranged from 3.3% for one year of experience to 12.2% for teachers who had 25 or more years of experience. Teachers who had three years of experience followed (10%). Science teachers primarily taught 3rd grade (43.3%), followed by 4th grade (30.0%), and then 5th grade (24.4%).

For the 77 science teachers that were not analyzed in this study, the majority of science teachers held a master's degree (66.5%), followed by a bachelor's degree (32.3%), and few had a doctoral degree (1.2%). Only 3.5% of these teachers had a degree in Science or Science Education and of these teachers 100% were a science specialist/coach. In terms of years of teaching experience, science teachers' experience ranged from 3.2% for one year of experience to 10.7% for teachers who had 25 or more years of experience. Science teachers primarily taught 3rd grade (41.4%), followed by 4th grade (32.5%), and then 5th grade (24.0%).

Instrument, Item Descriptives, and Reliability Assessment

Item means ranged from 1.55 (*SD* = 0.66) for 'doing work different from other students' (teacher use of differentiation) to 2.64 (*SD*= 0.55) for 'my teacher makes science interesting (teacher facilitation of student interest), with sample sizes for the items varying from 4,408 for teacher facilitation of student interest, and teacher facilitation of student discussion to 4,410 for teacher use of differentiation. Less than 1.0% (0.05%) of the participants in the Science sample were missing. Responses were approximately normally distributed, with skewness ranging from -1.20 to 0.81 and kurtosis values ranging from -0.91 to 0.46. Descriptive statistics for the items and scales can be found in Table 25 and responses to items can be found in Table 26.

Cronbach's alphas for the three scales described in Table 25, not taking into account the multilevel data structure, were .68, .62, and .62 respectively (Table 27). Given the multilevel nature of this data, these Cronbach's alphas represent a first look at the reliability of the data. Further below under the section entitled Multilevel Confirmatory Factor Analysis, the reliabilities are computed using the ICCs with the Spearman-Brown formula for the science sample of students nested within teachers.

In order to assess whether significant differences in the mean IP scores existed between students who had teacher ID's and students without teacher ID's (TIDs) an independent-samples t-test was conducted. For Teacher Facilitation of Student Discussion (IP2), there was not a significant difference in scores for students with TIDs (*M*=2.32, *SD*=0.45), and students without TIDs (*M*=2.34, *SD*=0.46; *t*[4408]=-1.30, *p*=.19). The magnitude of the differences in the means was very small (eta squared = .000). For Teacher Facilitation of Student Interest (IP7), there was

a significant difference in scores for students with TIDs (*M*=2.57, SD=0.42), and students without TIDs (*M*=2.53, *SD*=0.46; *t*[4393.09]=2.99, *p*=.00). The magnitude of the differences in the means was very small (eta squared = .002). For Teacher Use of Differentiation (IP10), there was not a significant difference in scores for students with TIDs (*M*=1.56, *SD*=0.49), and students without TIDs (*M*=1.55, *SD*=0.47; *t*[4406]=0.87, *p*=.38). The magnitude of the differences in the means was very small (eta squared = .000).

Table 25

Item Descriptives for the Science Student Fidelity of Implementation Questionnaire Instructional Pedagogical Domain

Subscale Item	***N***	**Number of Missing Cases**	***M***	***SD***	**Skewness**	**Kurtosis**	**ICC**
Teacher Facilitation of Student Discussion (IP2)							
My teacher asks us questions during science time. (2a)	4410	1	2.59	0.55	-0.92	-0.19	.07
My teacher wants us all to share ideas during science time. (2b)	4410	1	2.42	0.62	-0.57	-0.60	.14
My teacher asks me to talk to my classmates about their science ideas. (2c)	4410	1	2.20	0.67	-0.25	-0.80	.16
My teacher gives me the chance to talk to my classmates about my science schoolwork. (2d)	4410	1	2.13	0.69	-0.18	-0.91	.13

Table 25 (continued)

Subscale Item	*N*	Number of Missing Cases	*M*	*SD*	Skewness	Kurtosis	ICC
Teacher Facilitation of Student Interest (IP7)							
My teacher makes science interesting. (7a)	4410	1	2.64	0.55	-1.20	0.46	.13
My teacher tells us how things we learn in science can be used in the real world. (7b)	4410	1	2.42	0.63	-0.61	-0.58	.07
My teacher does things that make me like science. (7c)	4410	1	2.59	0.59	-1.09	0.17	.08
Teacher Use of Differentiation (IP10)							
All students in my science class do the same work at the same time. (10a-reverse coded)	4408	3	2.55	0.58	-0.88	-0.22	.03
During science time, some students do different work than others. (10b)	4408	3	1.67	0.67	0.49	-0.75	.06
During science time, I do work that is different from what other students are doing. (10c)	4408	3	1.55	0.66	0.81	-0.46	.06

Note. ICC = Intra-class correlation coefficient. ICCs are reported only for the sample of students who had a teacher ID (*N*= 2023). Response scale ranged from 1 (*Never or Hardly Ever*) to 3 (*A Lot*).

Table 26

Student Responses for the Science Student Fidelity of Implementation Questionnaire Instructional Pedagogical Domain

Subscale **Item**		***Never or Hardly Ever (1)***	***Sometimes (2)***	**A lot (3)**
	N	***%***	***%***	**%**
Teacher Facilitation of Student Discussion (IP2)	4410			
My teacher asks us questions during science time. (2a)		3.2	34.8	62.0
My teacher wants us all to share ideas during science time. (2b)		7.0	44.1	48.8
My teacher asks me to talk to my classmates about their science ideas. (2c)		14.4	51.4	34.2
My teacher gives me the chance to talk to my classmates about my science schoolwork. (2d)		18.1	50.7	31.2
Teacher Facilitation of Student Interest (IP7)	4410			
My teacher makes science interesting. (7a)		3.6	29.1	67.3
My teacher tells us how things we learn in science can be used in the real world. (7b)		7.4	43.0	49.6
My teacher does things that make me like science. (7c)		5.0	31.4	63.6
Teacher Use of Differentiation (IP10)	4408			
All students in my science class do the same work at the same time. (10a-reverse coded)		59.3	36.1	4.6
During science time, some students do different work than others. (10b)		43.9	45.0	11.0
During science time, I do work that is different from what other students are doing. (10c)		54.7	35.8	9.5

Table 27

Internal Consistency of Instructional Pedagogical Subscales (Cronbach's α) for Science

Scale	# of Items	Cronbach's α	N	Item-to-Total Correlation Range
Teacher Facilitation of Student Discussion (IP2)	4	.68	4410	.32 to .53
Teacher Facilitation of Student Interest (IP7)	3	.62	4410	.30 to .50
Teacher Use of Differentiation (IP10)	3	.73	4408	.25 to.55

Confirmatory Factor Analysis for the Science Instructional Pedagogical Student Model

Confirmatory Factor Analyses (CFA) and Multilevel Confirmatory Factor analyses (MCFA) were conducted using Mplus Version 7 (Muthen & Muthen, 1998-2014). As mentioned previously, at the beginning of the Math section, a categorical approach was used for the analyses and the overall goodness of fit for the models were evaluated using multiple fit indices.

CFA with corrected standard errors for nested data.

Prior to running the MCFA, I examined the factor structure using a single-level CFA with robust weighted least squares (WLS) approach (estimator = WLSMV in Mplus) and standard errors adjusted to take into account cluster sampling (i.e., nested data) to examine the three-factor measurement model underlying the Instructional Pedagogical domain. The data were clustered by teacher ID. In order to take into account the nested data structure (i.e., student data nested within teachers), it was necessary for the student to have an associated teacher ID. Students without a teacher ID were eliminated from this analysis and later for the multilevel analyses. The single level CFA does not take into account the two-level structure of the data; it is based on the total covariance matrix of the observed variables (i.e., the total covariance matrix

is not decomposed into between and within covariance matrices, which is the case for the MCFA).

The chi-square value for the single level, three factor CFA model, X^2 (32, N=2023) = 352.497, $p< .05$, indicated a statistically significant lack of fit. Alternative measures of fit, which are less sensitive to sample size, suggested marginally acceptable fit. The RMSEA of .07 was slightly greater than Hu and Bentler's (1999) cutoff of .06 and the CFI of .93 was slightly less than the .95 cutoff value for this index. A single level, three factor CFA for students without TIDs was also run to examine if differences existed. The model fit indices for the Student CFA models with TIDs can be found in Table 13 and the model fit indices for the Student CFA models without TIDs can be found in Table 14. As can be seen in the tables the models fit pretty similarly for both students with TIDs and students without TIDs.

All factor pattern coefficients (loadings) were significantly different from zero (p<.05). The standardized loadings for the items within the IP2 factor (teacher facilitation of student discussion) ranged from .46 to .75, from .55 to .75 for IP7 (teacher facilitation of student interest), and from .39 to .90 for IP10 (teacher use of differentiation). The correlations between the factors were positive and significantly different from zero (p<.05) for IP2 and IP7 (.56) , 1P2 and 1P10 (.12), and IP7 and IP10 (.07).

An alternative one-factor model was also considered. This model did not fit as well as the three factor model based on the chi-square value, X^2 (35, N=2023) = 2356.25, $p<.05$, and the other fit indices (RMSEA=.18, and CFI=.47). Standardized item loadings on the one-factor model ranged from .04 to .68.

Given that students were nested within teachers, thus violating the independence assumption, multilevel confirmatory factor analysis was used to further analyze the data for this study.

Multilevel Confirmatory Factor Analysis for the Science Instructional Pedagogical Student Model

Prior to conducting the MCFA, the variability between and within teachers on each item was examined by computing the intra-class correlations (ICCs) for each of the 10 items in the Instructional Pedagogical domain. Table 25 displays the ICCs for the 10 items in the Instructional Pedagogical domain for science. The ICCs for each of the observed items ranged from .03 (for item IP10a within the IP10 factor) to .16 (for item IP2c also within the IP2 factor). These values indicated that there was sufficient between teacher variability to warrant multilevel analysis.

As shown Figure 6, a three-factor multilevel model, in which the same number of factors at each level was run (3 within factors and 3 between factors). Results of the three-factor multilevel model with loadings freely estimated across levels indicated a reasonable fit of the model to the data. The RMSEA of .05 and CFI of .91 indicated reasonable fit overall. The SRMR fit indices at each level indicated that the fit of the level-1 (within) part of the model was better than at level-2 (SRMR within= .07 vs. SRMR between= .21; see Table 15 for measures of fit).

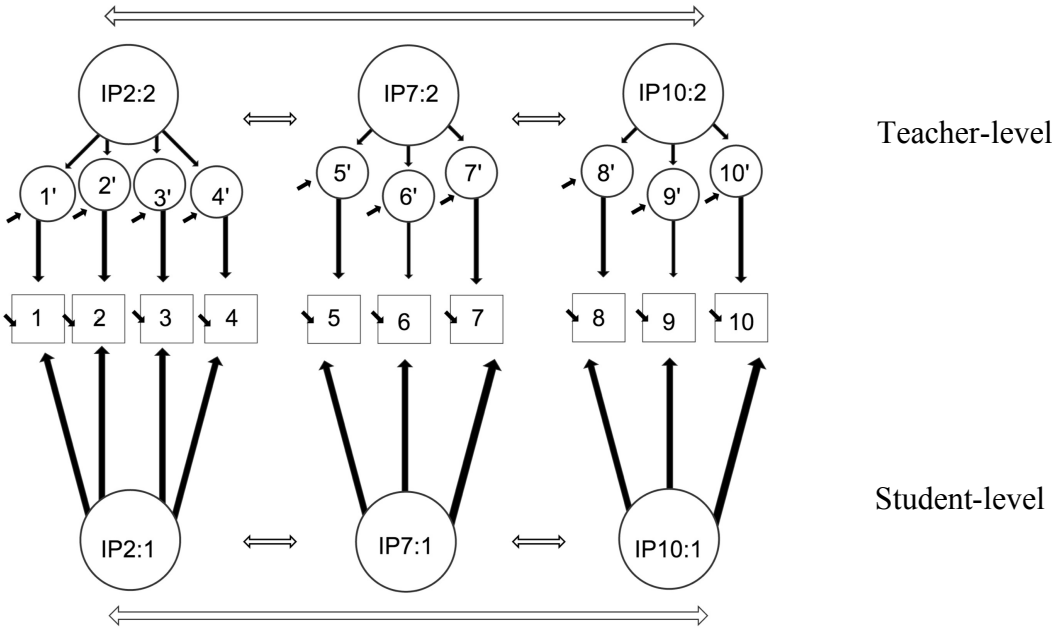


Figure 6. Three-Factor Multilevel Confirmatory Factor Analysis Model for Instructional Pedagogical in Science

At level-1 (student) all factor pattern coefficients (loadings) were significantly different from zero ($p < .05$). At Level 2 (teacher), all factor pattern coefficients (loadings) were significantly different from zero ($p < .05$), except for item 2a ($p = .20$). See Table 28 for the unstandardized factor loadings.

In MCFA, fixing residual variances to zero at the between level to zero is often necessary when sample sizes at level-2 are small and the true between-group variance is close to zero (Hox, 2002). In the case of IP for science, the residual variances for the Level 2 intercepts were fixed to zero for items 7a and 10c.

Table 28

Multilevel Confirmatory Factor Analysis: Unstandardized Factor Loadings and Residual Variances for the Three-Factor Model Underlying Student Ratings of Instructional Pedagogy

	Students with TID (*N*=2023)	**Teachers (*N*= 90)**	
Item on the Rubric	**Factor Loading**	**Factor Loading**	**Residual Variances**
Teacher Facilitation of Student Discussion			
2a	1.00[a] (--)	1.00[a] (--)	0.09 (0.03)
2b	1.91 (0.16)	7.11 (5.46)	0.10 (0.04)
2c	1.67 (0.15)	8.65 (6.91)	0.00 (0.04)
2d	1.67 (0.15)	7.69 (6.29)	0.00 (0.04)
Teacher Facilitation of Student Interest			
7a	1.00[a] (--)	1.00[a] (--)	0.00[b] (-)
7b	0.53 (0.05)	0.35 (0.12)	0.07 (0.03)
7c	0.99 (0.13)	0.69 (0.12)	0.05 (0.03)
Teacher Use of Differentiation			
10a	1.00[a] (--)	1.00[a] (--)	0.01 (0.01)
10b	4.82 (0.95)	2.82 (0.93)	0.04 (0.06)
10c	3.55 (0.49)	2.49 (0.76)	0.00[b] (-)

Note. Numbers in parentheses represent the standard error.
[a]Factor loading fixed to 1.0
[b] Residual variances were fixed to 0.

Inter-factor correlations were .58 (p<.05) between IP2 and IP7 at Level 1 and .15 (p=.23, not statistically significant) at Level 2; .11 (p<.05) between IP2 and IP10 at Level 1 and .29 (p<.05) at Level 2; and -.06 (p=.07, not statistically significant) between IP7 and IP10 at Level 1 and -.36 (p<.05) at Level 2.

Multilevel ICCs and Reliability

Using this model, it was possible to calculate the ICCs for the three latent variables and, subsequently, the reliability of each factor when aggregated at the teacher level. The ICC is the variation between teachers divided by the total variation. Total variation equals the combined within-and between- teacher variation. IP7 had the greatest amount of between teacher variability (ICC= .21), followed by IP10 (ICC= .16), and IP2 (ICC= .02). Using these ICCs with

the Spearman-Brown formula, [*k*(ICC)/ [(k-1)(ICC) +1], where k is the average number of students nested within teachers , the estimated reliabilities for the factors in this study, with an average cluster size of 22 respondents (students) per teacher, were .85 for IP7, .81 for IP10, and .31 for IP2. See Tables 36 and 37 at the end of this chapter for summary tables of internal consistency results by level.

Confirmatory Factor Analysis for the Science Instructional Pedagogical Teacher Model

In this section, the model fit based on teachers' self-reported data (rather than students' reports nested within teachers) is presented. The chi-square value for the single level, three-factor CFA model, X^2 (32, N=90) = 41.17, p< .05, indicated a statistically significant lack of fit. However, alternative measures of fit, which are less sensitive to sample size, suggested good fit. The RMSEA of .06 and the CFI of .99 were within the values for their respective indices.

All factor pattern coefficients (loadings) were significantly different from zero ($p< .05$). The standardized loadings for the items within the IP2 factor (teacher facilitation of student discussion) ranged from .71 to .92, from .79 to .92 for IP7 (teacher facilitation of student interest), and from .48 to .84 for IP10 (teacher use of differentiation). See Table 29 for the unstandardized factor loadings. The correlations between the factors were positive and significantly different from zero ($p< .05$) with IP2 and IP7, 1P2 and 1P10, and IP7 and IP10 correlating at .49, .42, and .50, respectively.

Table 29

Multilevel Confirmatory Factor Analysis: Unstandardized Factor Loadings and Residual Variances for the Three-Factor Model Underlying Student Ratings of Instructional Pedagogy

	Teachers with IDs (*N*=90)
Item on the Rubric	**Factor Loading**
Teacher Facilitation of Student Discussion	
2a	1.00[a] (--)
2b	0.95 (0.08)
2c	0.77 (0.09)
2d	0.93 (0.06)
Teacher Facilitation of Student Interest	
7a	1.00[a] (--)
7b	1.03 (0.08)
7c	0.88 (0.06)
Teacher Use of Differentiation	
10a	1.00[a] (--)
10b	1.02 (0.18)
10c	0.59 (0.14)

Convergent Validity

In order to examine convergent validity, meaning the correlation between student and teachers responses on the Instructional Pedagogical domain, the factor scores from the student perspective were correlated with the factor scores from the teacher perspective. The dataset consisted of 2,023 students (level-1) nested within 90 teachers of which all students had one teacher (level-2). Each of the 2,023 students provided data on instructional pedagogy from their perspective. These data constituted the lower-level (level-1) unit of analysis in this study. The second-level data included class instructional pedagogy scores for each of the 90 teachers. Data regarding instructional pedagogy were gathered from two sources: from the teachers (self-ascribed instructional pedagogy) and their students (perceived instructional pedagogy).

Preliminary analyses were conducted using the observed variables in SPSS. The student data were aggregated to create a teacher mean, as were the teacher data (although given that teacher responses for each student in a class were the same, the mean was the same as the teachers reported response). The correlations based on the observed variables between teacher and students on instructional pedagogical components Teacher Facilitation of Student Discussion, Teacher Facilitation of Student Interest and Teacher Use of Differentiation were .02, .10, and .15, respectively. Following that, the data were examined in M*plus* by estimating the correlation of the latent variables, taking into account the two-level framework (Figure 2).

The correlations between teacher and students on instructional pedagogical components Teacher Facilitation of Student Discussion, Teacher Facilitation of Student Interest and Teacher Use of Differentiation were .06, -.15, and .16, respectively. See Table 30 for Teacher and Student Correlations on the Instructional Pedagogical Domain.

Table 30

Correlations of Instructional Pedagogical Subscales from Science Student Questionnaire Compared with Teacher Questionnaire Based on the Multilevel Confirmatory Factor Analysis Model

		Teacher Questionnaire (*N*= 90)		
	Scale	**IP2**	**IP7**	**IP10**
Student Questionnaire (*N=2023)*	Teacher Facilitation of Student Discussion (IP2)	.06		
	Teacher Facilitation of Student Interest (IP7)		-.15	
	Teacher Use of Differentiation (IP10)			.16

Instructional Student Engagement Component in Science

Instrument, Item Descriptives, and Reliability Assessment

Item means ranged from 2.04 (*SD* = 0.69) for 'during science time, I talk to my teacher about what we are learning' (students engage in discussion) to 2.73 (*SD*= 0.49) for 'during science time, I work hard to understand a lesson' (students engage in cognitively demanding work), with sample sizes for the items varying from 4,102 for students contribute to small group work to 4,404 for students engage in discussion. Again, missing data for subscale Students Contribute to Small Group Work (ISE1) was greater than the other subscales due to a screening question ('Do *you ever work with a partner or in groups during science time?*') students answered prior to answering the ISE1 items. Responses were not normally distributed, items showed a negative skew, with skewness ranging from -1.50 to -0.49 and kurtosis values ranging from -0.90 to 1.26 (Table 31). The student questionnaire items utilized a 3-point frequency scale: *Never or Hardly Ever*, *Sometimes*, and *A Lot*. See Table 32 for student responses to this scale by item.

Cronbach's alphas for the four scales described in Table 31, not taking into account the multilevel data structure were .50, .60, .63, and .55, respectively (Table 33). Given the multilevel nature of this data, these Cronbach's alphas represent a first look at the reliability of the data. Further below under the section entitled Multilevel Confirmatory Factor Analysis, the reliabilities are computed using the ICCs with the Spearman-Brown formula for the math sample of students nested within teachers.

Table 31

Item Descriptives for the Science Student Fidelity of Implementation Questionnaire Instructional Student Engagement Domain

Subscale Item	*N*	Number of Missing Cases	*M*	*SD*	Skewness	Kurtosis	ICC
Students Contribute to Small Group Work (ISE1)							
When we work in science groups, we work as a team. (1a)	4102	309	2.68	0.50	-1.15	0.17	.11
During science time, I learn from other students when working in groups. (1b)	4102	309	2.45	0.60	-0.57	-0.61	.08
When we do group work in science, I cooperate with other students. (1c)	4102	309	2.62	0.54	-1.03	0.01	.08
Students Engage in Discussion (ISE2)							
I talk to other students about our science work. (2a)	4404	7	2.20	0.64	-0.21	-0.68	.11
Students talk with each other about what we're learning during science time. (2b)	4404	7	2.25	0.69	-0.28	-0.60	.07
During science time, I talk to my teacher about what we are learning. (2c)	4404	7	2.04	0.53	-0.05	-0.80	.09
I am a good listener when my classmates are talking during science time. (2d)	4404	7	2.64	0.64	-1.12	-0.91	.03
Students Engage in Cognitively Demanding Work (ISE3)							
During science time, I explain how I get my answer. (3a)	4397	14	2.31	0.64	-0.39	-0.70	.06
When I come up with an answer in science class, I make sure that it makes sense. (3b)	4397	14	2.64	0.54	-1.12	0.23	.02
I explain why I agree or disagree with things my classmates say in science. (3c)	4397	14	2.32	0.64	-0.42	-0.71	.05
During science time, I work hard to understand a lesson. (3d)	4397	14	2.73	0.49	-1.50	1.23	.02

Table 31 (continued)

Subscale Item	*N*	Number of Missing Cases	*M*	*SD*	Skewness	Kurtosis	ICC
Students Take Risks							
When working on science problems, I am willing to try something new or different. (4a)	4382	29	2.53	0.58	-0.80	-0.36	.04
I say what I think in science even if it's different from other students. (4b)	4382	29	2.34	0.63	-0.41	-0.67	.02
During science time, I ask questions when I am confused. (4c)	4382	29	2.40	0.63	-0.56	-0.63	.04
I am not embarrassed to answer questions during science time. (4d)	4382	29	2.22	0.74	-0.38	-1.08	.04

Note. ICC = Intraclass correlation coefficient. ICC's are reported only for the sample of students who had a teacher ID (*N*= 2021). Response scale ranges from 1(*never or hardly ever*) to 3 (*a lot*).

For Students Engage in Cognitively Demanding Work (ISE3), there was not a significant difference in scores for students with TIDs (M=2.51, SD=0.39), and students without TIDs (M=2.49, SD=0.40; t[54395]=1.13, p=.26). The magnitude of the differences in the means was very small (eta squared = .000). For Students Take Risks (ISE4), there was not a significant difference in scores for students with TIDs (M=2.37, SD=0.42), and students without TIDs (M=2.38, SD=0.42; t[4380]=-1.08, p=.28). The magnitude of the differences in the means was very small (eta squared = .000).

Confirmatory Factor Analysis for the Science Instructional Student Engagement Model

CFA with corrected standard errors for nested data.

Before running the MCFA, I examined the factor structure using a single-level CFA with robust weighted least squares (WLS) approach (estimator = WLSMV in Mplus) and standard

errors adjusted to take into account cluster sampling (i.e., nested data) to examine the four-factor measurement model underlying the Instructional Student Engagement domain. The data were clustered by teacher ID.

Table 32

Student Responses for the Science Student Fidelity of Implementation Questionnaire Instructional Student Engagement Domain

Subscale **Item**		***Never or Hardly Ever (1)***	***Sometimes (2)***	**A lot (3)**
	N	**%**	**%**	**%**
Students Contribute to Small Group Work (ISE1)	4102			
When we work in science groups, we work as a team. (1a)		1.7	29.0	69.3
During science time, I learn from other students when working in groups. (1b)		5.3	44.4	50.3
When we do group work in science, I cooperate with other students. (1c)		3.0	32.4	64.6
Students Engage in Discussion (ISE2)	4404			
I talk to other students about our science work. (2a)		12.9	54.5	32.6
Students talk with each other about what we're learning during science time. (2b)		11.4	52.4	36.2
During science time, I talk to my teacher about what we are learning. (2c)		21.9	52.3	25.7
I am a good listener when my classmates are talking during science time. (2d)		2.7	30.4	66.9
Students Engage in Cognitively Demanding Work (ISE3)	4397			
During science time, I explain how I get my answer. (3a)		9.8	49.3	40.9
When I come up with an answer in science class, I make sure that it makes sense. (3b)		2.8	30.4	66.8
I explain why I agree or disagree with things my classmates say in science. (3c)		9.7	48.1	42.2
During science time, I work hard to understand a lesson. (3d)		1.9	23.5	74.6
Students Take Risks	4382			
When working on science problems, I am willing to try something new or different. (4a)		4.5	38.2	57.3
I say what I think in science even if it's different from other students. (4b)		8.3	49.0	42.7
During science time, I ask questions when I am confused. (4c)		7.9	44.5	47.6
I am not embarrassed to answer questions during science time. (4d)		18.3	40.9	40.8

Table 33

Internal Consistency of Instructional Student Engagement Subscales (Cronbach's α) for Science

Scale	# of Items	Cronbach's α	N	Item-to-Total Correlation Range
Students Contribute to Small Group Work (ISE1)	3	.50	4102	.30 to .34
Students Engage in Discussion (ISE2)	4	.60	4404	.20 to .49
Students Engage in Cognitively Demanding Work (ISE3)	4	.63	4397	.38 to.45
Students Take Risks (ISE4)	4	.55	4382	.27 to .39

The chi-square value for the single level, four factor CFA model, X^2 (84, N=2021) = 699.83, $p<.05$, indicated a statistically significant lack of fit. Alternative measures of fit, which are less sensitive to sample size, suggested the fit was not acceptable. The RMSEA (.07) was slightly greater than the .06 cut-off and the CFI of .89 was less than the .95 cutoff value for this index. A single level, four-factor CFA for students without TIDs was also run to examine if differences existed. The model fit indices for the Student CFA models with TIDs can be found in Table 13 and the model fit indices for the Student CFA models without TIDs can be found in Table 14. As can be seen in the tables the models fit pretty similarly for both students with TIDs and students without TIDs.

All factor pattern coefficients (loadings) were significantly different from zero (p< .05). The standardized loadings for the items within the ISE1 factor (students contribute to small group work) ranged from .52 to .65, from .51 to .68 for ISE2 (students engage in discussion), from .62 to .70 for ISE3 (students engage in cognitively demanding work) and from .35 to .65 for ISE4 (students take risks). The correlations between the factors were positive and significantly different from zero (p< .05) with ISE1 and ISE2, 1SE1 and 1SE3, and ISE1 and

ISE4 correlating at .76, .87, and .77, respectively, then ISE2 and ISE3, ISE2 and ISE4, ISE3 and ISE4 correlating at .79, .78, and .95, respectively.

An alternative one-factor model was also considered. This model did not fit as well as the four factor model based on the chi-square value, X^2 (90, N=2021) = 778.92, $p<.05$, and the other fit indices (RMSEA=.06 and CFI=.90) but the fit of the one-factor model was marginally acceptable. Standardized item loadings on the one-factor model ranged from .33 to .69.

Given that students were nested within teachers, thus violating the independence assumption, multilevel confirmatory factor analysis was used to further analyze the data for this study.

Multilevel Confirmatory Factor Analysis for the Science Instructional Pedagogical Student Model

Prior to conducting the MCFA, the variability between and within teachers on each item was examined by computing the intra-class correlations (ICCs) for each of the 15 items in the Instructional Student Engagement domain. Table 31 displays the ICCs for these 15 items. The ICCs for each of the observed items ranged from .02 (for item ISE4b within the ISE4 factor) to .11 (for item ISE2a within the ISE2 factor). These values indicated that there was sufficient between teacher variability to warrant multilevel analysis.

Initially a four-between group and four-within group factors model was run, like the model for ISE in mathematics, but the standardized solution showed ISE1 correlations greater than 1.0 with ISE2 and ISE3 (between level). So, as is shown in Figure 7, a multilevel model, in which the number of factors varied at each level (4 within factors and 1 between factor) was run. Results of this multilevel model with loadings freely estimated across levels indicated a reasonable fit of the model to the data. The RMSEA was .04 and the CFI was .91. The SRMR

fit indices at each level indicated that the fit of the level-1 (within) part of the model was better than at level-2 (SRMR within= .06 vs. SRMR between= .27; see Table 15 for measures of fit).

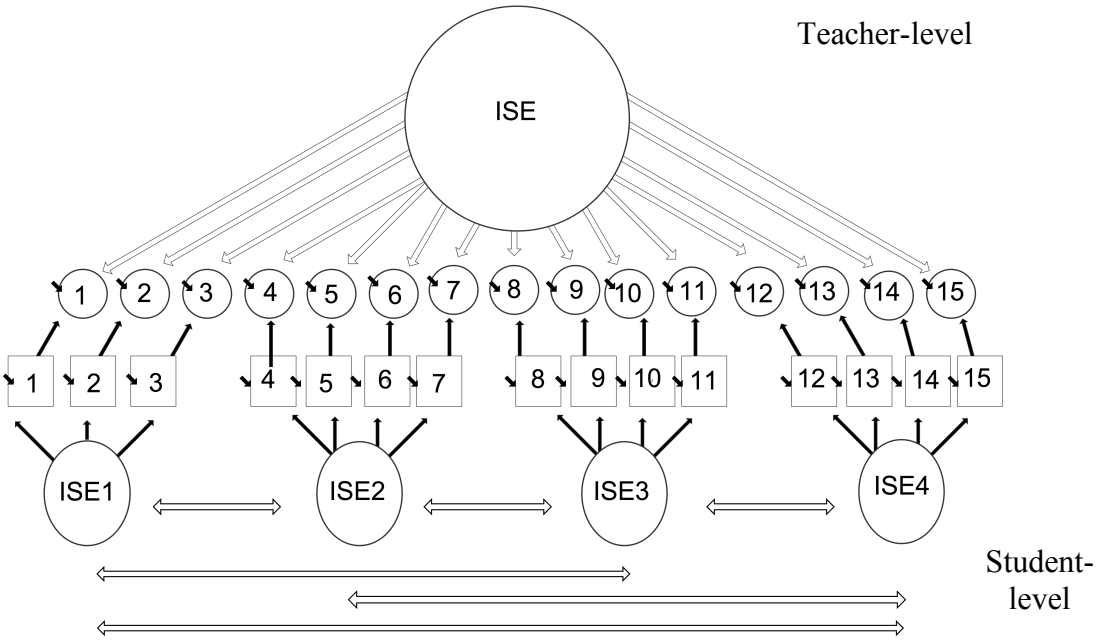


Figure 7. One-Between Group and Four-Within Group Factors for the Multilevel Confirmatory Factor Analysis for Instructional Student Engagement in Science

At level-1 (student) all factor pattern coefficients (loadings) were significantly different from zero (p<.05). At Level 2 (teacher) all factor pattern coefficients were also significantly different from zero (p<.05) except for three items: 1c (p=.30), 2d (p=.09), and 4d (p=.36). See Table 34 for the unstandardized factor loadings and residual variances.

Inter-factor correlations for Level 1were .68 (p<.05) between ISE1 and ISE2, .81 (p<.05) between ISE1 and ISE3, .74 (p<.05) between ISE1 and ISE4, .77 (p<.05) between ISE2 and ISE4 2, .75 (p<.05) between ISE2 and ISE3, and .94 (p<.05) between ISE3 and ISE4.

Table 34

Multilevel Confirmatory Factor Analysis: Unstandardized Factor Loadings and Residual Variances for the Four-Factor Model Underlying Student Ratings of Instructional Student Engagement

	Students with TID (*N*=2021)	Teachers (*N*= 90)	
Item on the Rubric	**Factor Loading**	**Factor Loading**	**Residual Variance**
Students Contribute to Small Group Work			
1a	1.00[a] (--)	1.00[a] (--)	0.74 (0.14)
1b	1.13 (0.10)	1.40 (0.56)	0.33 (0.19)
1c	1.58(0.16)	0.34 (0.33)	0.97 (0.06)
Student Engage in Discussion			
2a	1.00[a] (--)	2.26 (.80)	0.03 (0.17)
2b	0.84 (0.07)	1.32 (0.46)	0.37 (0.20)
2c	0.66 (0.0)	1.05 (0.50)	0.66 (0.14)
2d	0.80 (0.08)	0.37 (0.24)	0.89 (0.13)
Student Engage in Cognitively Demanding Work			
3a	1.00[a] (--)	1.33 (.51)	0.30 (0.14)
3b	1.04 (0.07)	0.57 (0.27)	0.64 (0.24)
3c	0.93 (0.06)	1.08 (0.44)	0.38 (0.16)
3d	1.10 (0.08)	0.51 (0.26)	0.74 (0.24)
Students Take Risks			
4a	1.00[a] (--)	0.65 (.30)	0.76 (0.15)
4b	0.91 (0.07)	0.72 (0.27)	0.28 (0.31)
4c	0.87 (0.06)	0.92 (0.43)	0.47 (0.18)
4d	0.47 (0.04)	0.20 (0.25)	0.97 (0.07)

Note. Numbers in parentheses represent the standard error.
[a]Factor loading fixed to 1.0

Multilevel ICCs and Reliability

It was not possible to calculate the multilevel reliability for this model, since the number of factors in the between and the number of factors in the within varied. In order to calculate the multilevel reliability I ran a one factor between and one factor within model. The fit of this model was not better than the one factor between and four factors within model [X^2 (180, *N*=2021) = 832.26, $p<.05$, RMSEA =.04, CFI =.88 , and SRMR= .07/.27]. Using this model

then it was possible to calculate the ICCs for the one latent variable and, subsequently, the reliability of the factor when aggregated at the teacher level. The ICC is the variation between teachers divided by the total variation. Total variation equals the combined within-and between-teacher variation. The ICC for ISE was .14. Using this ICCs with the Spearman-Brown formula, [*k*(ICC)/ [(k-1)(ICC) +1], where k is the average number of students nested within teachers , the estimated reliability for ISE in this study, with an average cluster size of 22 respondents (students) per teacher, was .78 (Table 37).

Confirmatory Factor Analysis for the Science Instructional Student Engagement Teacher Model

In this section, the model fit based on teachers' self-reported data (rather than students' reports nested within teachers) is presented. The chi-square value for the single-level, four-factor CFA model, X^2 (146, N=90) = 356.14, p< .05, indicated a statistically significant lack of fit. The alternative measures of fit, which are less sensitive to sample size, also suggested that the fit wasn't good. The RMSEA of .13 was much higher than Hu and Bentler's (1999) cutoff of .06, and the CFI of .93 was just slightly lower than the .95 cutoff values for this index.

All factor pattern coefficients (loadings) were significantly different from zero (p< .05). The standardized loadings for the items within the ISE1 factor (students contribute to small group work) ranged from .70 to .93, from .78 to .90 for ISE2 (students engage in discussion), from .50 to .88 for ISE3 (students engage in cognitively demanding work) and from .53 to .77 for ISE4 (students take risks). See Table 35 for the unstandardized factor loadings. The correlations between the factors were positive and significantly different from zero (p< .05) with ISE1 and ISE2, 1SE1 and 1SE3, and ISE1 and ISE4 correlating at .56, .55, and .33, respectively, then ISE2 and ISE3, ISE2 and ISE4, ISE3 and ISE4 correlating at .87, .71, and .80, respectively.

Table 35

Confirmatory Factor Analysis: Unstandardized Factor Loadings for the One-Factor Between and Four-Factor Within Model Underlying Teacher Ratings of Instructional Student Engagement

	All teachers with IDs (*N*=90)
Item on the Rubric	**Factor Loading**
Students Contribute to Small Group Work	
1a	1.00[a] (--)
1b	0.76 (0.09)
1c	0.82 (0.10)
Student Engage in Discussion	
2a	1.00[a] (--)
2b	1.12 (0.08)
2c	0.98 (0.07)
2d	1.10 (0.08)
Student Engage in Cognitively Demanding Work	
3a	1.00[a] (--)
3b	1.72 (0.22)
3c	1.58 (0.21)
3d	1.40 (0.20)
3e	1.70 (0.22)
3f	1.54 (0.22)
3g	1.81 (0.22)
3h	1.74 (0.23)
Students Take Risks	
4a	1.00[a] (--)
4b	1.45 (0.16)
4c	1.07 (0.22)
4d	1.31 (0.18)

Note. Numbers in parentheses represent the standard error.
[a]Factor loading fixed to 1.0

Convergent Validity

In order to examine convergent validity, meaning the correlation between student and teachers responses on the Instructional Student Engagement domain, the factor scores from the student perspective were correlated with the factor scores from the teacher perspective. The dataset consisted of 2,021 students (level-1) nested within 90 teachers of which all students had one teacher (level-2). Each of the 2,021 students provided data on instructional student

engagement from their perspective. These data constituted the lower-level (level-1) unit of analysis in this study. The second-level data included class instructional student engagement scores for each of the 90 teachers.

Table 36

Summary Table of Indicators of Internal Consistency for Mathematics

	Indicators of Internal Consistency		
Component Scale	**Cronbach's α**	**ICC(1)**	**ICC(2)**
Instructional Student Engagement			
Students Contribute to Small Group Work (ISE1)	.46	.08	.62
Students Engage in Discussion (ISE2)	.55	.37	.92
Students Engage in Cognitively Demanding Work (ISE3)	.57	.16	.79
Students Take Risks (ISE4)	.48	.10	.69
Instructional Pedagogy			
Teacher Facilitation of Student Discussion (IP2)	.62	.06	.56
Teacher Facilitation of Student Interest (IP7)	.56	.07	.60
Teacher Use of Differentiation (IP10)	.65	.38	.92

Note. ICC = Intraclass correlation coefficient. ICC is the reliability of individual level score as representation of group. ICC(2) is the reliability of group mean score to distinguish among groups. ICCs and ICC(2)s are reported only for the sample of students who had a teacher ID.

Preliminary analyses were conducted using the observed variables in SPSS. The student data were aggregated to create a teacher mean, as were the teacher data (although given that teacher responses for each student in a class were the same, the mean was the same as the teachers reported response). The correlations based on the observed variables between teacher and students on the Instructional Student Engagement components of Students Contribute to Small Group Work, Students Engage in Discussion, Students Engage in Cognitively Demanding Work, and Students Take Risks were -.05, .04, .06, and .05, respectively. Following that, the

data were examined in M*plus* by estimating the correlation of the latent variables, taking into account the two-level framework. Again as was mentioned prior in the section on multilevel reliability, a multilevel correlation between the teacher and student scores for the Instructional Student Engagement component could not be calculated for a model with varying factor structures across levels, so the one factor between, one factor within model was used to calculate the multilevel convergent validity. The correlation between teacher and student scores was .08.

Table 37

Summary Table of Indicators of Internal Consistency for Science

	Indicators of Internal Consistency		
Component Scale	**Cronbach's α**	**ICC (1)**	**ICC(2)**
Instructional Student Engagement			
Students Contribute to Small Group Work (ISE1)	.50	-.13	
Students Engage in Discussion (ISE2)	.60	.13	
Students Engage in Cognitively Demanding Work (ISE3)	.63	.11	
Students Take Risks (ISE4)	.35	.14	
Instructional Pedagogy			
Teacher Facilitation of Student Discussion (IP2)	.68	.02	.31
Teacher Facilitation of Student Interest (IP7)	.62	.21	.85
Teacher Use of Differentiation (IP10)	.62	.16	.81

Note. ICC = Intraclass correlation coefficient. ICC is the reliability of individual level score as representation of group. ICC(2) is the reliability of group mean score to distinguish among groups. ICCs and ICC(2)s are reported only for the sample of students who had a teacher ID. Multilevel reliability was calculated with a one-factor between one-factor within model. The correlation between teacher and student scores was .78.

CHAPTER 5
DISCUSSION AND CONCLUSION

The purpose of this study was to evaluate the reliability and validity of the student Instructional Pedagogical and Instructional Student Engagement scores for use in assessing teachers' fidelity of implementation. This chapter presents a summary of the study, discussion of the results, limitations, implications for the field, and recommendations for future directions.

Summary of the Study

Students cannot benefit from what they do not experience so assessing whether and how an intervention is delivered is important. There are multiple reasons why an intervention may not be delivered in its entirety or as it was designed. For example, it would be impossible to determine if an intervention designed to improve student outcomes in math failed because it was ill conceived and based on a faulty model, or if it failed because the theory was sound but the intervention was implemented poorly. In this era of educational accountability and limited dollars to go around, understanding how an intervention is delivered in the classroom is key to understanding why a program succeeds or fails. In order to assess how and whether a program has been implemented as intended an assessment of fidelity is needed. As noted in earlier chapters, the consequences of not assessing fidelity extend beyond methodological issues to substantive issues related to student performance when students do not 'experience' an intervention due to issues in intervention delivery and engagement.

Measuring fidelity is challenging for many reasons. Although five components that comprise fidelity (adherence, exposure, quality of delivery, participant responsiveness, and program differentiation) have been identified in the literature (Dane & Schneider, 1998; Dusenbury et al., 2003; Durlak & DuPre, 2008) definitional inconsistency and varying conceptual interpretations undermine what constitutes the core components of fidelity. This in turn fosters inconsistent application of methods to measure the construct (Gearing et al., 2011). Adherence and exposure are frequently the most assessed dimensions, perhaps in part due to ease of translation as they can be determined more objectively (e.g., intervention completion, determining if components of an intervention were delivered). In contrast, quality of delivery and participant responsiveness are less frequently assessed, given their process orientation and focus on assessing the interactions between the deliverer of services and the consumer. Even the methods and sources for collecting information on fidelity are challenging. Relying on the deliverer to accurately report activity (or lack thereof) may limit actual or perceived validity, through a social desirability bias, especially if staff suspect that the ratings may be a reflection of their performance. There is a significant potential for positivity bias among teachers (Lillehoj et al., 2004), which may be related to concerns that fidelity data might be used to evaluate performance (Donaldson & Grant-Vallone, 2002). Observation is thought to be more objective, valid and reliable than self-report (Rohrbach et al., 2007) but observation is costly and not always feasible as observers need to be identified and trained. Also, those conducting the observations may also pose validity issues as they are not blind to the program they are rating or why they are doing the rating. This holds true with the use of consumers as a fidelity data source, since some information may not be attainable from anywhere else besides directly from the consumer (Baldwin, 2000). For example, when examining the process or interactional piece

of fidelity as represented by participant responsiveness and engagement, consumers are likely to be the best source. Assessing participant responsiveness from the perspective of the participant may provide a more feasible, more objective, and less biased method of assessing fidelity when studying participant responsiveness, compared to observation and teacher self-report. When compared to other dimensions of fidelity, fewer studies have assessed participant responsiveness, especially outside the confines of a research study. Given its limited use as a measure of fidelity, the need to attend to procedural fidelity, and the potential benefits (greater objectivity and feasibility), there is an emerging interest in assessing participant responsiveness from the consumer's perspective as a way of complementing the multiple sources and methods that can serve to increase reliability and validity in fidelity ratings (Emshoff et al., 1987; Ruiz-Primo, 2005; Summerfelt, 2003; Vartuli & Rohs, 2009; Zvoch, Letourneau, & Parker, 2007).

As with all measures, an evaluation of the psychometric quality of participant responsiveness and engagement measures is also needed. Although researchers have identified critical steps in the development of fidelity measures (Bond et al., 2000; Century, Rudnick, & Freeman, 2010; Mowbray et al., 2003; O'Donnell, 2008), methods for validating fidelity measures are still unclear and there are fewer studies in the literature on fidelity focused on validation. According to Mowbray et al. (2003), five different approaches have been used to assess the psychometric quality of fidelity measures. The first approach has focused on reliability in terms of assessing consistency in respondents' perceptions (i.e., inter-rater reliability) and internal consistency of responses to multi-item scales as measured by Cronbach's alpha. The second approach, which focuses more on validity, has involved examining the internal structure of the data using exploratory and confirmatory factor analysis (Henggeler et al., 2002), or cluster analysis (Mills & Ragan, 2000). The third approach is the method of known

groups where one examines differences in fidelity scores across programs that are expected to be different (Bond et al., 2001; Hernandez et al., 2001; Lucca, 2000; Teague et al., 1995). Convergent validity is the fourth approach to validation. In convergent validity the focus is on examining the strength of the relation between two different sources of information about the program and its operations. The fifth approach is to examine the relationship between fidelity measures and expected outcomes for participants (e.g. Becker, Smith, Tanzman, Drake, & Tremblay, 2001).

This study was conducted because of the need to better understand the psychometric properties of fidelity measures used to assess fidelity of interventions designed to enhance student outcomes. The goal of this study was to move the field towards fidelity measures that examine the procedural aspects of fidelity (interactions), that use multiple methods and sources to assess fidelity, and that use appropriate methods to evaluate the psychometric quality of fidelity instruments.

This is a secondary data analysis study. The data for this study consisted of responses from students and teachers in mathematics and science across three school districts and 41 schools to an online fidelity of implementation questionnaire focused on assessing student engagement. The data for this study were hierarchically structured with students‘ responses nested within teachers. Single level and multilevel confirmatory factor analyses were used to evaluate the measurement models underlying the Instructional Pedagogical and Instructional Student Engagement fidelity measures. These models were examined separately for the science and mathematics instructional interventions. Reliabilities were determined for the scales underlying the Instructional Pedagogical and Instructional Student Engagement fidelity measures taking into account the multilevel data structure, as well as by ignoring this structure. Finally,

the relationships between students' and teachers' responses to the Instructional Pedagogical and Instructional Student Engagement domains, as a measure of convergent validity, were evaluated.

Discussion of the Results

Research Question 1

The first research question was addressed in two parts. The first part examined the internal consistency reliability of the scores for the IP and ISE components (for both Mathematics and Science) on the student instrument using the students as the unit of analysis and ignoring the multilevel structure of the data (i.e., students nested within teachers). The single level approach was used because of its frequent use in the field; it is only recently with the introduction of multilevel modeling techniques that reliability has been calculated using a multilevel framework. The single level reliability indicators were calculated using the entire sample of students in each content area (Mathematics and Science). The Cronbach's alphas for the scales in the Instructional Pedagogical Domain ranged from .55 to .62 for Mathematics and .62 to .68 for Science. Cronbach's alphas for the scales in the Instructional Student Engagement Domain in Mathematics, ranged from .46 to .57, and .55 to .63 for Science. For both domains (IP and ISE) and both content areas (Mathematics and Science) the alphas were lower than .70, which is considered acceptable reliability in social science research (Nunnally, 1978). This may in part be due to the fact that both IP and ISE contained only 3-4 items. In IP there were three factors, two of which only had three items; for ISE, there were four factors with three factors having four items each and one factor having three items. The Cronbach's alphas increased and were acceptable when ISE was treated as one factor with 15 items (.76 in Mathematics, .80 in Science). However, when IP was treated as one factor with 10 items (.62 in Mathematics, .63 in

Science) the single level reliability was not much better. The single level reliability for the ISE domain may also be higher because there are more items than in the IP domain. In addition to the Cronbach's alpha, I looked at the item-to-total correlations. Item-to-total correlations, which are the correlations between individual items and the total score, were also low, implying poor internal consistency.

Lower reliability is consistent with the poorer model fit as was determined using confirmatory factor analysis. It should be noted that the reverse can also be true, reliability can appear to be good, but model fit may not be acceptable. These analyses underscore the need to examine the psychometric quality of the measures from multiple perspectives.

Given that this instrument is still in its first generation of development, there may still be issues with items and wording that need to be resolved. These issues are particularly salient when considering the age of the students (i.e., 7-12 years) taking the measure and their interpretation of the items and the scale. Also another issue to be considered is the method by which data were collected from these 3rd to 5th grade students. The use of computers to administer the online survey was not tested (pilot test was paper and pencil) and may have impacted reliability.

The second part of the question was examined using multilevel reliability given that students were nested within teachers/classes. Estimating standard reliability estimates (e.g., Cronbach's alpha) from data collected at multiple levels (e.g., students nested within teachers) can confound the within-group variance and between-group variance and lead to biased reliability estimates as the assumption of independence is violated. For example, we may compute reliability on student achievement scores, but when researchers aggregate those achievement scores at the school level and talk about school achievement, an incorrect

assumption is made that the reliability of student achievement scores at the student-level is equal to the student achievement scores at school-level. As a consequence, single level reliability estimates may not reflect the true scale reliability at any single level of the analysis as it assumes a single level factor structure (Geldhof et al., 2013).

Multilevel analyses helps to avoid the forced choice of unit of analysis when the data to be analyzed are hierarchical. In many studies, the scale scores for individuals are used, ignoring the clustering or nesting of the data within a group. Other studies have averaged the individual responses to come up with a group mean, thereby ignoring the variability in individual responses. Further, regardless of the choice of unit of analysis for the study phase, many researchers have used the individual as the unit of analysis for the psychometric phase. This is problematic in that student perceptions of teacher facilitation of student discussion (IP2 factor in the Instructional Pedagogy domain) may reflect differences among teachers/classes in their organizational properties and contexts, but may also reflect differences among students who share membership in the same class. A multilevel analysis enables one to adjust for the effects of variables measured at the individual level when estimating effects of variables measured at the teacher/class level (Raudenbush et al., 1991).

In the computational analyses of single level reliabilities clustering by teacher/class is not taken into account. When students are clustered by teacher, reliabilities using a multilevel framework are based upon the average number of students in a class. In this study, the average number of students per class ranged from 20 (for Math) to 22 (for Science). Multilevel reliability then varies depending upon the number of student informants; fewer informants in a class would decrease the multilevel reliability estimate. For teacher facilitation of student interest (IP2), teacher facilitation of student discussion (IP7), and teacher use of differentiation

(IP10) in Mathematics the multilevel factor reliabilities were .56, .60, and .92, respectively, and for IP in Science the multilevel factor reliabilities were .31, .85, and .81, respectively. For students contribute to small group work (ISE1), students engage in discussion (ISE2), students engage in cognitively demanding work (ISE3), and students take risks (ISE4) in Mathematics the multilevel factor reliabilities were .62, .92, .79, and .69, respectively, and for ISE in Science .78 (in order to calculate the multilevel reliability for ISE a one-factor between and one-factor within model was run). Depending on the measures used in the multilevel analysis, values between .70 and .85 are usually taken to indicate acceptable levels of reliability (LeBreton & Senter, 2008; Lüdtke, Trautwein, Kunter, & Baumert, 2006). The greater than acceptable reliability for ISE Science was probably related to it being a one-factor mode with 15 items, as opposed to ISE in Mathematics, which had four factors with 3-4 items in each factor. The multilevel reliabilities for Mathematics ISE2 (students engage in discussion) and ISE3 (students engage in cognitively demanding work) were acceptable, however. The lowest reliability was for the IP2 scores, teacher facilitation of student discussion, which had limited between teacher variance and high within group variance (error) for this construct. Multilevel reliabilities for IP2, IP7, ISE1 and ISE4 in Mathematics and IP2 in Science fell below the .70 minimum criteria. The ICC values for these factors are the lowest of the ICC values and range from .02 to .10 (see Tables 36 and 37). When these ICC values are taken into consideration, and using the Spearman-Brown Prophecy Formula, the number of informants needed per teacher/class to obtain a .70 reliability in mathematics for IP2 was 37, for IP7 was 31, ISE1 was 27, and ISE4 was 21, and 114.4 for IP2 in Science. The larger number of informants, particularly for Science, needed is because students within the same teacher differed in their perceptions and also because there was limited true score variability between teachers/classes on these factors. One implication of this finding

of large within teacher/class variability is that researchers studying teachers/classes who use few informants within a class will produce scores with low reliabilities at the teacher/class level, resulting in attenuated relationships with other variables.

In general, the multilevel reliabilities for both IP and ISE in both content areas were higher than the single-level reliabilities (with the exception of IP2-teacher facilitation of student discussion). In Mathematics for example, the reliability of the factor teacher use of differentiation (IP10) at the teacher level was .92, whereas the reliability of this factor at the student level was .65. The reliability of the factor students take risks (ISE4) at the teacher level was .69, whereas the reliability of this factor at the student level was .48. This higher reliability is due to the fact that the reliability of an aggregated score (i.e., mean of students' scores for a teacher) is due in part to the number of students in a class who provide ratings; with more student ratings within a class the more accurately the class-mean rating will reflect the true value of the construct being measured. Summary tables by content area of these indicators of internal consistency can be found in Chapter 4 (Tables 36 and 37). When comparing the single to multilevel reliabilities, it was surprising that for the teacher facilitation of student discussion factor (IP2) that the single level reliability was stronger than the multilevel reliability for both mathematics and science.

It was expected that for the Instructional Student Engagement component the student level reliabilities would be stronger than the teacher level reliabilities, given that students were reporting on their own participation and engagement behaviors. But the teacher student relationship is interactional and student participation and engagement are related to the teaching practices that teachers deliver to their students in the classroom.

Internal consistencies at both the teacher/class and student levels have several important properties that have to be taken into consideration when interpreting results, as well as when developing and testing instruments. Given that many studies do not compute multilevel reliability, researchers are unaware of the true score variation of their measures. Having this awareness is especially important when constructing new scales so that researchers can demonstrate that the new scale reliably captures true score variation at each possible level of analysis (Geldhof et al., 2013). According to Raudenbush, Rowan and Kang, (1991), the teacher/class level internal consistency depends upon four quantities: the number of items in a scale, the level of inter-correlation among the items at the student level within the scale, the level of inter-subjective agreement between students within the teacher/class, and the number of students sampled within that teacher/class. Given the relationship between ICC(1), group size, and group mean reliability when considering refining an instrument for increased reliability at the teacher/class level, researchers may need to increase the number of students who provide information (Bliese, 1998) about treatment fidelity (even when inter-subjective agreement is low), as opposed to increasing the number of items in an instrument, which will have limited benefit. In contrast, when researchers are creating measures to examine individual level differences researchers need to focus on the degree of inter-correlation among items and the number of items in a scale.

Since reliability is a pre-requisite to validity, appropriate reliability analyses of multilevel data are important in understanding the psychometric quality of a measure. In moving beyond the psychometric phase of a study to a phase where hypotheses are tested, researchers need to know the reliability of the scores in order to understand the extent to which the strength of the relationship between variables may be attenuated.

Research Question 2

The second research question was focused on examining the factorial validity of the three-factor Instructional Pedagogical (IP) model and the four-factor Instructional Student Engagement (ISE) model in Mathematics and then the three-factor Instructional Pedagogical model and the four-factor Instructional Student Engagement model in Science. For each domain (IP and ISE), the factorial validity analyses began with an examination of the factor structure using single level Confirmatory Factor Analysis (CFA). Ignoring multilevel data structures when evaluating the factor structure of latent variables will likely result in models that exhibit more misfit (i.e., inflated chi-square test), hypothesis tests that are overly optimistic (i.e., deflated standard errors leading to increased Type I error), and inflation of the parameter estimates (e.g., factor loadings) when the ICCs are ≥ .10 (Myers, Feltz, Maier, Wolfe, & Reckase, 2006). Following confirmation that all the models were suitable for multilevel analyses (i.e., ICCs > .05), the a priori three factor IP model and four factor ISE models were tested for model fit using Multilevel Confirmatory Factor Analysis (MCFA). MCFA should be considered when individuals are meaningfully nested within groups and evaluation of the factor structure of a set of indicators is desired (Muthén, 1994). Given the ordinal nature of the data, the models were run using the Weighted Least Squares Means and Variance (WLSMV) adjusted estimation method. The Multilevel Confirmatory Factor Analyses for both IP and ISE scores were conducted using the sample of students who were assigned to a teacher (*N*=3,103 in mathematics, *N*=3,096 in science) and the sample of teachers who had a teacher ID (*N*=152 in mathematics, *N*=90 in science).

For Instructional Pedagogy, the models which were evaluated using multiple measures of fit, indicated that the three factor model fit the data more appropriately than other models, but the

results based on the fit indices for both Mathematics and Science were mixed, given that the chi-square fit was statistically significant and the CFIs were slightly lower than the .95 CFI cutoff used as a measure of good fit. Overall the models fit reasonably well (the results of the single level confirmatory factor analyses were similar) but fit was not excellent. With respect to variance, greater within group variance (than between group variance) was noted for both content areas for teacher facilitation of student interest (IP7).

Similarly, the Instructional Student Engagement models were evaluated using multiple measures of fit. For Mathematics, the four-factor model fit the data more appropriately than other models, but the results based on the fit indices for Mathematics were mixed, given that the chi-square fit was statistically significant and CFI values were slightly lower than .95 (again, the results of the single level confirmatory factor analyses were similar). For Science, the four factor model appeared to fit, but the correlations of the latent variables at level-2 (i.e., teacher) showed correlations greater than 1.0 for ISE1 with ISE2 and ISE3. The ISE Science model was rerun as a one-factor between, four-factor within model. The fit was similar to that of the four-factor between and four-factor within model. Like all the other models, the chi-square was statistically significant indicating a statistically significant lack of fit and the CFI values were slightly lower than .95. Similar to Instructional Pedagogy, there was greater within group variance than between group variance for the Instructional Student Engagement factors, particularly for students engage in discussion (ISE2) and students engage in cognitively demanding work (ISE3).

The Instructional Pedagogical Domain describes three dimensions of instructional pedagogy: teacher facilitation of student discussion, teacher facilitation of student interest, and teacher use of differentiation. A priori the decision was made to test the models based on a belief that these were three dimensions of instructional pedagogy. The same was true for the

Instructional Student Engagement domain, which describes four dimensions: students contribute to small group work, students engage in discussion, students engage in cognitively demanding work, and students take risks. Given the mixed fit results, the question of whether the dimensions were broken down into three and four factors as I had hypothesized a priori led me to re-run the models as one-factor between and one-factor within models for IP and ISE in Mathematics and Science. The decision to examine the models as one-factor models was based in part on prior single level exploratory and confirmatory factor analyses that were conducted during pilot testing of the measures (n = 252; the small sample size precluded running multilevel models) and the results of the current study. Despite the statistical lack of fit, it was decided by the CEMSE team that conceptually a one-factor model was more appropriate. The preliminary EFAs and CFAs in the pilot study had showed mixed results with minimal support for the one-, three- and four-factor models. In addition, in the pilot study items from each factor did not load together as they were conceived to load and there were many items with factor loadings lower than .30.

Each of the one-factor between and one-factor within models had poorer fit than the three- and four-factor solutions, and in the case of ISE for Science the fit for both the four-factor between/four-factor within and the one-factor between/four-factor within was still better than the one-factor between/one-factor within model.

Using .40 as a cutoff for meaningful factor loadings (Henson & Roberts, 2006), when the standardized factor loadings are examined for IP in Mathematics, there were two level-1 indicators, and one level-2 indicator with factor loadings less than .40. For IP in Science, there was one level-1 indicator, and one level-2 indicator with factor loadings less than .40. At level-1, item 2a and item 10a had factor loadings lower than .30 in Mathematics; Item 10a also had a

factor loading less than .40 for Science. For level-2 the factor loadings for the IP model for both content areas were reasonable aside from the lower factor loading for item IP2a in math, 'my teacher asks us questions during math/science time.' When the standardized factor loadings are examined for ISE in Mathematics and Science, there were quite a few items with loadings lower than .40; these were item 4d at level-1 and items 1c, 2d, 3d and 4d in both math and science at level-2.

It was hypothesized that the three factors that comprise Instructional Pedagogy would be correlated for Mathematics and Science at both levels. Though when the correlations between the IP factors for both mathematics and science, were reviewed the correlations were predominately weak and in the case of IP7 and IP10, at both levels, the factors were negatively correlated (and not statistically significant at level-1). Aside from the correlation between IP2 and IP7 at level-1 for both Mathematics and Science, all other factors were weak suggesting good discriminant validity and little shared variance. Although, low reliability in measures attenuates relationships, making it appear as if there is discriminant validity. Similarly, it was hypothesized that the four factors that comprise Instructional Student Engagement would also be correlated for Mathematics and Science at both levels. For ISE in Science, the strong correlations of .94 between ISE3 and ISE4 at level-1 and for ISE in mathematics, the strong correlations of .92 between ISE1 and ISE2 at level-2 suggests that these factors shared considerable variance and have limited discriminant validity. In general for ISE across both content areas, there were strong positive correlations between all four factors at both levels with .71 being the lowest correlation between factors; these large correlations indicate that there is limited discriminant validity between the ISE factors.

The standardized loadings for IP in mathematics and science suggest that the items with weak loadings, as well as factors that were negatively related to one another may have contributed to the mixed results related to model fit. The small standardized loadings for ISE in mathematics and science also indicated that there were weak items. The weak loadings contributed to the mixed results in model fit. These findings are not surprising given the single level and multilevel reliability results presented earlier. It is important to note here that issues with the factor structure also have implications for reliability estimates. Single-level reliability was examined prior to determining the factor structure in this study, based on the a priori model that was tested in this study. Researchers may inappropriately use Cronbach's alpha as a measure of unidimensionalty. For more accurate results when examining a measure, the factor structure should be explored first (before looking at reliability). This is especially the case when analyzing multilevel data as the factor structure may vary at different levels. For example, if the hypothesis is that a model has three factors at the between level and three factors at the within level, but the model does not fit, then looking at the reliability does not make much sense.

Additional psychometric analyses could have been done, such as model modification procedures to improve the fit of these models, but the purpose of this study was to test the a priori models. Future research will need to determine if the present results will be replicated with new samples of students and teachers.

Research Question 3

The third research question was focused on convergent validity and was evaluated using the correlation between students' and teachers' responses for the IP and for ISE in both mathematics and science. Convergent validity has been described by Mowbray et al. (2003) as a

promising method for validating fidelity measures. Convergent validity occurs when the fidelity measure being studied correlates highly with other fidelity measures of the same construct (Calsyn, 2000). Convergent validity is a validation method that may limit the bias and confounding that can come with the use of one source of information for validation and assessment. Convergent validity involves examining the relation between two different sources of information about the program and its operations (e.g., compare records and documents with on-site observations) and/or comparing the same measures of fidelity across diverse information sources (teachers, students, observers). The use of multiple sources and methods serves to increase confidence in fidelity ratings. Multiple studies within the field of mental health have used convergent validity methods (Blakely et al., 1987; Lucca, 2000; Macias et al., 2001; McGrew, Pescosolido, & Wright, 2003). For example, in a study by Lucca (2000), on a vocational program for adults with psychiatric disabilities, the fidelity measure used was a 15-item checklist derived from the literature of program components essential to the clubhouse model. Convergent validity was established by examining the relationship of the fidelity score assigned to each program by non-staff evaluators with staff members' responses on a scale measuring how consistently their programs followed psychosocial rehabilitation principles. There was a statistically significant correlation (.59) between the number of model components a clubhouse had in place and the programs adherence to rehabilitation principles, as reported by staff members. In a recent study by Snyder, Hemmeter, Fox, Bishop, and Miller (2013), convergent validity was assessed by comparing scores in 50 teachers/classrooms on the TPOT-P (a fidelity instrument to assess practitioners' fidelity of implementation of professional development practices associated with the Pyramid model and scores on the CLASS (Classroom Assessment Scoring System). Both instruments were observation based. Correlations between

the TPOT-P scores and CLASS domain scores ranged from .64 to .74, for key practice subcomponents scores ranged from .70 to .76, and for red flags and domain scores, correlations ranged from -.70 to -.55, and correlations between environmental arrangement scores ranged from .08 to .13. The negative correlations between the measures for red flags was explained by associations between lower red flags and higher instructional and interactional quality as measured by CLASS, and the low correlations between TPOT-P scores and CLASS scores on environmental arrangements was not unexpected given CLASS does not measure environmental features. Although some fidelity studies have used convergent validity methods, the existing research literature on fidelity lacks studies of convergent validity using consumers as an information source (Mook, 2010). An exception is a study by Mowbray et al. (2006), which attempted to use convergent validity to validate a fidelity rating instrument for consumer-operated drop in services. Fidelity ratings by trained observers were validated in relation to reports from interviews about similar concepts from users of the center. Out of the 31 relationships between program level fidelity ratings and consumer reported program attributes, the reliabilities of most consumer reported variables were above .50, and half were above .60, with two single item measures near .20 and one measure with a reliability of .01.

For this study, the factor scores from the student perspective (aggregated to the teacher level) were correlated with the factor scores from the teacher perspective. Students are informants relaying information about the teacher on the two student engagement constructs (instructional pedagogy and instructional student engagement), but students also have their own factor model, as do teachers. For example, data regarding the instructional pedagogy domain was gathered from two sources: from the teachers (self-ascribed instructional pedagogy) and their students (perceived instructional pedagogy). Preliminary analyses of convergent validity

for the observed variables were conducted using SPSS. Students' responses for a teacher were aggregated to create a teacher mean. As mentioned previously, aggregating student responses to create a mean score to correlate with teacher reports is a common approach researchers use to create scores with observed variables.

When looking at IP for mathematics, the correlations based on the observed variables between teacher and students on the Instructional Pedagogical components of Teacher Facilitation of Student Discussion, Teacher Facilitation of Student Interest and Teacher Use of Differentiation were .25, .15, and .42, respectively). For ISE in mathematics, the correlations based on the observed variables between teacher and students on the Instructional Student Engagement components of Students Contribute to Small Group Work, Students Engage in Discussion, Students Engage in Cognitively Demanding Work, and Students Take Risks were .03, .23*, .07, and .18*, respectively (* correlation is significant at the .05 level). When we turn to science, the correlations based on the observed variables for IP between teacher and students were .02, -.10, and .15, respectively. For ISE in science, the correlations based on the observed variables between teacher and students were -.05, .04, .06, and .05, respectively. None of the observed correlations for science were statistically significant.

Although, the correlations for the observed variables were not expected to be very strong, it was surprising that the correlations were so weak. Additionally, the correlations between the Instructional Student Engagement component were much smaller than the Instructional Pedagogical component, which was also surprising given the ISE component had a greater number of items in both the student and teacher measures. The convergent validity coefficients of the observed variables for both student engagement components (instructional pedagogy and instructional student engagement) in both content areas were predominantly weak, aside from

IP10 (teacher use of differentiation). Teacher and student reports for Students Contribute to Small Group Work (ISE1 in the Instructional Student Engagement component) were negatively correlated; there is not enough evidence to determine whether teachers' self-ascribed instructional student engagement and their students' perceived instructional student engagement are related (to reject the null hypothesis that the correlation is zero).

As was noted earlier in the discussion, the aggregation of individual level data is not appropriate when data are nested, such as students nested within teachers/classes as in this study. So, following the preliminary convergent analyses of observed variables, multilevel confirmatory factor analysis was used to estimate the correlation of the latent variables, taking into account the two-level framework. The dataset consisted of students (level-1) nested within teachers of which all students had one teacher (level-2). Each of the students provided data on the student engagement constructs (instructional pedagogy and instructional student engagement), from their perspective. These data constituted the lower-level (level-1) unit of analysis in this study. The second-level data included class student engagement (instructional pedagogy and instructional student engagement) scores for each of the teachers.

In Mathematics, the correlations based on the latent variables between teachers' and students' scores on the instructional pedagogical components of Teacher Facilitation of Student Discussion, Teacher Facilitation of Student Interest, and Teacher Use of Differentiation were .38, .26, and .72, respectively. The correlations between the teachers' and students' scores on the Instructional Student Engagement components of Students Engage in Discussion, Students Engage in Cognitively Demanding Work, and Students Take Risks were -.07, .28, .20, and .41, respectively. In Science, the correlations based on the latent variables between teachers and students' scores on the Instructional Pedagogical components of Teacher Facilitation of Student

Discussion, Teacher Facilitation of Student Interest and Teacher Use of Differentiation were .06, -.15, and .16, respectively. In Science, the correlation between teachers' and students' scores from the one-factor between, one-factor within model was .08.

The convergent validity coefficients of the latent variables for both student engagement components (instructional pedagogy and instructional student engagement) in both content areas were as expected stronger than the convergent validity coefficients of the observed variables, with the exception of IP7 (teacher facilitation of student interest), which was weaker and negative. For the latent variables, the majority of the correlations were weak with the exception of IP10 again (teacher use of differentiation) which was strongly correlated. The hypothesis going into this study was that the convergent validity coefficients of the latent variables for both student engagement components (instructional pedagogy and instructional student engagement) would be more strongly correlated. The convergent validity coefficients of the IP and ISE latent variables for mathematics were larger than IP and ISE for Science, which may be because there were more teachers at level-2 in the Math sample.

The negative correlations and weak correlations for both observed and latent IP and ISE variables in both content areas may be related to issues mentioned previously: number of items per factor, poor item consistency, weak item loadings on factors. Another related issue is the differences in response scales for the teacher questionnaire and student questionnaire. Differences in response scales can sometimes introduce differences and lower correlations. The student scale is a 3-point scale, whereas the teacher scale is a 5-point scale. Additionally, there is some variation in the number of items by factor, specifically for the students engage in cognitively demanding work (ISE3) factor in the teacher questionnaire, which has 8 items when compared to the 4 items on the student instrument. There is variation in the means across

measures by factor. For example, for IP2 in mathematics, the mean for student items was 2.29, whereas the mean for the teacher items was 4.33 and for ISE3 in mathematics, the student mean was 2.53 and the teacher mean was 3.34.

Also, given that in the data set 50 or more students could have been associated with a teacher ID, it is assumed that teachers taught more than one class, but teachers only completed the teacher questionnaire once for all the classes they taught. For factors like ISE1 in which the items specifically ask what proportion of 'students contribute to group work, manage time efficiently, and work collaboratively with their peers' students likely answered the items with one teacher/class in mind, whereas teachers answered their items with all students, in all their classes in mind. This is problematic for assessing fidelity, given teachers' responses are not specific to a class, and particularly more problematic when we are assessing student engagement since lower scores cannot be acted upon for improvement when the target of the teacher's responses is unknown. When the teacher and student items for IP10 (teacher use of differentiation) are examined, the stronger correlation for this factor when compared to all the other factor correlations may also be related to the items used to assess teacher use of differentiation. The teacher items for IP10, teacher use of differentiation asks teachers to rate themselves on how often they 'scaffold ideas and activities for individual students, give students different activities based on ability and learning modality, and group students based on their ability and learning modality.' The student items ask students about the extent to which they 'do the same work at the same time, some students do different work than others, and do work that is different from what other students are doing'. The stronger correlation may be because both the teacher and student items ask for ratings from their own experience, as opposed to ISE3, students engage in cognitively demanding work where students are asked about their own experience but

teachers are asked to report on the proportion of students. Although the correlation is not very strong, it is promising that teachers' and students' scores on this factor converge, especially given that teachers are self-reporting on their teaching practices on differentiation and students are rating 'receipt' of those teaching practices. Also, it is important to note here that the low reliabilities, and issues in the specification of the factor structure, result in attenuation, which has implications for convergent validity. Attenuation weakens the relationship between variables, in that there is a large amount of randomness in the data that will not correlate. It is possible that the correlations between teachers and students might have been stronger had there not been attenuation (see the correlations for the latent variables which do not include random error for possible approximations of the true correlation between student and teacher report).

Conclusion

In conclusion, the psychometric results of the student fidelity of implementation questionnaire assessing the student engagement components of fidelity (i.e., instructional pedagogy and instructional student engagement) were mixed in this study. The single and multilevel internal consistency reliabilities of the scores from the Instructional Pedagogy component and Instructional Student Engagement component in Science and Math were not acceptable with a few exceptions. Support for the factorial validity of the multilevel student models (IP for Mathematics, ISE for Mathematics, IP for Science, ISE for Science) was less than acceptable, with model fit indicating that some of the measured variables did not load strongly on their respective factors and some of the factors lacked discriminant validity. Lastly, the correlations between students' and teachers' scores for both the observed and latent IP and ISE variables displayed limited convergent validity.

Implications of the Study

The results of this validity study indicate that caution should be taken in the use of this student questionnaire, especially when assessing fidelity of implementation (i.e., instructional pedagogy and instructional student engagement). This caution is based on several limitations related to instrumentation and design that had implications for validity and reliability.

This study also demonstrates the importance of attending to multilevel analyses in the psychometric phase that has implications for the appropriate reporting and interpretation of reliability and validity findings. The use of multilevel psychometric analyses in validating fidelity measures is key when using consumers as a source of fidelity data, given that the individuals who receive the service delivery will be nested within the person delivering that service. However, even when we consider the common practice of assessing fidelity by self-report from the individual delivering the services, the individual is typically still nested within an organization. The organization in which they are nested may provide varying levels of support, which may impact fidelity between others delivering services within the organization or between organizations delivering the same intervention. For example, if a program developer or funder is interested in understanding how well services for a particular intervention are delivered across agencies in a specific region, looking only at individual scores or scores aggregated to the agency level will not be as meaningful as looking at scores that factor in the agency (and for a three level model-region), especially when a funder or program developer wants to understand poor program outcomes. For this study, although students were nested within teachers/classes who were nested in schools, it was not possible to include school in the model because there were a limited number of schools and teachers came from across all 41 schools.

Instrumentation decisions have important implications for the reported reliability and validity results in this study. The psychometric analyses for this study were limited in this study by the few items (3-4) used to measure each of the factors. When compared to other engagement instruments (e.g. TROFLEI – 8 items, MCI Short- 5 items) the number of items per scale is lower. In an effort to limit the number of items or questions asked of students, the reliability at the individual level suffers. This highlights the need for researchers to think about the context (nesting) in which intended respondents exist when developing an instrument. Prior to this study the teacher questionnaire had been previously developed and validated with a 5-point scale. The purpose of the original CEMSE study was to validate the student measure, not to examine convergent validity. So for the purpose of their study, CEMSE appropriately selected a 3-point scale for the student measure (when compared to other measures of children of the same age and grade, e.g., Child Behavior Checklist, MCI-Short Form). The variation in response scales for the student questionnaire and teacher questionnaire may have also created differences and reduced correlations.

The psychometric analyses as well as the possible interpretations of study results were limited in this study by the larger study's design decisions (from which these data came). Students who participated in the study were in grades 3-5. To my knowledge students this young have not been used in other studies to assess fidelity, but have been used in other survey based studies. Also students took an online survey. Previous administrations of the survey were paper and pencil, so the method used to collect the student data was not tested and may have contributed to the less than acceptable results.

As was explained previously, a large number of students completed the online student questionnaire, but not all of their respective teachers completed the teacher questionnaire. So

only students with a related teacher ID could be used in the multilevel analyses. This severely limited the sample of teachers for the multilevel psychometric analyses (approximately half of the entire teacher sample was included), as well as limited the number of students. Students identified their teachers by name in the questionnaire, so even having the student data grouped for a teacher without the teacher data could have been beneficial for establishing factorial validity. There are greater implications though for how teacher data were collected than just numbers; teachers were permitted to self-report on their teaching practices for all their classes rather than complete one teacher questionnaire for each class. So all students connected to that teacher than formed a 'class' that may not have represented the true membership of a class. So the extent to which teacher and student groupings reflected actual classrooms is unknown. This is problematic when we think about assessing fidelity of implementation. Teachers are not responding to the questionnaire with one class in mind, but rather all the classes in which they taught the mathematics or science module. In contrast, students responded to the items in the questionnaire with their particular teacher in mind for the Instructional Pedagogy items, and with themselves in mind for the Instructional Student Engagement items. Student responses were then more specific than those of teachers. Had teachers responded to the teacher questionnaire Instructional Student Engagement items with a particular class in mind or completed a questionnaire for each class they taught there might have been greater convergent validity between the student and teacher scores.

Classroom context plays a key role in student engagement and perceptions of teacher instructional practices. Variation can exist when delivering specific teaching practices from class to class at a minimum based on the types of students in a particular class. This variation goes undetected when a teacher responds to the questionnaire with all classes in mind. Even if

one could successfully argue that teachers' teaching practices do not differ across their classes, student engagement as assessed by the Instructional Student Engagement component will differ between classes. For example, teachers are asked in the Instructional Student Engagement component of the teacher questionnaire to report the proportion of students who 'regularly worked collaboratively with their peers,' 'responded to questions in a group setting,' and 'conversed with the teacher about a topic.' When teachers respond to these items across all their classes the data become less meaningful, especially when we think about the use of fidelity data. If a use of fidelity data, in this case assessing participant responsiveness, is to improve student engagement than knowing specifically what classes student engagement needs improvement is key. Given the era of accountability we are in and the push for pay–for performance in education, fidelity data that are meaningful is the key to understanding how interventions are delivered and program outcomes.

Contributions to the Literature

Although the results from this study were mixed, this study has made some important contributions to the literature in measurement and fidelity. As was mentioned previously in the review of the literature, organizational researchers are typically savvy and know to use multilevel analyses in the study phase of their research, but fewer attend to multilevel analyses in the psychometric phase. In many studies where there is nesting, the scale scores for individuals are used, ignoring the clustering or grouping within an organization. Other studies have averaged the individual responses to come up with a group mean, thereby ignoring the variability in individual responses. This is problematic in that student perception of a teacher's use of specific teaching practices may reflect differences among teachers/classes, but may also reflect differences among students who share membership in the same class. A multilevel analysis

enables one to adjust for the effects of variables measured at the individual level when estimating effects of variables measured at the teacher/class level (Raudenbush et al., 1991). This study attended to the multilevel analyses in the psychometric phase of the study, which provides an important contribution to the measurement literature. Additionally throughout this study, single level psychometric analyses were also conducted so that readers could compare findings and better understand the implications when interpreting results.

With respect to assessing fidelity, this study is an important contribution to the literature for multiple reasons. First, as was described in the review of the literature, fidelity measures are typically not validated. This study collected validity data in an attempt to validate the student fidelity measure (but the data did not provide strong support for the validity of these scores). Also this study used multilevel psychometric approaches to validate the measure. The larger study from which these data are from only planned to do single level psychometric analyses. The use of multilevel psychometric analyses for fidelity, as was done in this study is important, as the assessment of fidelity typically occurs in settings in which nesting is inherent. Second, this study used multiple sources (teachers and students) to assess fidelity. As is noted in the literature, the use of multiple sources and methods can serve to increase reliability and validity in fidelity ratings (Emshoff et al., 1987; Ruiz-Primo, 2005; Summerfelt, 2003; Vartuli & Rohs, 2009; Zvoch, Letourneau, & Parker, 2007). Third, when fidelity is typically assessed in a study, the dimensions of fidelity studied are dosage and adherence, and sometimes quality. This study was focused on examining the participant responsiveness dimension (engagement) of fidelity. The teacher and student questionnaires included items related to instructional student engagement and teaching practices, which presented an opportunity to not only understand whether/ how teaching practices were delivered by teachers, but also whether/how it was

received by students. The final contribution to the literature on fidelity is the use of consumers (students) as fidelity informants. In addition to limiting teacher self-report bias, an added advantage to using consumer self-reports of fidelity is that some information (like student interactions and student engagement) may not be attainable from anywhere else besides directly from the consumer (Baldwin, 2000). Consumers are not going to know about all the activities going on in a program (Mowbray et al., 2006), but when examining the process or interactional piece of fidelity (participant responsiveness, engagement) consumers are likely to be the best source. Assessing participant responsiveness from the perspective of the participant may provide a more feasible, more objective, and less biased method of assessing fidelity when studying process and interaction, compared to observation and teacher self-report.

Recommendations for Future Research

This student fidelity of implementation instrument is still in its first generation of development. The student questionnaire was developed in fall of 2011 and then cognitive interviews and pilot testing of the measure followed in the spring of 2012. The student questionnaire was revised based on the cognitive interview feedback from students and the pilot testing, but given the smaller pilot sample, multilevel analyses could not be conducted and so revisions to the instrument may have been limited. This study began the process of providing information related to the multilevel validity of these scores. The results and conclusions from this study on the psychometric properties of the instrument can guide the further development of this instrument.

Future researchers should consider the use of consumers when assessing fidelity. Although the convergent validity findings were mixed and reliability was low, so it is possible that there was attenuation, positive, but not strong correlations were found for teacher and

student scores on Instructional Pedagogy items. This may imply that some relationship exists between what practices teachers say they deliver and what students perceive they are receiving. These correlations possibly might have been stronger had teachers completed a teacher questionnaire for each class. This finding though holds some promise for the use of students (consumers) in assessing fidelity.

Although this study used two sources for fidelity data, the study could have been strengthened by an additional data source and data triangulation. In the larger study, observation was used to assess teachers' fidelity of implementation but observations could not be connected to the specific classes for students who participated in the study and items did not overlap. Future researchers may want to consider adding observers as a source, so that convergent validity between observer ratings and student and teacher ratings can be computed. This additional source may provide empirical evidence that consumers can be used as informants to assess fidelity.

Given the mixed results, as a first step, future researchers who want to further develop these scales will need to determine if the present results will be replicated with new samples of students and teachers. This was only one application of the student measure and the mixed model fit could be due to other factors. Also, since the purpose of this study was to test the a priori model, modifications were not made. Following replication, if results remain unclear or there is little variability, then future researchers may want to bring together an expert panel to review items, determine what were the weaknesses in the models (i.e., weak loadings, strong correlations of errors, loadings on more than one factor) and make decisions about item revision, addition, and deletion. Future researchers may determine there is sufficient evidence from this that modification to the instrument will be necessary and may instead begin with pulling together

an expert panel, for the purpose of reviewing the results and determining the weakness in the models. Following that they may decide to replicate the study, but put certain items on a watch list. With respect to reliability, given the study findings, future researchers will need to consider whether the addition of items is warranted for strengthening reliability at the individual level or whether reliability should be strengthened at the group level with additional respondents.

Future researchers should carefully consider what was presented in this study with respect to the procedures, methods and sources used to collect fidelity data and its potential impact on the findings. In terms of psychometric analyses for fidelity instruments, there is still very little reported in the literature. Future research on fidelity should include the psychometric analyses of fidelity instruments. To increase reliability and validity when developing fidelity instruments, future researchers should attend to the levels in which their respondents may be nested and plan their research to include multiple data sources and data triangulation. Future research on fidelity assessment should attend to the issues of multiple levels of analysis in both the psychometric and study phase. "Although use of multilevel modeling is not without complications and complexities remain regarding interpretation of the resulting reliability estimates, the ability to examine those estimates at multiple levels of analysis allows for the use of theories and investigation of effects in which individual-and group-level processes are distinguished" (Bonito et al., 2012, p. 461).

REFERENCES

Abbott, R., O'Donnell, J., Hawkins, J., Hill, K., Kosterman, R., & Catalano, R. (1998). Changing teaching practices to promote achievement and bonding to school. *American Journal of Orthopsychiatry, 68*(4), 542 - 552.

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Burlington, VT: Department of Psychiatry, University of Vermont.

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

Backer, T. E. (2001). *Balancing program fidelity and adaptation in substance abuse prevention: A state-of-the-art review*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Baldwin, R. L. (2000). Implementation of an evidence-based guideline for bronchiolitis. *Journal of Investigative Medicine, 48*(1), 168A-168A.

Becker, D. R., Smith, J., Tanzman, B., Drake, R. E., & Tremblay, T. (2001). Fidelity of supported employment programs and employment outcomes. *Psychiatric Services, 52*, 834-836.

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science, 12*(1), 23-33.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D. B., & Emshoff, J. G. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology, 15*, 253-268.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 1*(4), 355-373.

Bond, G., Williams, J., Evans, L., Salyers, M., Kim, H. W., Sharpe, H., & Leff, H. S. (2000). *Psychiatric rehabilitation fidelity toolkit*. Cambridge, MA: Human Services Research Institute.

Bond, G. R., Becker, D. R., Drake, R. E., Rapp, C. A., Meisler, N., Lehman, A. F., . . . Blyler, C. R. (2001). Implementing supported employment as an evidence-based practice. *Psychiatric Services, 52*(3), 313-322.

Bonito, J. A., Ruppel, E. K., & Keyton, J. (2012). Reliability estimates for multilevel designs in group research. *Small Group Research, 43*(4), 443-467.

Botvin, G. J., Griffin, K. W., Diaz, T., & Ifill-Williams, M. (2001). Preventing binge drinking during early adolescence: One- and two-year follow-up of a school-based preventive intervention. *Psychology of Addictive Behaviors, 15*, 360-365.

Botvin, G., Baker, E., Dusenbury, L., Botvin, E., & Diaz, T. (1990). Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *Journal of the American Medical Association, 273*, 1106 - 1112.

Calsyn, R. J. (2000). A checklist for critiquing treatment fidelity studies. *Mental Health Services Research, 2*(2), 107-113.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(1), 40.

Century, J., Rudnick, M., & Freeman, C. (2010). A Framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge source. *American Journal of Evaluation, 31*(2), 199-218.

Century, J. R., Mollie; Freeman, Cassie. (2008). Accumulating knowledge on elementary science specialists: A strategy for building conceptual clarity and sharing findings. *Science Educator, 17*(2), 31-44.

Chambers, C. T., & Johnston, C. (2002). Developmental differences in children's use of rating scales. *Journal of Pediatric Psychology, 27*(1), 27-36.

Clarke, G. (1995). Improving the transition from basic efficacy research to effectiveness studies: Methodological issues and procedures. *Journal of Consulting and Clinical Psychology, 63*, 718-725.

Cronbach, L. J. (1971). Test Validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23-45.

Dedrick, R. F., & Greenbaum, P. E. (2011). Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Emotional and Behavioral Disorders, 19*(1), 27-40.

Domitrovich, C. E., & Greenberg, M. T. (2000). The study of implementation: Current findings from effective programs that prevent mental disorders in school-aged children. *Journal of Educational and Psychological Consultation, 11*(2), 193-221.

Donaldson, S., & Grant-Vallone, E. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology, 17*(2).

Durlak, J., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327-350.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237-256.

Dusenbury, L., Brannigan, R., Hansen, W., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research, 20*(3), 308 - 313.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly, 16*(1), 149-167.

Elliott, D., & Mihalic, S. (2004). Issues in disseminating and replicating effective prevention programs. *Prevention Science, 5*(1), 47 - 53.

Emshoff, J. G., Blakely, C., Gottschalk, R., Mayer, J., Davidson, W., & Erickson, S. (1987). Innovation processes in education and criminal justice: Measuring fidelity of implementation and program effectiveness. *Education Evaluation and Policy Analysis*, 9, 300-311.

Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy, 36*(1), 3-13.

Fullan, M. (2001). *The new meaning of educational change* (3rd ed.). New York, NY: Teachers College Press.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*(1), 79-88.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2013). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72-91.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C. R., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. Exceptional Children, 71, 149-164.

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. (1999). Opening the black box: Using process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology, 27*(5), 715-735.

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children, 79*(2), 181+.

Henggeler, S. W., Schoenwald, S. K., Liao, J. G., Letourneau, E. J., & Edwards, D. L. (2002). Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child & Adolescent Psychology, 31*(2), 155-167.

Hernandez, M., Gomez, A., Lipien, L., Greenbaum, P. E., Armstrong, K. H., & Gonzalez, P. (2001). Use of the system-of-care practice review in the national evaluation: Evaluating the fidelity of practice to system-of-care principles. *Journal of Emotional and Behavioral Disorders, 9*(1), 43-52.

Hox, J. J. (2002). *Multilevel analysis techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York: Routledge.

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

James Bell Associates. (2009, October). *Evaluation brief: Measuring implementation fidelity*. Retrieved from http://www.jbassoc.com/ReportsPublications/EvaluationBrief-MeasuringImplementationFidelity_Octob.pdf

Kalafat, J., Illback, R. J., & Sanders Jr, D. (2007). The relationship between implementation fidelity and educational outcomes in a school-based family support program: Development of a model for evaluating multidimensional full-service programs. *Evaluation and Program Planning, 30*(2), 136-148.

Kelly, J. A., Somlai, A. M., DiFranceisco, W. J., Otto-Salaj, L. L., McAuliffe, T. L., Hackl, K. L., . . . Rompa, D. (2000). Bridging the gap between the science and service of HIV prevention: Transferring effective research-based HIV prevention interventions to community AIDS service providers. *American Journal of Public Health, 90*(7), 1082-1088.

Lillehoj, C. J., Griffin, K. W, & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education and Behavior, 31*(2), 242-257.

Lucca, A. M. (2000). A clubhouse fidelity index: Preliminary reliability and validity results. *Mental Health Services Research, 2*(2), 89-94.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120-131.

Lynch, S., Taymans, J., Watson, W. A., Ochsedorf, R. J., Pyke, C., & Szesze, M. J. (2007). Effectiveness of a highly rated science curriculum unit for students with disabilities in general education classrooms. *Exceptional Children, 73*(2), 202-223.

Macias, C., Propst, R., Rodican, C., & Boyd, J. (2001). Strategic planning for ICCD clubhouse implementation: Development of the clubhouse research and evaluation screening Survey (CRESS). *Mental Health Services Research, 3*(3), 155-167.

Mantzicopoulos, P., Patrick, H., & Samarapungavan, A. (2008). Young children's motivational beliefs about learning science. *Early Childhood Research* Quarterly, 23, 378–394.

McGrew, J. H., Pescosolido, B., & Wright, E. (2003). Case managers' perspectives on critical ingredients of assertive community treatment and on its implementation. *Psychiatric Services, 54*(3), 370-376.

Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth, 4*, 83 - 105.

Mihalic, S., Fagan, A., Irwin, K., Ballard, D., & Elliott, D. (2002). *Blueprints for violence prevention replications: Factors for implementation success*. Boulder, CO: Center for the Study and Prevention of Violence.

Mowbray, C. T., Bybee, D., Holter, M., & Lewandowski, L. (2006). Validation of a fidelity rating instrument for consumer-operated services. *American Journal of Evaluation, 27*(1), 9-27.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315-340.

Muthén, B. (1994). Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study. In I. Westbury, C. Ethington, L. Sosniak, & D. Baker (Eds.), *In Search of more effective mathematics education: Examining data from the IEA Second International Mathematics Study* (pp. 293-324). Norwood, NJ: Ablex.

Myers, N. D., Feltz, D. L., Maier, K. S., Wolfe, E. W., & Reckase, M. D. (2006). Athletes' evaluations of their head coach's coaching competency. *Research Quarterly for Exercise and Sport, 77*(1), 111-121.

National Research Council. (2004). On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations. Washington, DC: The National Academies Press.

Noel, P. (2006). The impact of therapeutic case management on participation in adolescent substance abuse treatment. *American Journal of Drug Alcohol Abuse, 32*, 311 - 327.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

Paulson, R. I., Post, R. L., Herinckx, H. A., & Risser, P. (2002). Beyond components: Using fidelity scales to measure and assure choice in program implementation and quality assurance. *Community Mental Health Journal, 38*(2), 119-128.

Pentz, M. A., Trebow, E. A., Hansen, W. B., MacKinnon, D. P., Dwyer, J. H., Johnson, C. A., . . . Cormack, C. (1990). Effects of program implementation on adolescent drug use behavior: The Midwestern Prevention Project (MPP). *Evaluation Review, 14*(3), 264-289.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*: New York, NY: Sage.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to US high-school data. *Journal of Educational and Behavioral Statistics, 16*(4), 295-330.

Resnick, B., Inguito, P., Orwig, D., Yahiro, J. Y., Hawkes, W., Werner, M., . . . Magaziner, J. (2005). Treatment fidelity in behavior change research - A case example. *Nursing Research, 54*(2), 139-143.

Resnick, B., Neale, M., & Rosenheck, R. (2003). Impact of public support payments, intensive psychiatric community care, and program fidelity on employment outcomes for people with severe mental illness. *Journal of Nervous and Mental Disease, 191*, 139 - 144.

Ringwalt, C. L., Ennett, S., Johnson, R., Rohrbach, L. A., Simons-Rudolph, A., Vincus, A., & Thorne, J. (2003). Factors associated with fidelity to substance use prevention curriculum guides in the nation's middle schools. *Health Education & Behavior, 30*(3), 375-391.

Rogers, E. M. (1995). *Diffusion of Innovations* (4 ed.). New York, NY: The Free Press.

Rohrbach, L. A., Dent, C. W., Skara, S., Sun, P., & Sussman, S. (2007). Fidelity of implementation in Project Towards No Drug Abuse (TND): A comparison of classroom teachers and program specialists. *Prevention Science, 8*(2), 125-132.

Sanetti, L., & Kratochwill, T. (2008). Treatment integrity in behavioral consultation: Measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy, 4*(1), 95-114.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*(3), 259-280.

Skara, S., Rohrbach, L. A., Sun, P., & Sussman, S. (2005). An evaluation of the fidelity of implementation of a school-based drug abuse prevention program: Project Toward No Drug Abuse (TND). *Journal of Drug Education, 35*(4), 305-329.

Snyder, P. A., Hemmeter, M. L., Fox, L., Bishop, C. C., & Miller, M. D. (2013). Developing and gathering psychometric evidence for a fidelity instrument: The Teaching Pyramid Observation Tool–Pilot Version. *Journal of Early Intervention*, *35*(2), 150-172.

Teague, G. B., Drake, R. E., & Ackerson, T. H. (1995). Evaluating use of continuous treatment teams for persons with mental illness and substance abuse. *Psychiatric Services, 46*, 689-695.

Institute of Education Sciences. (2014). *What works clearinghouse: About us*. Retrieved from http://ies.ed.gov/ncee/wwc/AboutUs.asp

Weisman, A. G., Okazaki, S., Gregory, J., Tompson, M. C., Goldstien, M. J., Rea, M., & Miklowitz, D. J. (1998). Evaluating therapist competency and adherence to behavioral family management with bipolar patients. *Family Process, 37*, 107-121.

Zvoch, K. (2012). How does fidelity of implementation matter? Using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation*, *33*(4), 547-565.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice, 12*(2), 127.

APPENDIX A
INTERVENTION DESCRIPTIONS

Full Option Science System (FOSS): FOSS is a K–8 hands-on science curriculum created here at the Lawrence Hall of Science with support from the National Science Foundation. The FOSS national program (2005 edition) includes 35 modules and/or courses organized under four strands:

- Life Science
- Physical Science
- Earth Science
- Scientific Reasoning and Technology

The FOSS CA K–5 Program (2007 edition, state adopted) consists of 18 modules, three for each grade level. Program components include an extensive teacher guide, equipment kits, teacher preparation videos, and science resource books for students, and multimedia access. Delta Education publishes and distributes FOSS.
http://www.lawrencehallofscience.org/programs_for_schools/programs/foss

Science & Technology for Children: STC™ : Since 1988, the National Science Resources Center (NSRC) has been developing Science and Technology for Children (STC), an innovative hands-on science program for children in grades one through six.

The 24 units of the STC program, four for each grade level, are designed to provide students with stimulating experiences in the life, earth, and physical sciences and technology.

http://www.sempcoinc.com/scandteforch.html

Science Companion: Science Companion is curriculum for teachers, by teachers, built from the some of the strongest pedagogical constructs in hands-on learning in the world.

The Science Companion curriculum, developed by the Chicago Science Group (CSG) is a hands-on learning program that takes advantage of children's extensive knowledge of – and curiosity about – how things work in the world. The purpose of the curriculum is not only to provide children with the opportunity to wonder about their world, but to teach them science processes as they explore, quantify, and interpret the world. The children are also given the time and encouragement to draw, write, discuss, and reflect upon what they have done. The program's approach to primary education balances discovery-based learning with teacher-directed instruction.

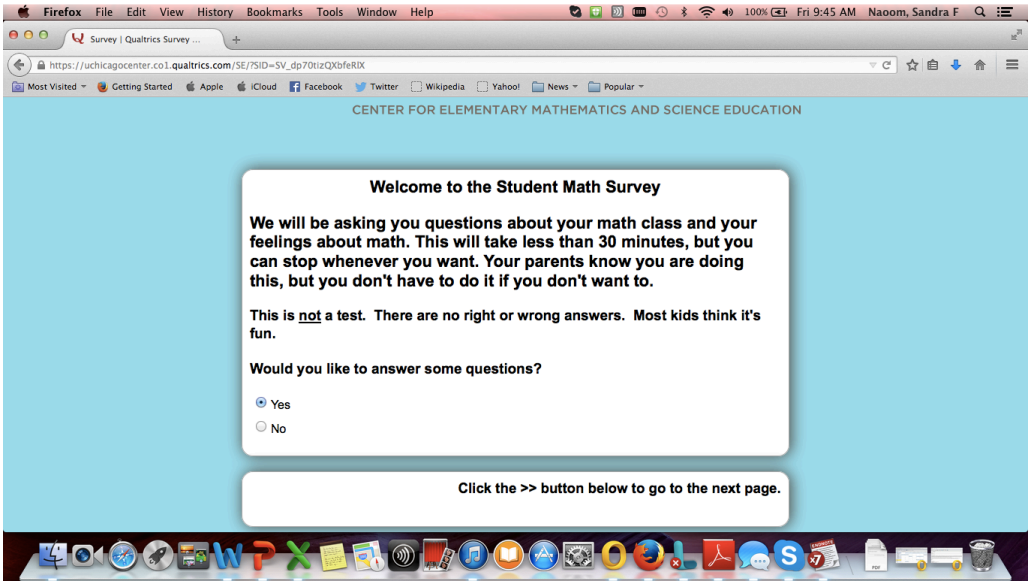
http://www.sciencecompanion.com/science-companion-story/

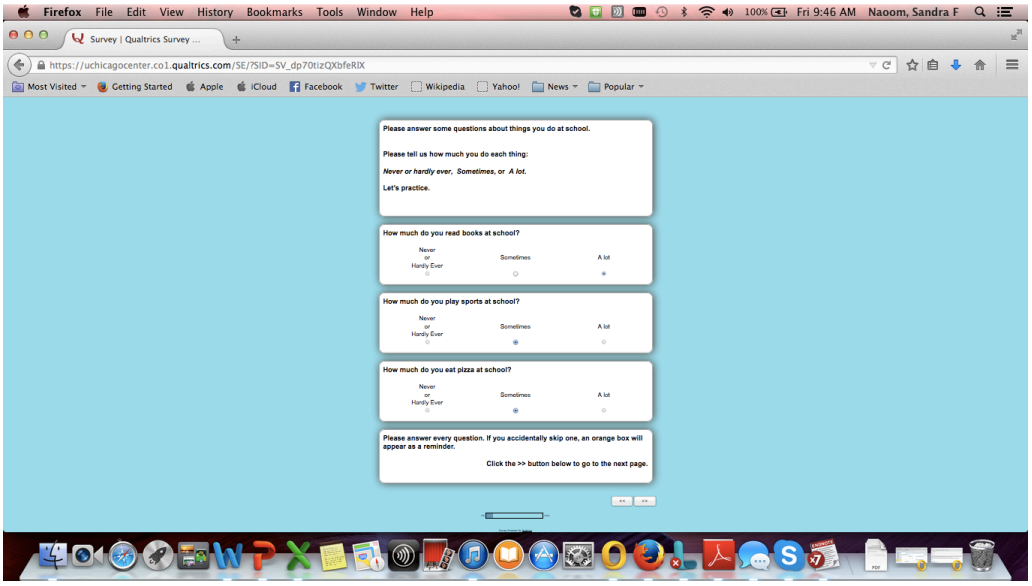
Everyday Mathematics: Everyday Mathematics is a comprehensive Pre-K through 6th grade mathematics curriculum developed by the University of Chicago School Mathematics Project and published by McGraw-Hill Education. It is currently being used by about 4.3 million students in over 220,000 classrooms.

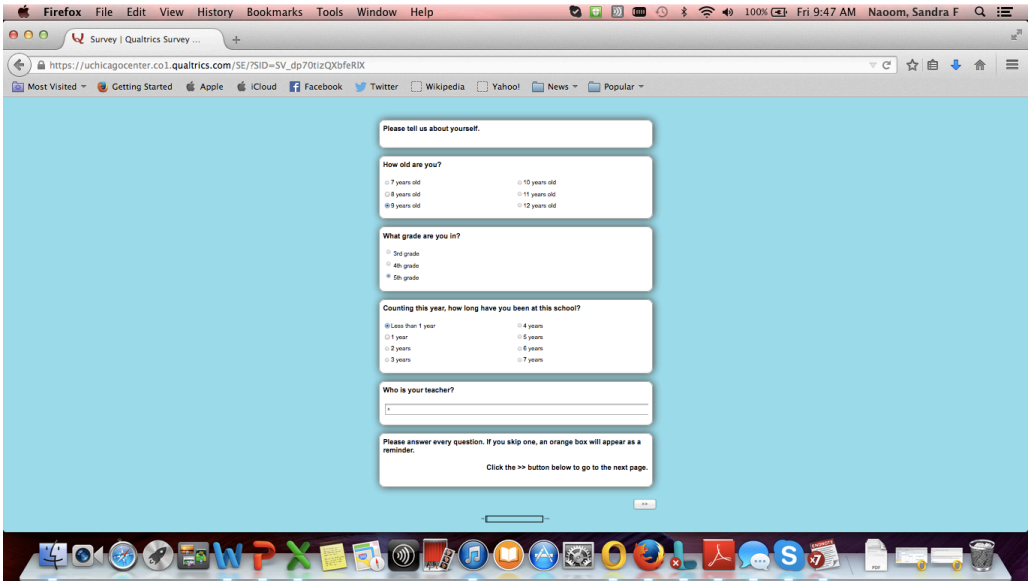
Everyday Mathematics is organized around Grade-Level Goals, Program Goals, and Content Strands. To guide curriculum development, the original Everyday Mathematics authors formulated a set of beliefs and principles based on research about what worked best in other

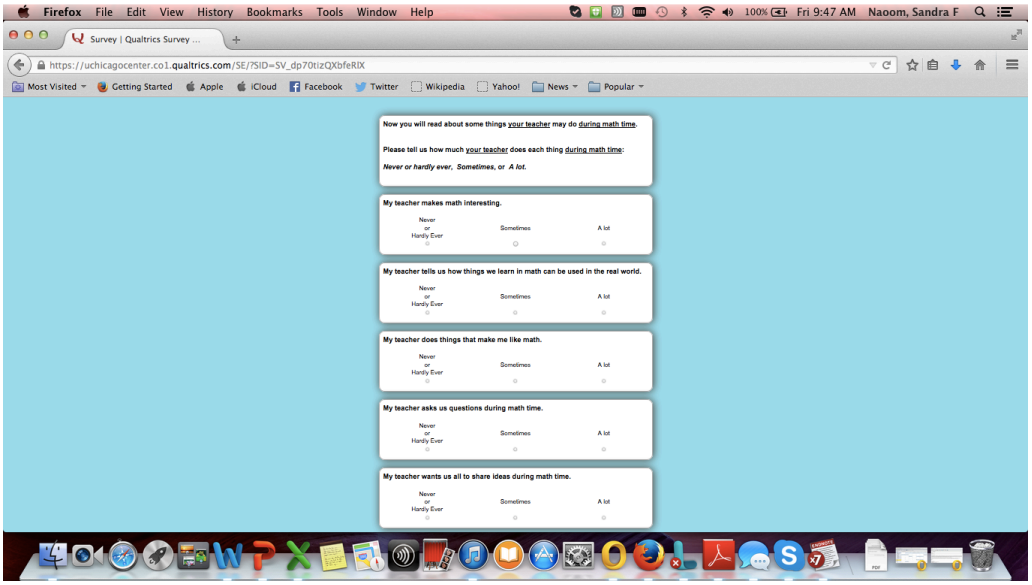
countries and in the authors' own field research. Based on these principles, the original Everyday Mathematics authors identified guidelines for the best teaching methods that help children build a strong mathematical foundation in their elementary education years. Based on this philosophy, the Everyday Mathematics authors created a curriculum that featured several specific pedagogical principles. These principles include emphasizing appropriate use of technology, teaching real-life problem solving, improving home/school partnership, and more. http://everydaymath.uchicago.edu/about/understanding-em/

APPENDIX B
STUDENT INSTRUMENT SCREEN SHOTS









APPENDIX C
IRB APPROVAL LETTER

USF
UNIVERSITY OF
SOUTH FLORIDA

DIVISION OF RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-5618

April 6, 2012

Sandra Naoom
Edu Measurement & Research
2025 Fluorshire Drive
Brandon, FL 33511

RE: **Not Human Research Activities Determination**

Activity Title: Validation of a Student Fidelity Measure

Dear Ms. Naoom:

I have reviewed the information you provided regarding the above referenced project and have determined the activities do not meet the USF definition of human subjects research activities; therefore, IRB approval is not required. If, in the future, you change this activity such that it becomes human subjects research activities, prior IRB approval is required. If you wish to obtain a determination about whether the activity, with the proposed changes, will be human research activities, please contact the IRB Office for further guidance.

All research activities, regardless of the level of IRB oversight, must be conducted in a manner that is consistent with the ethical principles of your profession and the ethical guidelines for the protection of human subjects. As principal investigator, it is your responsibility to ensure subjects' rights and welfare are protected during the execution of this project

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John A Schinka, Ph.D.

John Schinka, Ph.D., Chairperson

ABOUT THE AUTHOR

Sandra F. Naoom, M.S.P.H, Ph.D. Candidate, is Associate Director of the National Implementation Research Network (NIRN), located at the Frank Porter Graham Child Development Institute at the University of North Carolina in Chapel Hill. Sandra is a co-author of the highly regarded monograph Implementation Research: A Synthesis of the Literature and has spent the last ten plus years working with the NIRN to establish and build the science and practice of implementation. At NIRN, she leads projects and provides consultation, evaluation and technical assistance around implementation of innovations with fidelity. Sandra's areas of focus are in developing strategies for implementation in underdeveloped and under-resourced global settings, supporting organizations in the development of their implementation capacity through implementation system and infrastructure evaluation and planning, developing measures to assess fidelity and implementation, and studying the supply and demand issues involved in the implementation of innovations in various contexts.