REFINEMENT OF A MEASURE OF DATA USE

PRACTICES

By

SHANNON HAZEL STARK GUSS

Bachelor of Science in Sociology
Langston University
Tulsa, OK 2004

Master of Arts in Liberal Studies
University of Oklahoma
Norman, OK 2012

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2019

REFINEMENT OF A MEASURE OF DATA USE

PRACTICES

Dissertation Approved:

Dr. Laura Barnes

Dissertation Adviser

Dr. Ki Cole

Dr. Mwarumba Mwavita

Dr. Amanda Sheffield Morris

ACKNOWLEDGEMENTS

I am profoundly grateful to my husband, Alex Guss, who has supported me by being a single dad to our three children on days when I worked late into the night. I am thankful to my father, Gary Stark, for instilling in me a love of math and science, of critical thinking, and of learning; to my mother, Marti Stark, for teaching me to work hard, make plans, and be persistent – even when the first set of plans don't work out like you expected. I honor my brothers and sister, who anchor my identity in this world, reminding me when I need to be humble and when I need to get tough. I also want to thank my grandmothers for being examples to which I continuously aspire; Jean Ingram – mixing grit and humor to succeed in life's adventures and struggles, and Betty Beachler Wells – always poised and prepared to make dreams a reality.

I also want to thank those who have partnered in this project. The measure development team included Amanda Stein, Mary Sweet-Darter, and Brian Mangus. The Technical Expert Panel included Curt Adams, Tim Ford, Laura Barnes, Diane Horm, Kyong-Ah Kwon, Jordan Ware, Vicki Lake, Todd Reeves, and James Spillane. The Advisory Board across all years included Jennifer Ladner, Serenity Weeden, Lori Jablonsky, Noreen Yazejian, and Deborah Norris. This work was generously funded by the Buffett Early Childhood Fund.

I would also like to thank the teaching and leadership staff at sites across the Educare Learning Network, Oklahoma Public School teachers, Head Start and Early Head Start, and other teachers who found the study through social networks and contributed to this study with their time and experience. I am also grateful to the students at University of Oklahoma – Tulsa and the staff of the Early Childhood Education Institute who contributed their experiences as classroom teachers to the pre-testing process.

Finally, I am incredibly lucky to have had the opportunity to work with a number of excellent mentors and advisors. I am grateful to each of my committee members, who have contributed to unique and important parts of my academic journey; my patient and persistent advisor - Laura Barnes, IRT professor - Ki Cole, methods professor - Mwarumba Mwavita, and Amanda Sheffield Morris, Human Development and Family Sciences professor. I am also thankful to the researchers of the Educare Learning Network who have encouraged my enthusiastic curiosity. Without doubt, I would not be attempting this without those who have opened doors and pushed me to seek a ceiling to my potential, particularly my closest mentors, Diane Horm and Deborah Norris. While some would say that anything that exists in any amount is measurable, I am grateful beyond measure to you.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: SHANNON STARK GUSS

Date of Degree: DECEMBER, 2019

Title of Study: REFINEMENT OF A MEASURE OF DATA USE PRACTICES

Major Field: EDUCATIONAL PSYCHOLOGY

Abstract: The purpose of this study is to refine a measure of data use in early childhood schools, focusing on one section of a set of measures used to capture data use by teachers in schools serving children from birth to age 5. Data are defined as information collected purposefully by teachers or by others. Primary examples of data in early childhood settings include child assessment data, data on classroom quality, and data on families. A framework applied to refine items included learning from data at a low level to responding based on data at a high level. Items were developed to coincide with these levels.

Items were field tested through cognitive interviews and behavioral coding, and presented to panels of experts. These processes informed a version that was collected in the spring of 2019 with a sample of 359 teachers. The sample included early childhood teachers employed by a national network of early childhood programs that emphasize the use of data, early childhood teachers involved in similar evaluations, and teachers recruited from social networks. The purpose of this study was to review the psychometric qualities of six scales using factor analysis and item response theory (IRT) applications.

Results/conclusions: The study found that five of the six scales of the Informational Use set had acceptable psychometric qualities. Scales measuring the use of teacher-collected child assessment data and teacher-collected classroom quality data had items that had high estimates of discrimination and measured well across the latent spectrum. These are considered sufficient for use in research and professional development. Scales focused on other-collected child assessment data and other-collected classroom quality data had sufficient discrimination and bandwidth, but are considered primarily useful for descriptive purposes due to item fit issues that need additional study. One of the two Family scales is also sufficient for descriptive use. However, teacher-collected family data did not have sufficient evidence to recommend use as a scale. Model fit statistics had mixed results for all but one scale – the teacher-collected classroom data scale.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

When educational practices are considered important, they should be measured and included in the study of teaching and learning. Data use is a practice that has shown to have benefits in K-12 educational settings. Particularly since the No Child Left Behind (NCLB) Act, data have been infused into all aspects of schools' systems in the United States through teacher observation, student achievement tests, and student surveys. Observations of teachers and assessments of children's learning and development are also common in classrooms and programs serving very young children – from birth to age 8. While little research has focused on data-based decision-making in early childhood classrooms, it is considered a key part of successful early childhood education (Derrick-Mills, 2015; NCQTL, 2011; Sandstrom, Isaacs, & Rohacek, 2015). Research is needed on this key practice in early childhood classrooms but is difficult to conduct efficiently without a measure. Thus, it is important that the practice of data use be measured in early childhood classrooms by a measure that is adapted to this specific educational setting.

The benefits of data utilization in K-12 were reviewed by Stecker, Fuchs, & Fuchs (2005), who found that curriculum-aligned assessment was used effectively by teachers to bolster teaching practices, and thereby, student outcomes. For early childhood settings, this kind of research is less common than it is for K-12 schools. However, the use of data was written into the standards for the regulation of the large, federally-funded early childhood program – Head Start – in 2007, requiring grantees to set school readiness goals (NCPMFO, 2013). Child care centers and schools that receive Head Start funding are expected to use data from child assessments conducted in a previous year to set school readiness goals for the upcoming school year (Sandstrom, Isaacs, & Rohacek, 2015). These goals are thereby expected to drive practices and professional development plans of the school. Due to the expectation that data use should be part of early childhood education practice, as it is in education systems for older children, the measurement of data use in early childhood education is important. Only when data use practices are defined and understood through effective measurement will the empirical evidence for the benefits of data use to early childhood educational practices be understood. The measurement of data use in early childhood will allow for the empirical exploration of both the variation and contribution of this practice to the learning and development of young children.

Defining Data Use in Early Childhood Settings

A practical problem within the early childhood field is defining the construct of data use**.** High quality early childhood education and care (ecec) has been associated with positive outcomes for children (Boyd, Barnett, Bodrova, Leong, & Gomby, 2005; Campbell et al., 2008; NICHD Early Childcare Research Network, 2002), but whether data use is one of the components of high quality early childhood teaching practice is unclear. As the research in early childhood education evolves, it is important to clearly define the types of practices that are hypothesized to support children's learning and development, as well as whether and how they each uniquely contribute to outcomes.

Research in early childhood education tends to focus on classroom quality, or teacher-child interactions. The practice of data use has received attention more recently, as new standards for assessment and evaluation have been put in place. While there are many measures for classroom quality and teacher-child interactions, there is only one other measure of data use developed specifically for early childhood programs. This may be, in part, due to a lack of clarity on what data use means within early childhood practice in comparison to what it might mean in general educational settings. While K-12 settings have a variety of curriculum-aligned assessments developed for children – across both age groups and learning content, child assessments of young children are more difficult for a variety of reasons. Young children are developing rapidly, therefore children of a similar chronological age may have widely varying levels of skills. Additionally, early education and care is not as focused on learning content as it is on exposing children to a wide variety of developmentally appropriate opportunities. Further, unique types of data are frequently used in early childhood that are not used as frequently in K-12, such as data about classroom quality and data about families. Because young children are difficult to assess, getting information from families and from viewing classroom interactions is often used to inform teacher decisions. Thus, the types of data that are used in early childhood are unique and this must be taken into consideration when defining data use.

Because studies about data use often focus on data-based decision-making, this study defines data use as the application of systematically collected information that is gathered for the purpose of decision-making in schools (Gottfried, Ikemoto, Orr, & Lemke, 2011; Spillane, 2012; Walker, Carta, Greenwood, & Buzhardt, 2008). Data use in schools is similar to data use in other professions. Doctors use patient data to diagnose and treat. Counselors use individual assessments to better understand a set of symptoms from which a client is suffering. Meteorologists use patterns and trends in weather, coupled with their training about those patterns, to predict weather and advise action. Likewise, teachers can use data to make decisions about what to teach, how to teach it, to whom, at what time, and under what context. The use of data is a specific behavior that varies among professionals and can be measured.

Research, Evaluation, and Data Utilization

Data use, for the purpose of this study, is specific to a local context, such as a single child or a single classroom. This differentiates it from both research and evaluation. Data use in schools differs from research, as research seeks to use data to generalize to the wider public (Bickman & Rog, 2009). As data use in classrooms is localized and specific to one part of a school, it is not intended to be generalizable outside of this intended use. Data use also differs from program evaluation. While program evaluation often uses multiple connected pieces of data to understand how a program operates as a whole system, data use often applies a specific set of data to inform pieces of an educational program – such as a differentiated lesson plan for an individual student, lesson plans for a classroom, professional development for teachers, or materials and activities recommended to the families of young children (Fitzpatrick, Sanders, & Worthen, 2011;Krugly, Stein, & Centeno, 2014).

As an example of the differences between research, program evaluation, and data utilization, consider data that shows that adults facing food insecurity often feed their children rather than themselves and that the parenting styles of these adults are often characterized as harsh. In research, this finding would support a synthesis of family stress theory and risk and resilience theory (Patterson, 2002). The parents who go hungry rather than allow their children to go hungry are protective factors supporting their children's resilience (Henry, Morris, & Harrist, 2015). However, stress and hunger reduce parents' emotional capacities for patience, sensitivity, and self-awareness, and may increase the level of harsh interactions between parent and child (McCubbin, 1993). The generalizability would extend to food insecure parents, with wide-ranging implications for policy, practice, and additional evidence to support a theoretical perspective. In contrast, a program evaluation that found this pattern would draw implications based on the evaluation approach and program stakeholders' goals. If food insecurity was the focus of the program, then food might be provided with parenting outcomes receiving little attention. However, if the focus of the program was parenting, then community partners might be accessed to reduce food insecurity among program participants as a short-term output on the path to long-term outcomes (Fitzpatrick, Sanders, & Worthen, 2011).

The difference in data use is in the specificity with which the information can be applied. If a classroom teacher did a questionnaire with his students' parents and found that two of twenty mentioned that meeting basic needs is a stressor in their lives and that they wish they could be more laid back with their kids, then this is not generalizable to a wider population of students nor is it evidence for a school to change the way it interacts with the majority of parents. The teacher might, however, sign the two specific children up for a program that sends food home on the weekend or connect the parent to a local food bank. Data use is characterized by learning, considering choices, and tailoring responses to specific information.

Data Use Process as a Theoretical Framework

In schools, data use is characterized by data-informed goals that guide the focus of learning and drive concerted efforts of schools (Christman et al., 2009). Thus, data use is optimized when it occurs within a social context that facilitates the flow of information and cultivates informed practice as a pedagogical norm (Gerzon & Guckenberg, 2015; Johnson, 2002; Love, Stiles, Mundry, DiRanna, 2008; Spillane, 2012). Data use is characterized by learning and is more effective when applied with thoughtful pedagogical practice and age-appropriate content to intentional goals and specific plans (Bambrick-Santoyo, 2010; Means, Padilla, DeBarger, & Bakia, 2009). This can also be succinctly put as using data to build understanding, connecting knowledge, and responding – the basis of the construct framework used for this study, shown in Table 1.1 (Marsh & Farrell, 2015).

Table 1.1.

Developmental Framework of Data Use

Levels of Data Use	Definitions
Data as information	1. Data use contexts and experiences facilitate creation of information; supporting understanding and learning.
Data as an extension of knowledge	2. Data use contexts and experiences facilitate combination of data with other knowledge to generate ideas and new knowledge.
Data as a catalyst for action or stabilizer	3. Data use contexts and experiences facilitate application of ideas to create change or stasis in work behavior.

Note. Based on conceptualizations from Marsh & Farrell (2015), Spillane (2012).

The lack of theoretical clarity around data use at the individual teacher level and at the organizational school level hinder the accessibility of conversations about the potential benefits and consequences of data use. While many data use scholars have defined steps of using data, Marsh & Farrell (2015) also apply sociocultural learning theory to the process clarifying that data use occurs within social contexts through a process of shared meaning-making. Table 1.1 shows how the work of data use scholars has been applied to define a theoretical framework for data utilization that can be applied in early childhood settings (Marsh & Farrell, 2015; Spillane, 2012). The framework is primarily from the work of Marsh and Farrell (2015). The work of Spillane (2012) is also synthesized in the response portion – as data can be used to create *change* or *stasis*. That is, data does not always indicate that action needs to be taken.

Evidence for the Benefits of Data Use in Early Childhood

Data-based decision-making is the subject of many articles in education – both in early childhood education and general education (Gottfried, Ikemoto, Orr, & Lemke, 2011; Spillane, 2012; Walker, Carta, Greenwood, Buzhardt, 2008). While evidence exists of a relationship between the use of assessments for teacher decisions and stronger student achievement in K-12 education (see Stecker, Fuchs & Fuchs [2006] for a review), there is less evidence for this in early childhood education. Despite most evidence for data use coming from K-12 programs, the use of data has become expected in early childhood programs as well (Monahan, Atkins-Burnett, Wasik, Akers, Hurwitz, & Carta, 2016). Thus, it is possible that an educational practice that is beneficial in K-12 programs, but less beneficial in early childhood programs, is being encouraged without clear evidence of its value.

Statement of the Problem

The lack of measures of data use limits our ability to understand whether and how data use is important in the lives of young children. The lack of a measure of data use by teachers serving children birth to age 8 has practical, empirical, and theoretical problems for the understanding of teaching and learning in early childhood settings. A practical issue is the definition of data utilization in early childhood, which should be distinguished from data utilization in K-12 programs. Developing a measure of data utilization that is unique to early childhood will ameliorate a practical problem by capturing aspects of data use that are unique to early childhood programs. This type of measure could potentially inform professional development efforts.

Additionally, despite government-mandates on the use of data for program improvement in early childhood settings (e.g, Head Start programs), an understanding of both the presence and importance of data use in early childhood settings is limited due to the lack of measuring this important construct. Measuring data use in early childhood would facilitate generation of empirical evidence on the use of data in early childhood programs. The lack of a valid and reliable measure reduces the efficiency of collecting evidence on the relationship of data utilization to teaching and learning.

A theoretical issue that needs to be addressed is the development of a measure to test theory. If data use can be explained within current theory, a measure that is validly aligned to that theory will provide evidence to test the explanation empirically. Further, a measure based on a theoretical framework will have consistency in its content validity. Starting with theory benefits the research process by reducing the possibility that patterns found are found by chance and providing a clear path by which we can explain why and how we observe specific phenomena.

These practical, empirical, and theoretical problems can begin to be addressed through the development and testing of a measure of data use for early childhood programs.

Education researchers have a responsibility to students and teachers to clarify the most effective educational practices and discover how to support and implement practices in ways that maximize teachers' efforts and students' learning. While data use might be an effective educational practice in early childhood education, without a valid and reliable measure with which to use in research studies, the data use to child outcomes relationship (or lack thereof), remains a gap in our knowledge. Defining a construct with a theoretical framework solves a practical issue of clarity. Constructing a measure around a theoretical framework supports the validity of the measure. Further, the application modern measurement methods will support the sensitivity and specificity of the measure. The lack of measures capturing the variation of data use practices by teachers is a problem that hinders understanding of whether data use is important in the lives of young children. This problem is addressed in this study by refining and testing a measure developed for this purpose.

Measuring Data Use in Schools

The measurement of data utilization is complex. While the benefits of data utilization are clear in some instances, the clearest examples are in narrow instances of curriculum-based assessments being used to inform the depth and breadth of teaching that is aligned with the same curriculum (Stecker, Fuchs, & Fuchs, 2005). The ways in which information can be used become broader when moving away from curriculum-based assessments. For example, the use of research has been shown to have differing types of uses including use to change a practice (instrumental), use to confirm existing beliefs (symbolic), and use to change a view about a problem or practice (conceptual; Penuel, et al., 2017). The use of data has been theorized to have similar patterns, creating change or stasis through organizational routines (Spillane, 2012).

Additionally, the context in which data is used may motivate people toward differing actions depending on the perceived function of the data or *functional significance* (Ford, Van Sickle, Fazio Brunson, 2016). While there is some consensus in the literature that data use involves identifying problems, deciding on solutions, and following through with a response, these routines can occur within a variety of contexts (Firestone & Gonzalez, 2007; Marsh & Farrell, 2015; Spillane, 2012). One of the contexts is the type of data use culture that characterizes an organization – whether one of accountability or of learning (Firestone & Gonzalez, 2007; Gannon-Slater, LaLonde, Crenshaw, Evens, Greene, & Schwandt (2017). Macrostructures organizing the data use culture influence the perception of the function of data – whether the use of data have *controlling significance* or *informational significance* (Adams, Forsythe, Ware, & Mwavita, 2016; Ford et al., 2016). Accountability cultures would focus data use on monitoring and compliance, giving data controlling significance (Gannon-Slater et al., 2017; Ford et al., 2016). Organizational learning cultures would focus data use on identifying student needs and exploring pedagogical strategies to address those needs (Gannon-Slater, 2017). Hence, in organizational learning cultures, data has informational significance (Ford et al., 2016). Due to empirical support and theoretical foundations, the use of data as feedback that supports learning is regarded as more beneficial to educational practice, student outcomes, and teacher well-being. Thus, measuring the use of data that has informational significance, or *Informational Use*, is the focus of this study.

Currently, the only measure of data use developed for early childhood settings – the Examining Data Informing Teaching (EDIT) measure - requires interviewing, portfolio review, and video coding to gather enough data to complete a score (Monahan, et al., 2016). A survey measure of data use, while having the limitations of all self-report measures, may be more accessible to early childhood programs. A survey measure of data use would also allow early childhood programs to assess their schools' level of data use to inform professional development needs. A survey measure could also support program evaluation and research in early childhood settings – clarifying the contribution of this aspect of pedagogy. Specifically, a clearly defined measure of data use should allow researchers to understand the influence that the practice of data use has on young children's learning and development. This contribution may be difficult to separate from similar constructs of teacher sensitivity – the responsiveness of a teacher to the immediate needs of a child.

Because more effective data use entails organizational learning and concerted efforts, the process of applying data to learning and planning is a social effort at the organizational level of the school. This includes organizational routines and infrastructure that support the work of data use (Gerzon & Guckenberg, 2015; Spillane, 2012). Additionally, the social supportiveness of collaborative teams is an important piece of the data use process (Cosner, 2011; Ford et al., 2016). Finally, the function of data in a school will affect the data use process. While those aspects of data use context are covered in sections of the measure, the focus of this dissertation will be Informational Use of teacher-collected and other-collected data.

In addition to multiple contexts of use – whether collected by teachers or by others – educators also have multiple types of data that may be useful in making decisions about their classroom practices. Teachers can use child assessments to understand whether a child is struggling with specific learning objectives and adjust classroom practices to meet this need. Teachers can also use data from observations of their teaching practice to reflect on potential areas for improvement. In addition to applying data to children and classroom practice, education professionals, particularly those in early childhood programs, can apply data about children's families to their work as co-caregivers of young children. Thus, early childhood teachers may have different types of data – child, classroom, or family – that are collected in different contexts – either by teachers or by others – and are used in different school climates – either for controlling or informational purposes.

Overview of the Project

While data is considered important to education, including early childhood education, a measure of data utilization for research, evaluation, or professional development does not currently exist in a form that is accessible to the majority of early childhood programs. This project was originally funded as a response to this problem. It was expected to be a three-year project, from May 2015 and end May 2018. I applied as the principal investigator based on previous attempts at adapting a measure from self-determination theory literature to the process of data utilization in early childhood schools (i.e. Deci, Spiegel, Ryan, Koester, & Kauffman, 1982, Deci & Ryan, 2003, Ryan & Deci, 1987). I was joined on the project by two Ph.D. level colleagues and one school psychologist – all with previous experience in measurement development. As the lead developer, my role included writing the original grant, writing the year one literature review, designing the study timeline and plan of analysis, and recruiting participants. The other investigators provided feedback and editing of all project products (reports, power point slides, and conference presentations), consultation on design, and assistance with recruitment. The team and I reviewed analyses to make decisions on next steps and collaborated on item writing and re-writing. I received a no cost extension to end the project in October 2019. The co-investigators have been minimally involved in this extended arm of the project. An overview of the project, including what is and is not considered part of this dissertation study is described below.

The project aims are to develop a comprehensive set of measures to capture various aspects of early childhood teachers' beliefs about, attitudes toward, and utilization of data. The set of instruments is called the Data Use Contexts and Experiences Scales (DUCES). This dissertation study focuses on the portion of scales dedicated to Informational Use practices.

Pre-Dissertation. The generation of the DUCES set of measures has undergone two previous iterations – with this study focusing on the testing of a third iteration. Before describing the methodology for the testing of the third iteration, the development and testing of the first two iterations will be described.

Literature Review. Version 1 was based on an extensive literature review around four concepts of data use – organizational, individual, motivational, and practices. At this point in the measure development, a breadth of items were being generated based on what the literature suggested could be important to understanding data utilization. This literature review can be viewed in Appendix 1. Items were developed around these definitions, pretested, and piloted in the Summer of 2016.

Pilot 1. This administered survey (Version 1) had 64 items and collected qualitative information about the types of data teachers used. Analysis of Version 1's pre-testing data led to significant changes, including dropping of two scales, the addition of new items, and a revision in the response categories of the scale.

After pre-testing, Version 1 had response categories on a four-point Likert-type scale, with some scales being on a Strongly Disagree-Disagree-Agree-Strongly Agree continuum, one from Not at all Confident to Very Confident with two unanchored categories, and one being frequency-based from Never, Quarterly, Monthly, to Weekly. It was collected using an online survey from early childhood teachers (N=164). Analyses included exploratory factor analysis and the application of the generalized partial credit model (GPCM) which is further described in Section 3. Results of this first iteration led to a reorganization of the constructs and the addition of a new construct. A primary problem that the developers wanted to address was the low item locations for the highest categories – indicating that people with low levels of the construct (e.g., low experience of social support, low frequency of data use) were likely to respond in the high categories. Because the higher categories were expected to capture those with higher levels of the construct, the developers wanted to calibrate items to be more difficult or harder to endorse by those at average levels of the construct. Two strategies were used to increase the location estimates – increasing the number of categories from which respondents could choose and re-wording or creating new items that were expected to reach to higher levels of the construct.

Pilot 2. Version 2 was also pretested in cognitive interviews and behavioral coding prior to data collection. Within the data collection of Version 2, the qualitative data collected in Version 1 was categorized into themes and provided as a choice to teachers. The types of data teachers chose generated skip logic patterns. Respondents were only asked to complete items about data they used (either data about children, classroom quality, or families). Version 2 also forced respondents to explain their reason if they chose "Not applicable" rather than a frequency category. The qualitative and quantitative data from Version 2 data collection was used to generate Version 3.

Analyses of Version 2 indicated that item locations were still quite low, with few items reaching past a theta of 0. It was at this point that items were evaluated and at times re-written to better fit a data use construct based on Marsh & Farrell's (2015) theoretical framework. In the development of Version 2, the framework for three levels of data use was added to generate items that would be calibrated toward higher levels of the construct and to clarify a spectrum of data use based on a theoretical framework.

Version 2 was administered to 313 early childhood teachers from Early Head Start/Head Start (serving children birth to age 5 from low-income families) and community childcare programs (serving children birth to age 8 with varying incomes). The sample included both Bachelor-degreed lead teachers and assistant teachers with Associates degrees or Child Development Certificates. Most items were answered, but some respondents were removed due to missing data on a whole scale or evidence of inattentiveness (all answers the same across multiple scales). The final sample was 295 teachers.

The GPCM was again applied to data from Version 2, which was collected from teachers – both lead teachers, assistants, and aides (N=225). Refer to Appendix 3 for Version 2 items full scale. Results indicated that changes to the Likert scale from a 4- to 6-point scale did not necessarily increase the category locations, but the estimated probabilities of responses did provide information about the response categories that are likely to give the best information across frequency locations. For example, if two adjacent categories had little distance between the item-location, these could be combined. Similarly, if an item had low discrimination estimates, there might be a very low probability of participants responding in some categories, -so the categories may need to be broken up into more granular categories or the item may need to be reworded. Response categories for the DUCES were drawn from common standard measure categories (Sue & Ritter, 2013). These included frequencies of 1=Not applicable/Not in the last 6 months, 2=Once in the last 6 months, 3=Every 2 to 3 months, 4=Once a month, 5=Two to three times a month, 6=Once a week, 7=Daily.

Changes made from the analysis of Pilot 2 were often based on the examination of discrimination and location – viewed through numerical estimates and graphical descriptions of each item. For Informational Use items, many items were eliminated due to low discrimination. The reduced set was examined within an Exploratory Factor Analysis – using principal axis factoring with Varimax rotation. Informational Use items within child, classroom, and family decision content area performed well – with each scale containing only one factor and factor loadings between .62 - .89.

Data Use Contexts and Experiences Scales (DUCES)

The piloting of items related to data use has led to a shorter and more refined version of the measure. The measure fully incorporates the theoretical framework developed by Marsh & Farrell (2015) by structuring items to the context – intrapersonal, interpersonal, and organizational-structuralenvironmental – as well as the process of data use – learning, connecting knowledge, and responding based on data. Table 1.2 shows the constructs captured in the full set of measures.

The first six scales focused on Informational Use are subject of this dissertation study.

Table 1.2.

Definition and Example Items of the Data Use Contexts and Experiences Scales (DUCES)

	Scale	Definition	Example items
Informational Use			
1	TC Class	Reported frequency of behaviors applying data to inform decisions in each content area.	• I used data to evaluate progress toward student learning goals.
2	OC Class		
3	TC Child		• I identified areas where I needed to strengthen my instructional practice.
4	OC Child		
5	TC Fam		• I looked at data as a way to learn about the family.
6	OC Fam		
Motivational Experiences			
7	Motivation	Extent to which data users' experiences were intrinsically or extrinsically motivating.	• I wished I could do things differently in my classroom, but I feel pressure based on data.
Structural Context			
8	Tools and Resources	Rating of the effectiveness of tools & resources facilitating data use	• Leaders who facilitated understanding of data.
9	Time & Training	Rating of the effectiveness of time & training facilitating data use	• Necessary training on interpreting data appropriately.
Intrapersonal Context			

10	Attitudes	Level of agreement that data use is beneficial to teachers.	• Data can help me identify my own strengths and weaknesses.
11	Confidence	Reported feelings of confidence in ability to apply data to work with children and families.	• Apply your early childhood expertise to create plans with data.
Interpersonal Context			
12	Social Supports	Level of agreement that work colleagues are supportive of data use.	I have people at work who would help me if I did not understand data.

Note. TC = teacher-collected, context of data is data that is teacher-collected; OC = other-collected, context of data is data that is collected by someone other than teachers.

Purpose Statement

The purpose of this dissertation study was to refine and test the psychometric qualities of the first six scales of the Data Use Contexts and Experiences Scales (DUCES) – the Informational Use scales focused on practices of teachers in the last six to twelve months for using teacher-collected and other-collected data. This dissertation study included decisions based on the pretesting processes, as well as analyses of the survey measure. The study covered content validity, construct validity and discriminant validity, as well as unidimensionality of the scale, and discrimination and location of the items. To meet the purpose of this study – examining psychometric qualities - the hypotheses below will be tested using the related plan of analysis. The study will review the psychometric qualities of six scales using factor analysis and item response theory (IRT) applications.

Hypotheses

Table 1.3.

Hypotheses and plan of analysis

		Hypotheses	Related analyses
Phase 1: Pretesting	H1:	Scale items are interpreted as intended.	Cognitive interviews indicate that questions can be easily interpreted and answered by respondents with experience in early childhood classrooms (Ruel, Wagoner, & Gillespie, 2016)
	H2:	Scale items are congruent with their intended level of calibration.	Technical expert panelists rate items as fitting within the intended level/definition using the Index of Item Objective Congruence formula (Turner & Carlson, 2003).
Phase 2: Analysis of Survey Data	H3:	Scales are normally distributed	Skewness is calculated between +/-1.0. Excess kurtosis is calculated at .75. (Ho & Yu, 2014)
	H4:	Scales are unidimensional and measure unique constructs	Principle axis factor analysis show only one factor with an eigenvalue greater than 1, or parallel analysis confirms a single factor structure (Costello & Osborne, 2005). Scales are correlated with each other

			at no greater than .50 using Pearson's correlation coefficient (Ruel, Wagoner, & Gillespie, 2016).
	H5:	Items are discriminating	Discrimination of items is greater than .80 within an Item Response Theory generalized partial credit model (De Ayala, 2009).
	H6:	Items measure across the construct levels	Item location parameters estimate item thresholds' (δ_{jh}) differences of $\lvert 2 \rvert$between the first and last threshold with no less than a .25 gap between adjacent thresholds (δ_{jh}). (De Ayala, 2009)
Phase 3: Contrasting groups analysis	H7:	Scales discriminate between high- and low-level data users.	Ratings from key insiders will be used to place survey respondents into high and low use groups (Ruel, Wagoner, & Gillespie, 2016). Independent t-tests will show significant differences on DUCES scores between high and low data use groups

CHAPTER II

REVIEW OF LITERATURE

In the previous chapter, a case was briefly made for the need of a measure of data use for early childhood settings, as well as an overview of plans to fill this gap by continuing the refinement of a measure of data use practices. This chapter further explains the need for a measure by reviewing current literature and introduces the need for each of the six scales refined in this study to flexibly measure data use based on multiple types of data used by early childhood educators and differing contexts of data collection.

Theories and Frameworks of Data Use

The work of Marsh & Farrell (2015) outlines a framework that clearly aligns to many other data use scholars conceptualizations of the complex process of using data (e.g. Mandinach, 2012; Krugly, Stein, & Centeno, 2014). Their theory of action posits that the first step from having data alone to using data is in the form of generating information. This entails understanding data and interpreting it accurately. Next, the information gained from the data is joined with knowledge that already exists within the individual or group who are using data.

The existing knowledge teachers have from their previous training/education and experience with children provides a basis from which they interpret the data. Teachers and other data users usually have knowledge about the interests, strengths, or struggles of the children in their classroom. Teachers and principals bring information about the curriculum or content of teaching targets to the conversation about data. In total, the extant knowledge from a group of data users includes their previous training and education, their understanding of current needs of children and families, and the expectations of their schools' curriculum or expectations of developmentally appropriate practice. Their framework for data use capacity-building is also a framework for measuring the process.

This process of moving from data to information, to knowledge, and then response characterizes effective data utilization and serves as an excellent the basis for the measurement of data use. As shown in Table 1.1, the Marsh & Farrell (2015) framework gives an understanding of data use that defines lower and higher aspects of data use. The lower and higher aspects of the process align well to the goal of measuring the process, as one of the goals of measurement is to understand the nature of a construct by using rules to capture its variation (De Ayala, 2009). As applied to measurement, Marsh and Farrell's framework may help us distinguish the lowest levels of data use from that which is more complex, as defined by the types of activities occurring in the lower versus higher levels of data use.

While Marsh and Farrell (2015) were not the first to explain a process of data use, they uniquely rooted the process in sociocultural learning theory. This is an important distinction, particularly when considering data use in schools, as the impact of data use on children is often not from the teacher alone, but from multiple agents working in concert Other educational researchers have suggested that the sociocultural learning theory should be considered the modern approach to data utilization as it moves away from the incoherence of accountability practices that narrowed teaching practices without improving child outcomes (Lee & Reeves, 2017; Shepard & Penuel, 2018). Sociocultural applications emphasize that learning occurs through meaningful relationships and about things that are of value to the culture. This again brings the issue back to the context of data use, for which the scales related to interpersonal, intrapersonal, and structural aspects of data use were developed. The Informational Use scales were developed to focus on process.

As mentioned, Marsh and Farrell (2015) were not the first to discuss the process of data use. Table 2.1 shows the alignment between Marsh and Farrell's work and the work of several other authors studying data utilization. This shows that, while many different versions of data process and protocols have been suggested, they can all be summarized succinctly by the Marsh & Farrell (2015) process of Information, Knowledge, and Response.

Table 2.1.

Processes of Data Utilization

Marsh & Farrell (2015)	Johnson (2002)	Krugly et al. (2014)	Derrick-Mills et al. (2014)	Mandinach et al. (2012)	Virmani (2013)	Spillane (2012)
Data	Summarize		Create goals Gather data	Collect Organize		
Information	Analyze	Prepare	Analyze	Analyze Summarize	Acknowledge	Routines for Identification of problems
Knowledge	Interpret	Interpret	Synthesize	Synthesize	Ask	Diagnosis

Response	Plan Act Follow Up	Implement Evaluate	Prioritize Implement Monitor Evaluate	Prioritize	Adapt	Prognosis

Current Available Measures

While a measure of data use in early childhood is needed, some measures of data use do exist – one in early childhood (Monahan, et al., 2016) and two developed for general use (Dunn, Airola, Lo, & Garrison, 2013; Jimerson, 2015). These measures have their shortcomings, such as their lack of accessibility to general early childhood programs or their lack of specific adaptation to the early childhood context or too little focus on the practice of data use.

Examining data informing teaching (EDIT). The EDIT incorporates the selection and collection of data, the organizing and interpreting of data, and using the data for overall and individualized teaching (Monahan, et al., 2016). It uses checklists, ratings, and rubrics to gather data about how teachers are using data. EDIT raters review documents and assessments, review administration of assessments and instruction. Data collection includes a one-hour interview. This extensive process, while generating comprehensive information on the use of formative assessment, is likely inaccessible to most early childhood programs. While the EDIT will likely be an effective tool for research studies focused on data use, another measure may be needed to provide more accessible information for programs and sufficient evidence for research and evaluation. The measure is very new and no studies have been published using it as a measure.

Data-driven decision-making efficacy and anxiety (3D-MEA). The 3D-MEA measures teachers' sense of efficacy and anxiety related to data. Many of the items are related to teachers' feelings of confidence and intimidation in terms of data use. The items of the 3D-MEA loaded on five factors – identification, technology, interpretation, application, and anxiety. Studies of the 3D-MEA indicate that measurement of teacher confidence in data use, as measured by the 3D-MEA, may fall into two factors – application and interpretation (Dunn, et al., 2013). The 3D-MEA is based in part on a definition of data use skills as being 1) teacher ability to select and collect data and 2) teacher ability to apply data. The 3D-MEA expands this definition to break down other skills such as interpreting, using technology, and overcoming anxiety.

One distinction of this measure is that it does not measure practices. The 3D-MEA does not ask about data use practices but does ask about confidence in applying data to practice. While this may be perceived as an effective proxy, confidence could be a moderator of practice – with more confidence supporting the effectiveness of data use practice. While confidence may be an important construct to measure, particularly in its application to professional development of teachers, it is unclear whether or how it is linked to data use practices.

Survey of data use and professional learning (S-DUPL). Items of the S-DUPL ask teachers about the extent to which they agree with statements related to technology, collaboration, inquiry, interpretation, and application (Jimerson, 2015). The scales include Confidence, Effectiveness, Construal, Beneficence, Anxiety, and Collaboration, along with some non-scale items that provide descriptive information. Each of the scales asks about interpersonal feelings (Confidence, Construal, Beneficence, and Anxiety) or about organizational supports (Effectiveness of professional development, Collaborations). The S-DUPL, like the 3D-MEA, place aspects that are out of the teachers control (e.g., access to technology to view data) on a different subscale than that which is in the teachers' control (e.g., create lesson plans with data). The scales' internal consistency estimates, as measured by Cronbach's alpha, ranged from .64 for the Collaboration scale to .89 for the Effectiveness scale. Most of the test-retest reliabilities for items were .50 or greater.

In addition to separating what is specific to the school and what is specific to the teacher, a comprehensive measure of data use would also capture the interaction of these two – the interpersonal context of data use. None of the current measures of data use assess the interpersonal aspect. Additionally, it is unclear whether an underlying theory was applied to consider lower versus higher levels of data use beyond attitudes and confidence in the S-DUPL and 3D-MEA.

Other Means of Measuring Data Use. An innovative way in which data use micro-processes have been measured is in the access to a computer system. The researcher counted the number of teacher interactions with technology built to support data use – counting the number of clicks related to certain actions (Wayman, Shaw, & Cho, 2017). Others have looked at practice in a broader way, with data used to inform instruction, to give feedback to students, or to support peer- or self-assessment (Kippers, Wolternich, Schildkamp, & Poortman, 2018).While students may benefit from feedback about their data or be able to self-assess their own learning in the upper grades, these types of data use are not appropriate for children who are still learning to read or even talk.

Surveys used in other research had acceptably high internal consistency when asking about use of curriculum-based benchmark assessment data (Christman et al, 2009). Others studied data use more broadly and had low variability on items asking about access to data and changes made with data (Pierce & Chick, 2011). The U.S Department of Education has studied data use through the Office of Planning, Research, and Evaluation (OPRE) and found that data use increased over time – these data were gathered through phone interviews using a questionnaire (Sandstrom, Isaacs, & Rohacek, 2015). None of these efforts focused on the broad types of data used by early childhood teachers, nor incorporated a framework of data use at different levels.

Challenges of Measuring Data Use: Content & Type

Classroom Content. Beyond data on student learning, two other topics are considered extremely important in high quality early childhood education and care – information about classroom quality and information about the family. Classroom quality data, while also used by teachers of older children, has significant importance in early childhood due to studies indicating that high quality early childhood classrooms influence the development of children through their lifespan, with better outcomes still evident at 40 years of age (Campbell et al., 2008). These studies were specifically on children from low-income homes. More recently, studies are finding similar results with analysis of children from a diversity of backgrounds, with recent studies of the children in their adolescence (Love et al., 2003; Vandell, Burchinal, Pierce, Eccles, & Dubow 2016). Because the quality of a classroom is empirically tied to the ongoing learning and development of young children, early care and education teachers are expected to produce measurably high quality classrooms. One way to improve the quality of a classroom is to allow measures of classroom quality to inform practice (Zweig et al., 2015). That is, to use data from classroom quality measures to identify sources of strength and areas for improvement, followed by professional development or other supports to adjust classroom practice as needed.

Child Content. While other data use studies have examined the use of data related to measures of student outcomes (Dunn, Airola, Lo, & Garrison, 2013; Reeves, Summers, & Grove, 2016), early childhood professionals rely less heavily on child assessment than teachers of older children. This is, in part, due to the difficulty of assessing young children and the historical dearth of measures available for non-clinical assessment (Stevenson-Garcia, Horm, Atkins-Burnett, 2019). Some early childhood professionals resist data use because they feel that child assessments are not developmentally appropriate (Ford, et al., 2016). However, as more measures have been developed for young children, the use of play-based assessments has been welcomed, as they capitalize on the playful behavior of young children (Walker, Carta, Greenwood, & Buzhardt, 2008). These play-based measures are more comfortable for teachers to collect as frequent formative assessments, which is also a recommended practice for effective data use (Bambrick-Santoyo, 2010; Carta, Greenwood, Baggett, Buzhardt, & Walker, 2012; Stecker, Fuchs, & Fuchs, 2005).

Family Content. A third area of data that can be used by early childhood educators is data on families. For example, data on types of adversity children may have experienced through their family life may help teachers or other specialists decide on the types of supports that a child needs – supporting a child who has experienced trauma can occur through play that allows them to express their fears and process confusing events. A more common example may be information about what a child's typical schedule is like at home and whether a child is regularly read to outside the classroom setting. This kind of information may help teachers adjust to a child's needs – allowing a child to take an earlier nap if needed or giving a child more one-on-one time with books to allow him to explore the process of text and pictures individually. A value of most early childhood programs is partnership with families, under the standard that parents are children's first and most important teacher (National Center on Parent, Family, and Community Engagement; NCPFCE, 2011). Organizations that offer accreditation also have requirements for partnership and involvement of families. This is due to the unique relationships teachers often have with parents when children are young. Daily communication related to diapering, feeding, and naps is highly important between caregivers at home and in early childhood settings. Thus, early childhood professionals are, at times, in a position to support and encourage parents to work toward goals. Large programs, such as Head Start, require that families are involved in setting goals for their children (National Center on Program Management and Fiscal Operations; NCPMFO, 2013). As mother's education and family income are some of the strongest predictors of student achievement (e.g., Ayoub, O'Connor, Rappolt-Schlictmann, Vallotton, Raikes, & Chazan-Cohen, 2009), Head Start staff often work with families to set goals related to employment or education. Thereby, data on families because part of the information used to make individual and program decisions.

Two Types of Data: Who Collects

To further specify the information that could be considered when using 'data', the manner of collection can be important. In a qualitative study on the use of data in Louisiana, researchers found that teachers were more likely to use data they had collected themselves rather than data that was collected via a standardized computer format (Ford et al., 2016). Additionally, the self-collection of data may increase the accessibility of data. Some studies have noted that child outcomes improved when teachers had technology that facilitated access to data (as cited by Akers, Grosso, Atkins-Burnett, Boller, Carta, and Wasik [2014]). This generates two types of data – teacher-collected and other-collected. Altogether, data can be defined both by the content of data (classroom, children, or families) and by the type of data (teacher-collected or other-collected).

Research on Data Use Practices

While few measures of data use exist, research on data has been done through observation of data use or by assessing change in child outcomes following a data use intervention. Aligning curriculum-based assessment has produced gains in student outcomes (Stecker, Fuchs & Fuchs). Targeting children who need additional assistance through data and providing additional learning supports has led to larger gains for preschoolers (Buysee, Peisner-Feinberg, Soukakou, Fettig, Schaaf, & Burchinal, 2016). A data use intervention showed positive effects, particularly for children at risk (Van Geel, 2016). The use of a classroom observation measure of literacy practices was used to guide instruction, resulting in an increase in children's literacy skills (Beecher, Abbott, Petersen, & Greenwood, 2017). Research supports that data use is important for educational outcomes including teacher practices and student learning.

When data use was measured based on the number of clicks teachers made on a website that was built to support teachers' data use, a relationship between practice and outcomes was observed. The teachers used the website to provide students with resources, to pull reports, and track student work. This was associated with higher gains on assessments of elementary school reading (Wayman, Shaw, & Cho, 2017). However, the amount of interaction with the website did not have a relationship to elementary math, junior high reading, or junior high math. Some studies have shown that data use practices of teachers are related to student outcomes. Thus, it seems important to focus on this aspect of data utilization – what teachers are doing in response to data – whether learning, connecting their learning to other information, or responding to better meet the needs of children in their classroom.

Research Questions

Current measures are limited in several ways, including accessibility and affordability, comprehensive assessment of data use practices, and types of items that appropriately capture the data use experience of early childhood teachers. I expect that some of these limitations can be overcome through further refinement and testing of the Informational Use scales of the DUCES.

The DUCES is based on a theoretical framework that can be applied to other research, practice, and policy. The currently available measure developed for early childhood is focused primarily on child assessments and collection may be burdensome for some programs. Measures that are less burdensome were not developed for teachers of younger children and do not focus on practices - an important driver of teaching and learning. Thus, refining and testing a measure of data utilization that is appropriate for early childhood programs will fill a gap in current research. Research questions for this study include:

RQ1: Are scale items interpreted as intended?

RQ2: Are scale items congruent with their intended level of calibration?

RQ3: Are scale scores normally distributed?

RQ4: Are items discriminating at an acceptable level?

RQ5: Are items measuring across the construct's latent continuum?

RQ6: Do scales discriminate between higher- and lower-level data users?

Considering that multiple education scholars have outlined similar processes of data use that align with the Marsh & Farrell framework, their framework was chosen as a model to organize analysis and interpretation. This chapter also outlined existing measures and argued for the need of a measure that is more accessible and more comprehensive than current options. Finally, the comprehensiveness of a measure was outlined by arguing for the need to measure within two context of data collection – data collected by teacher or others – and three types of data, including child assessment, classroom quality, and family data.

CHAPTER III

METHODOLOGY

The previous chapter reviewed literature about data use in early childhood and made a case for a theoretical framework to organize a measure. The literature also supported the need for a measure to cover contexts and types of data that are important in early childhood settings.

The scales being developed for this study stem from a previous iteration of the tools, described in Chapter 1. Decisions about the refinement of this measure were made in accordance with Standards for Educational and Psychological Measurement (AERA, APA, & NCME, 2014) by subjecting the items to cognitive interviewing and item paneling for the purposes of pretesting and refining the items. Decisions about a recommended set of items were made in January through March of 2018 in the pre-dissertation phase of the project. This dissertation study began in April 2018 with pretesting. Following refinement based on pretesting, the refined items were tested with 359 teachers in February -- August of 2019. The survey data were analyzed using factor analysis and item response theory applications to assess item quality in September of 2019.

A subset of teachers, N=61, received ratings of their data use abilities by key insiders in February through March of 2019. This data was gathered as part of the validation process for the measure.

Procedures

This research was funded by the Buffett Early Childhood Fund. The IRB of record was with the University of Oklahoma, with Oklahoma State University reviewing the entire application as part of the agreement to allow for this. The study was reviewed and defined as Exempt. Data from survey respondents was kept on a secure online website (Qualtrics) or on secure servers at the university. Data from other respondents was not identifiable. Survey respondents were compensated with a $20 gift card for completing the survey. Raters of teachers were compensated with $10 per rating.

Four phases of data collection occurred – pre-testing, technical expert paneling, teacher survey completion, and insider perspective ratings of the teachers' data use abilities. All participants were informed that their participation was voluntary and that the purpose of the study was to learn about the items and not to learn about participants themselves. Survey participants signed a consent online, with identifiable data linked to their answers, including race, gender, age, school, education, and experience. Cognitive interviewees and behavioral coding participants were volunteers about whom no identifiable information was collected. Technical expert panelists were compensated or provided their time in-kind.

Pre-Testing

Cognitive Interview Sample. The cognitive interviewees were recruited from a group of current and former early childhood teachers within the social network of the researcher. Two former teachers and one current teacher were recruited to complete the cognitive interviews.

Cognitive Interviewing and Analysis. Cognitive interviews are a means of gathering the thought process of respondents while they respond to test items to ensure that the intended cognitive process is occurring (AERA, APA, & NCME, 2014). The cognitive interviewees were recruited from a group of current and former early childhood teachers within the social network of the researcher. Two former teachers and one current teacher were recruited to complete the cognitive interviews. The researcher explained the protocol – that the participants should read the question out loud, then continuing talking about how they were interpreting it until they could answer it. They were asked to give a 'stream of consciousness' or to continue talking about their thoughts while they decided on their answer. Once they answered, they could go on. The researcher took notes while the participants completed this process.

This process involves a single participant responding to each item while talking about their interpretation of the item, considerations they are having as they generate an answer to the item, and their rational for a final answer (Ruel, Wagoner, & Gillespie, 2016). A section of the set of measures was used, as it would be too time consuming to conduct interviews on the entire measure. Issues discovered through this process were documented and considered in the further refinement of the items. See Appendix 2 for a description of protocol.

Behavioral Coding Sample. While the original proposal did not call for behavioral coding, the small number of cognitive interviewees that were recruited motivated the addition of behavioral coding to be conducted in order to test whether items were being interpreted as intended. An early childhood education night class was recruited to participate in the study. All students (N=16) worked during the day in an early childhood setting.

Behavioral Coding and Analysis. Participants were asked to complete the survey and mark any question that they struggled to understand with a star so that the researcher could work to improve the understandability of the item. Three of the sixteen returned surveys were marked with a star. Those items were discussed. The participants in the night class were watched for signs of difficulty while they completed the test. These signs were also brought up and discussed. Identified issues were used for refinement of items.

Technical Expert Panel

Technical Expert Panel Sample. The technical expert panel was recruited through research contacts and outreach via email. The panelists were recruited based on their expertise in three areas important to the overall set of scales – data use in schools, self-determination theory applications to data use, and early childhood education. The panel included two experts in each of these areas. Data use experts had published work or conducted work related to the use of data for decision-making in school settings. Early childhood education experts had degrees and publications related to the learning and development of young children in group settings. Self-determination experts had publications in which self-determination theory was applied to data utilization.

Technical Expert Paneling and Analysis. All six panelists were asked to rate the set of items that had Child Content (applicable to TC Child and OC Child). Due to time concerns, the rest of the measure was broken up between the group so that they were not asked to rate the entire measure. The two experts in early childhood were also asked to rate items related to Classroom Content and Family Content (applicable to TC Class, OC Class, TC Fam, OC Fam). Thus, the number of ratings for the Classroom and Family scales is smaller than that for the Child scales. Because many items were refined to better fit with the Marsh & Farrell (2015) framework, the refined items needed to be examined to evaluate whether others could evaluate and categorize items to the intended location on the construct table (Ruel, Wagoner, & Gillespie, 2016). This allowed for an evaluation of items separate from the developer's view. Panel members were oriented to the construct definitions through a presentation by the developer and through online descriptions.

The method for item review was an online survey presented so that panelists could rate each item as clearly fitting in, somewhat fitting in, or not fitting in three categories – learning, connecting knowledge, or responding based on data. Thus, for each item, respondents had nine choices and were asked to choose three per item. See below for an example item and Appendix 4 for the entire measure.

Q1 I used data to evaluate progress toward student learning goals.

	Does not fit this description (-1)	Somewhat fits this description (0)	Clearly fits this category (1)
Data is used to understand children, without clear connection to plans. (1)	○	○	○
Data connects to existing knowledge or to consider children's status within current plans. (2)	○	○	○
Data is used to respond, plan responses, or evaluate previous plans/actions. (3)	○	○	○

Figure 3.1. Example item from Technical Expert Panel rating instrument.

As shown in Figure 3.1, the item "I used data to evaluate progress toward student learning goals" could be evaluated as fitting, not fitting, or somewhat fitting in each of the three definitions. This allowed for the respondents to give rich data about their view of the items – both where they do and do not fit and where their fit is unclear.

While this did not show on the survey that panelists viewed, you can see on the example item that a rating of 'does not fit' receives a -1, somewhat fits = 0, and clearly fits = 1. These values were used in the calculation of the Index of Item-Objective Congruence (Crocker & Algina, 1986). The equation is shown in the equation

$$I_{ik} = \frac{N}{2N-2}(\mu_x - \mu), \tag{3.1}$$

where *I* is the Index of Item Objective Congruence for *i* items having *k* content areas. The mean μ_x is the judges' rating on the item within its intended category, while μ is the judges' ratings on the item across all categories. This measure allows multiple judges' ratings on the intended item content to be compared to the judges' ratings on across all the item's content areas. This distance, accounting for the number of objectives, gives a number between 0 and 1. An IIOC of .75 is considered acceptable as it can be interpreted as agreement among three-fourths of the judges (Turner & Carlson, 2003).

Due to feedback from panelists, this data was not used to refine items. However, the data is helpful in assessing fit with the theoretical construct. A comparison between where panelist calibrated items and where location estimates calibrated items will be an important exercise in considering the measures' fit with the theoretical framework of an underlying data use construct.

DUCES Survey

Sampling and Recruitment. The early childhood teachers who completed the survey tool were recruited in several ways – by providing their contact information while completing other surveys for the Educare Learning Network, through Educare emails, through contacts generated by participation in similar early childhood evaluations, through newsletters disseminated by early childhood evaluators and grant funders, and through social networking sites (Facebook groups) that were aimed at early childhood teachers. While this did not enable the recruitment of thousands of respondents, as is ideal for item response theory applications, it was hoped that at least 300 would be recruited. A suggested sample size for a general partial credit model is five people for each parameter estimated (De Ayala, 2000). With the general partial credit model, each item has its own discrimination and a location parameter is estimated for k-1 categories. Thus, for my largest scale (OC Child), that had nine items and seven parameters per item, I would have generated 63 parameters. The recommended sample size would have been 315. However, the categories were collapsed to five, generating 45 parameters and necessitating a sample size of 225. The final sample size for OC child was 286. The remaining scales also had a sufficient sample size using a five-cases-per-parameter rule.

Sample Description. Demographics of 338 teachers who participated in the survey were diverse in age and race, but not in gender. The sample of respondent reports are described in Tables 3.2 and 3.3.

Race and Ethnicity. Teachers could report on multiple races/ethnicities. Those who marked more than one category are grouped into the multi-racial category. Almost half (45 %) of teachers were White/Caucasian/European American, 21.7% were Black/African American, and 9.5% were Hispanic/Latinx. As shown in table 3.2, the remaining teachers were of another race/ethnic, were multiracial, or did not have data reported. Race/ethnicity data was not collected from some participants (n=33) and thus, their data are missing.

Table 3.2.

Race/Ethnicity of Teachers (N=359)

	N	%
European American	162	45.1%
African American	78	21.7%
Native American	12	3.3%
Asian	7	1.9%
Pacific Islander	0	0.0%
Hispanic	34	9.5%
Other	6	1.7%
Multi-racial	27	7.5%
Missing	33	9.2%
Total	359	100.0%

Teachers also reported age and gender. Some gender data was not reported (9.5%). Of those that were reported, the majority of teachers were female, with only 1.7% of teachers being male. This is representative of the measures' target population, as 98% of Head Start teachers were reported to be female in a 2016 study and a recent report focused on the ece workforce described the population as "predominantly female" (Institute of Medicine & National Research Council, 2015; Whitebook, McLean, Austin, & Edwards, 2018). There was more diversity in age, with about a third (29.6%) of teachers being between 25 and 34, a quarter (25.9%) between 35 and 44, and almost another third (29.2%) between 45 and 64. The exact percentages are shown in Table 3.3.

Table 3.3.

Gender and Age Group of Teachers (N=359)

Gender		n	%
	Female	319	88.9
	Male	6	1.7
	Not reported	34	9.5
	Total	359	100
Age Group		n	%
	18-24	20	5.6
	25-34	106	29.5
	35-44	93	25.9
	45-54	68	18.9
	55-64	37	10.3

65+	3	.8
Missing	32	8.9
Total	359	100

Survey Data Collection and Analysis. Teachers completed scales based on the types of data they use, so that the content of items relevant to them were shown. TC Classroom Content was shown to those who said they had access to data collected by other teachers about classroom quality and teacher-child interactions. OC Classroom Content was shown to respondents who said they had access to data collected by evaluators, researchers, or quality rating and improvement systems (QRIS) data collectors about classroom quality or teacher-child interactions. More teachers had access to data collected by others than collected by teachers. The most common type of data to which teachers had access was data about children that teachers collected themselves – only one teacher in the sample did not have access to this kind of data. The second most common type was data about children that others had collected. Teachers collected formative assessments, natural observations, and screenings, while others collected direct child assessments and standardized measures.

Table 3.4.

Frequency and Percent of Total Sample Completing Each Scale by types of data accessed (N=359)

Teacher Collected Context	N	%	Other Collected Context	N	%
Classroom Content	254	70.8	Classroom Content	268	74.7

(TC Class scale)			(OC Class scale)		
Child Content (TC Child scale)	329	91.6	Child Content (OC Child scale)	286	79.7
Family Content (TC Family scale)	227	63.2	Family Content (OC Family scale)	223	62.1

The type of data that the fewest teachers had access to was data about families. This included parent interest surveys collected by teachers or formal family needs assessments collected by others.

Normality. Normality is an important assumption of parametric statistics. Thus, the normality of the data is an important point to check as we begin analysis of a measure. The data were checked in three ways – skewness, kurtosis, and the Shapiro-Wilk test. Skewness was expected to be between -.1.0 and 1.0, while excess kurtosis was expected to be between greater than .75 (Ho & Yu, 2014). The Shapiro-Wilk test has a null hypothesis that the data in the distribution are identical to a normal distribution. Thus, a statistically significant Shapiro-Wilk test would indicate that the data are not normal.

Unidimensionality. Because changes to the measure were significant through pretesting, exploratory factor analysis was conducted rather than confirmatory factor analysis. Similar to the previous version, principal axis factor analysis with varimax rotation was used, as suggested by Costello and Osborne (2015). Scales were considered unidimensional if all items load on a single factor with an eigenvalue greater than 1, also known as the Kaiser criterion. This method is not the most accurate method; it is critiqued because it tends to generate too many factors (Costello & Osborne, 2015). This effectively makes it a more conservative approach to testing unidimensionality with a single factor considered evidence for unidimensionality.

Item response theory. Item response theory (IRT) is a modern approach to measure development (De Ayala, 2009; Devellis, 2016). IRT applications provide a sample-independent means of examining the information contributed by an individual item to a scale. The information contributed by an item can be estimated based on the item's estimated discrimination value and each category's location estimate, further described below. These methods can be applied to both dichotomous and categorical, or polytomous, ordered data. This study applied IRT methods to polytomous items to which there is no correct answer.

The items of the measure examined for this study are on an ordered scale ranging from Never, Once in the last 6 months, Every 2 to 3 months, Once a month, Two to three times a month, Once a week, and Daily. Potential methods for handling this type of data include the graded response model (GRM) and the generalized partial response model (GPCM). The GPCM was chosen because it is an adjacent model, rather than a cumulative model. That is, the GPCM estimates the location on the latent continuum at which there is a .5 probability of crossing from one adjacent category to the next, but does not assume categories are in the 'correct' order by person location. In contrast, the GRM uses an additive model that does assume correct order of response categories.

While the frequency response scale from Never to Daily does have a natural ordering in terms of understood timing, the 'ordering' referred to and tested by the GPCM is that those with a lower level of the construct, i.e., low-level data users, will respond in lower categories, such as never and once-a-month, and those with a higher level of the construct will respond in higher categories, such as Weekly or Daily. The equation for the GPCM is

$$P\left(X_{jk}\middle|\theta, \alpha_j, \delta_{jk}\right) = \frac{exp\left[\sum_{h=1}^{k_j} \alpha_j(\theta - \delta_{jh})\right]}{\sum_{c=1}^{m_j} exp\left[\sum_{h=1}^{c} \alpha_j(\theta - \delta_{jh})\right]}, \tag{3.2}$$

which shows the probability of score X for item j in category k, given person location (θ), discrimination (α) of item j, and item j location (δ) of each category k (De Ayala, 2009). The equation is essentially an odds ratio where probability is determined by the occurrence of an event over all possible occurrences. In the generalized partial credit model, the event is the logit of the sum of person location differences from item locations for each item j's threshold h, multiplied by the discrimination estimate (De Ayala, 2009).

The GPCM model was applied to the dataset of item responses from the set of examinees' response patterns on the set of items for each of the six scales. Following the evaluation of the unidimensional assumption of IRT, each of the 6 scales had item parameters estimated using marginal maximum likelihood estimation.

Model fit was tested using the C2 adaptation of Maydeu-Olivares M2* by Cai and Monroe (2014). The developers of the C2 goodness of fit statistic demonstrate that C2 is more powerful than M2 and M2*, particularly when there are a small number of items (Cai & Monroe, 2014). The statistic was tested with unidimensional and ordinal data. The data in this study fit both of these characteristics – small number of items per scale and unidimensional, ordinal data – making C2 an appropriate model fit statistic for this analysis.

These model fit indices were run in R with the mirt package. The model was adapted from R defaults for the generalized partial credit model to replicate the modeling of Stata software. Specifically, Stata software uses a Newton-Raphson algorithm and sets quadrature points to seven (Yang & Zheng, 2018). The code that was used to estimate the model in R is shown below.

```
Mod1 <- mirt(data, 1, itemtype = 'gpcm', method = "EM", optimizer = 'NR', calcNull = TRUE,
quadpts = 7, verbose = TRUE)
```

Item fit was assessed with SX2. The SX2 item fit statistic was shown to be have the lowest Type 1 error rate when compared to G2 and RISE (Liang & Wells, 2009).

Model fit and item fit were also conducted in R. The remaining analyses were conducted in Stata V.14, 2015.

Discrimination. Discrimination is an item's ability to distinguish between respondents who differ from each other on the latent continuum that is being measured. In the GPCM model, each item has a discrimination parameter. The quality of the items of each scale was evaluated by examining that discrimination of items were .80 or greater. This is a recommended point to identify items having unacceptably low discrimination in a generalized partial credit model (De Ayala, 2009). Items with a low discrimination parameter (less than .80) were considered for elimination (De Ayala, 2009).

Location. Location of items is the estimate of where on the latent continuum an item captures information about respondents. In the GPCM, location is estimated at the threshold between category responses (e.g. between Never and Rarely). The quality of the items of each scale was evaluated by examining that estimated locations were ordered and ranged between -1(δ) and 1(δ) units of the spectrum. Items with categories that do not follow the expected order or are redundant in their locations will be considered for elimination or refinement.

Insider Perspective Ratings

As part of the validation process for the data use measure, ratings of teachers by key respondents (teachers' supervisors or local evaluation partners) were used to examine discriminant validity. While discriminant validity and divergent validity are sometimes used interchangeably, discriminant validity is here defined as a measure's capacity to distinguish between two different groups. This differs from divergent validity, which should be defined as a measure's divergence from a dissimilar construct (Raykov & Marcoulides, 2010). Discriminant validity was examined with a mean comparison between two groups that were determined to be higher and lower data users based on the insider perspective ratings. Insider perspective ratings were used to assess discriminant validity. The ratings of key insiders were used to create high and low groups for discriminant validity.

Sampling and Recruitment. The insider perspective ratings were gathered from key individuals who had experience working with teachers during data use processes at the school. The individuals were supervisors of teachers or Local Evaluation Partners (LEPs). The supervisor or LEP was recruited based on being in a position to be familiar with the teachers' use of data. In order to protect the confidentiality of teachers, supervisors and LEPs were not told which teachers were participating in the study. Raters provided a rating to as many teachers as they wanted, with the expectation set that they would only rate teachers with whom they had worked in a data use context. Eleven raters provided 127 ratings. About half of these were matched with a completed survey (N=61). This sample size was smaller than the expected 100 matched ratings.

Data Collection and Analysis. The key respondents (insiders) rated each individual teacher in classrooms they had worked with using a brief 10-item rating, created by the researcher, asking how likely the teacher was to use data in a variety of ways. Example items include "How likely is this teacher to… use data to learn more about a child; use data to identify areas of classroom practice they want to improve." The full rating instrument can be found in Appendix 5.

To protect confidentiality, key respondents were not provided with information about which teachers had participated in the study and had no knowledge of teachers' scores on the DUCES. The rating was scored by summing all 10 items. The scores were used to create high and low groups based on an approximate mean split prior to matching the insider rating to the DUCES survey completed by the rated teacher.

Because those who were rating the teachers were unaware of who was in the study, the ratings and surveys were not matched in one-to-one correspondence. In fact, only 48% of teachers who were rated completed the DUCES. Attrition from staff turnover, promotion, surveys filtering as spam, or just being ignored due to lack of recognition reduced the potential sample. While issues of job changes and email problems could be considered somewhat random, if the email was seen and understood, but participants selected not to participate, this would bias the sample. Thus, a sensitivity check was conducted to assess the appropriateness of the validity data by analyzing mean differences (independent t-test) between the insider ratings of teachers who had completed a DUCES survey (and therefore had a match) and the insider ratings of teachers who had not completed a survey (and therefore had no matched DUCES survey).

CHAPTER IV

FINDINGS

This chapter covers results for each hypothesis, in order of the research questions outlined in the first chapter. A synthesis of the findings and decisions about items and scales are then shown to as a final refinement of the set of measures.

RQ1: Are scale items interpreted as intended?

H1: Scale items are interpreted as intended.

Pre-Testing: Cognitive Interviews

The cognitive interviewing process led to minor revisions in the measures. Most items were interpreted as intended. Some items were misinterpreted by multiple interviewees; these items were dropped. Other items were reworded.

Use of Classroom Data. When answering about use of classroom data collected by others, one interviewee noted that she received feedback from supervisors on a schedule. Thus, her answers were based on systems and lacked variability in regard to frequency.

She also struggled to interpret one item. The second interviewee also answered classroom data based on other-collected data, as that was all that she had experienced during her time as a teacher. Her explanations of why she answered were less detailed than her answers about using child data. She seemed to interpret all items generally, rather than focusing them on how data was used when the items did not begin with "I used data to…" She also mentioned that she would have taken action more frequently if the system had provided feedback more frequently (e.g. she had been given more information about interpreting and responding to the data). There was little variability in her scores. The third teacher did not answer the classroom set of questions because she said that teachers rarely observed other teachers and formal observations only occurred once a year. The scale was ordered from 'Not in the last 6 months' to 'Once a week'.

These results informed decisions to shorten one item and to begin all items with "I used data to…" There is also evidence that systems – programs or schools – may influence this measure more than the scale assessing use of child data. The response category, 'once in the last 12 months' was also added.

Use of Child Data. The first participant was asked to only think of teacher-collected data when answering questions. She struggled with one question, wondering 'if they aren't doing well everyday…' but interpreted it based on whether she had data to group children, as the item intended. As she answered questions about child data, she explained that her answers were primarily based on structured monthly meetings, so many of her answers were 'monthly'. For answers that were 'daily', she noted that she was basing it on informal data collection. Specifically, "daily data in some form is constant, especially with babies." She referenced what some early childhood professionals would consider knowing a child's schedule and reading an infant's cues as pieces of data that she learned from experience or noted on a Daily Sheet and made decisions based on throughout the day. These are high-level skills in an early childhood professional, thus an answer of daily in this context would align with a high ability level, as intended.

The second participant was asked to answer questions about teacher-collected data. All questions were interpreted as expected. An example is the interpretation of "I used data to identify whether student learning needs were met using data" as using data to provide different types of instruction based on abilities. This shows that the teacher is connecting data to other knowledge she has – that children may need differentiated instruction. This item was designed to capture the mid-level of use – Connecting to Existing Knowledge.

The third participant answered based on teacher-collected data. She noted that some time frames for her use were systematic, depending on software systems generating data. She noted that she used data weekly in most areas and had little trouble interpreting the data. Despite her reliance on systems, she was using data at almost the highest level on the scale.

Use of Family Data. The first participant answered about teacher-collected data. She was confused about two items, felt that she could not answer a third due to lack of applicability to her situation, and needed additional information to answer a third. The second respondent also responded about teacher collected data. Three of the questions were easy to interpret – learned about the family, reflected on data, and noticed changes in the child, but the other three were difficult to understand. The third respondent interpreted the first two questions easily but struggled to answer three. The results of this interview led to dropping two items and slightly rewording a fourth.

Pre-Testing: Behavioral Coding

Based on the information gathered from the 16 students in the behavioral coding sample, x items were reworded. The process also confirmed the need to keep the scales separated between

Teacher-collected and Other-collected. The number of response categories was increased from four to seven, partially based on information received during this meeting.

Participants said that they struggled to answer some questions based on consideration of the type of data, saying “I do things differently with my own data (than with data that others collect).” Some questions did not apply to their experience (family progress). The researcher followed each of these comments with further questions. The researcher asked if the questions would have been easier to answer if the survey had started out by saying – ‘Think about data you collect’ and the teacher confirmed that this was true. The teacher who felt that the family questions did not apply to her noted that she was a public school teacher, rather than a Head Start or Early Head Start teacher – whose organizations focus heavily on family engagement. These findings confirmed that the questions needed to be split by teacher-collected and other-collected types of data. The findings also supported the use of logic embedded into the online survey that would only show questions about families to teachers who had access to family data for the purpose of informing practice.

The researcher noted that some finished very quickly, while others took a long time. The researcher also noted who erased their answers, had facial expressions denoting a cognitive struggle (frowns, furrowed eyebrows, turned head), and who had long pauses before going on to the next answer. Follow up with the participants who had signs of struggling to complete the assessment were asked about their interpretation of questions. These participants asked about specific phrases, such as ‘inform practice’ and one explained that she does not use much data. The participants who were pausing for a long time said that they felt the question did not apply to them (home childcare). They also felt that there should be more response categories to provide for more or less frequent use of data. This was part of the reason the scale was moved from a four-point scale to a seven-point scale. Overall, most items were interpreted as intended. Items that were not interpreted as intended were revised.

RQ2: Are scale items congruent with their intended level of calibration?

H2: Scale items are congruent with their intended level of calibration.

Technical Expert Panel Analyses

The Index of item-objective congruence (IIOC) calculations for items was not at a high level for most items. Turner and Carlson (2002) suggest an IIOC of .75 is acceptable, as that would indicate that 3/4ths of the judges agreed that the item was congruent with its intended content. Average scores for each item's content area are shown in Table 4.1 under Objectives. As an example, the first TC Classroom item, "I identified areas where I needed to strengthen my instructional practice", was intended to be about learning, apart from connect to other knowledge or responding – hence the rating under the objective Learning is highlighted in grey.

Application of the Turner and Carlson rule was not possible, however, due to the lack of judges to rate some of the scales. Specifically, only the items for TC Child and OC Child were reviewed by more than two judges. Thus, the Classroom and Family items only have four scores - both said that the item clearly fell in one objective, giving it a rating of (1), with their scores averaged to equal (1); both said that the item clearly did not fall in one objective/level, giving it a rating of (0), with scores averaged to (0); the two disagreed on whether the item fell in the objective, with one saying it clearly did/did not (1/-1) and one saying it did somewhat (0), for an average rating of (.50/-.50). The IIOC is calculated accounting for the number of objectives.

While three of the five Classroom Data items were rated at 1.0 (both raters agreed), only one of the Child Data items were rated above the suggested criteria of .75. One of the four Family Data items had IIOC values at 1.0. Refer to Table 4.1 for more information. Most items on the Child and Family scales were not congruent with intended levels of calibration based on IIOC values.

Most of the items on the Classroom scale did meet the expected IIOC value, indicating that these items were congruent with their intended level.

Table 4.1

Index of Item Objective Congruence Estimates and Judges' Average Ratings by Objective

Item	IIOC	Objectives 1) Learning	2) Knowledge	3) Response
Classroom Data (TC or OC) N=2				
I identified areas where I needed to strengthen my instructional practice.	-0.62	-1.00	1.00	-0.50
I used data to make plans (e.g. professional goals, work objectives).	0.75*	-1.00	1.00	0.00
I made changes to the types of activities I was doing with students.	0.75*	-1.00	0.00	1.00
I made changes to the strategies I used to support learning and development.	0.75*	-1.00	0.00	1.00
I used data to evaluate the effectiveness of my plans for improvement.	0.00	-1.00	1.00	0.00

Child Data (TC or OC) N=6

I used data to evaluate progress toward student learning goals.	0.25	-0.50	0.17	0.33
I used data to evaluate the effectiveness of my instruction. (e.g. lessons and/or projects)	0.71	-0.67	-0.50	0.83
I used data to identify a group of children who weren't doing well.	0.13	0.17	0.17	-0.33
I identified students' strengths and areas for improvement.	0.25	0.33	0.17	-.050
I identified students who needed more learning challenges.	-0.13	0.33	-0.17	-0.17
I identified whether student learning needs were met using data.	0.08	-0.17	0.17	0.17
I used data to determine children's knowledge or skills before teaching.	-.17	0.67	-0.17	-0.33
I formed small groups based on data about children's learning and development.	0.75*	-1.00	0.00	1.00
I identified reasons a child struggled to learn.	0.62	-0.50	1.00	0.00

Family Data (TC or OC) N=2

I looked at data as a way to learn about the family.	1.00*	1.00	-1.00	-1.00
I reflected on data with others (e.g. co-workers, supervisors) to think about the strengths of families.	0.25	0.00	0.00	-1.00
I used data to plan family engagement activities with other staff	0.00	-1.00	1.00	0.00
I used data to assess progress in the well-being of families.	0.00	0.00	0.00	0.00
I noticed that family data explained changes (positive or negative) in a child.	0.00	0.00	0.00	0.00

Note: Grey coloring indicates the valid objective defined to be measured by the item. *Marks IIOC that meet the criteria of .75.

In addition to the IIOC, Table 4.1 reports the average score across judges for each level of data use. The intended objective is highlighted in grey. This shows how judges calibrated the items across objectives. Judges thought that the last item on the Classroom Data Scale fit best under "knowledge" – or connecting to knowledge. While the act of evaluating plans with data was placed in a higher level in development than learning to make plans, it is easily understood how this item could be calibrated at the 'Knowledge' level.

While most items on the Child scale did not reach the .75 criterion, the average ratings on the intended objective give some indication of how well the item was calibrated. Items 2, 8, and 9 had high ratings, with most judges rating them as clearly fitting in their intended category. However, they did not all reach the .75 criterion because the items were also rated under other objectives. Items 1, 3, 4, and 6 also had positive values, indicating some agreement with the intended objective. Items 5 and 7 were both rated in a lower objective than intended. Most of the Family items, with the exception of item 1, seemed unclear to judges, with average ratings of zero across many categories.

The information from this process was not clear enough to be used for changing the items. One panelist mentioned that for the design of the content, objectives logically included multiple levels. He noted that he had marked multiple categories and why. Specifically, if data use occurred at a high level, the understanding was that it was also occurring at the lower levels. Judges were given the ability to rank the data in this way for precisely this reason – I wanted to check whether my assumptions about category fit matched others' assumptions. This feedback gave clear initial indication that my assumptions were not validated – judges saw many items fitting in multiple categories, rather than just one. While this data was not used to refine the measure, the IIOC and the average ratings indicated that some items were calibrated as intended and others were not.

RQ3: Are scale scores normally distributed?

H3: Scales are normally distributed.

Evaluation of the DUCES Distribution and Structure

Normality. Scales were scored by summing the items. The means, standard deviations, Shapiro-Wilk test statistic, skewness, and kurtosis were examined. Many scales lacked normality. The frequencies were examined to learn what categories might be collapsed in order to increase the normality of the items and, thereby, the scale. Very few teachers responded on the lowest ends of the scale, such as “Never” or “Once or twice a year.” For this reason, many scales were collapsed at the low end to improve the normality of the scale.

The means, standard deviations, Shapiro-Wilk test statistic, skewness, and kurtosis following the collapse of categories are shown in Table 4.2. For each scale, the original skewness and kurtosis was examined, as well as frequencies of responses within each category to determine the most appropriate combination of categories. Practical considerations of interpretability were also part of the decision making process. All scales initially had 7 categories. TC Child and TC Class were collapsed to 3 categories; TC Fam, OC Class, and OC Fam were collapsed to 4 categories; and TC Child was collapsed to 5 categories.

In the collapsed scales, the Shapiro-Wilk test of normality is significant for all scales, indicating that the distribution of scores are statistically significantly different from a normal distribution. However, the W value is high. The W value calculated by the Shapiro-Wilk formula is between 0 and 1. Distributions with a W value close to 1 are close to normal (Shapiro & Wilk, 1965). Thus, while the Shapiro-Wilk test indicated a non-normal distribution based on the p-value, the W value indicates a distribution approaching normality. This is likely due to the Shapiro-Wilk’s sensitivity to sample size (Field, 2005).

The level of excess kurtosis that was considered acceptable was set at +/- .75 and the level of acceptable skewness was set between +/- 1.0 based in a demonstration of normality testing on raw and scaled scores (Ho & Yu, 2014). While the original data collected had issues of negative skewness, data were collapsed at the lower end by combining categories such as ‘Not in the last 6 months’ and ‘Once in the last 6 months’. This served to improve the normality of items, and thus the normality of the overall scales. The scales were also somewhat platykurtotic, with excess negative kurtosis. The collapsing of categories also seemed to improve this issue. As shown in Table 4.2, the collapsed categories had acceptable skewness and kurtosis in the scores.

Table 4.2.

Evaluation of Normality of Distribution of Scales Following Category Collapse

	Variables	Mean	SD	Categories	Shapiro-Wilk	Skewness	Kurtosis	N
1	TC Class	11.44	2.38	3	.94*	-.25	.20	254
2	TC Child	23.60	6.81	3	.97*	-.18	.41	329
3	TC Fam	10.91	2.93	4	.95*	-0.35	.21	227
4	OC Class	13.01	4.21	4	.95*	-0.13	.67	268
5	OC Child	25.45	8.41	5	.95*	.55	.17	286
6	OC Fam	13.27	4.00	4	.94*	-0.29	.52	223

*p < .05

Another way to check the data was through graphing the distribution. Figure 4.1 indicates that there are some outliers at the low end of TC Class and TC Child. The scales all had a wide spread of scores across their relative ranges. OC Child had a particularly normal looking boxplot, with the median score at the middle of the inner quartiles and the legs of the box plot showing even distribution from both ends. Still, this distribution did not meet criteria sufficient for a non-significant Shapiro-Wilk test.

Thus, the normality statistics were mixed. While skewness and kurtosis were acceptable for all scales, the Shapiro-Wilk was significant, indicating non-normal distribution. However, the W value was high, which is often associated with a non-significant Shapiro-Wilk. For this reason, I determined that the data were sufficient for analysis. As noted in the title of Table 4.2, the analysis of normality was conducted following category collapse, which was necessary to meet this first assumption. All subsequent analyses are also administered with the collapsed scales to continue testing assumptions with the same set of data.

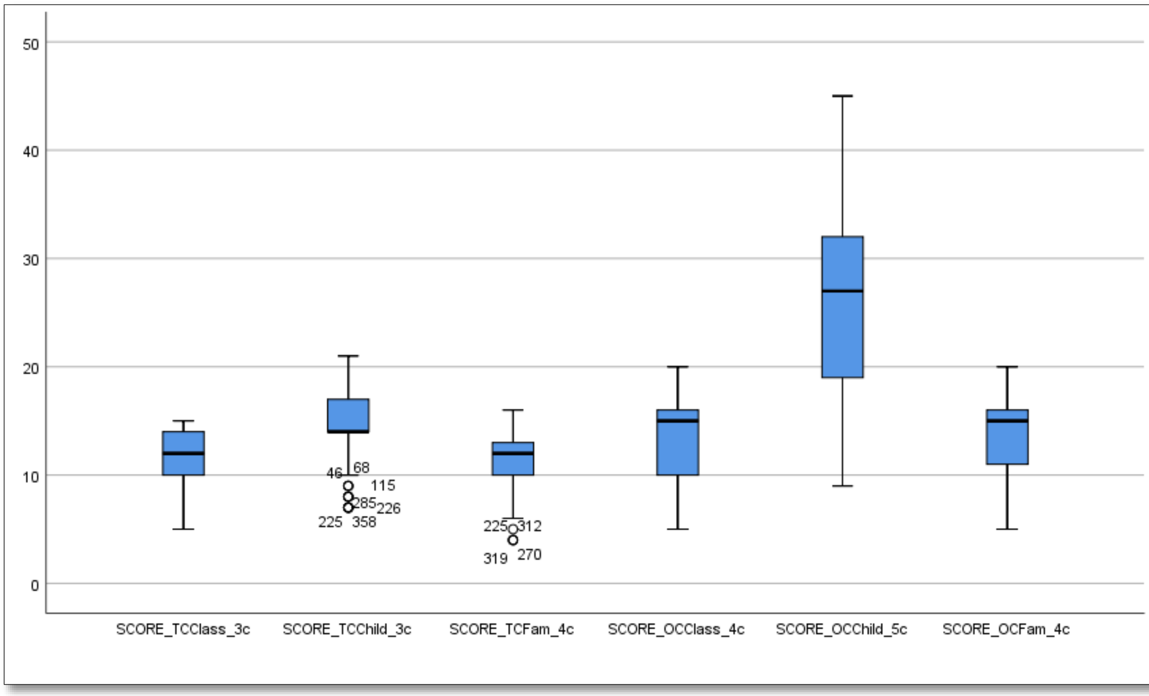


Figure 4.1. Boxplots of the Six Scales' Scores

H4: Scales are unidimensional and measure unique constructs

Unidimensionality.

Preliminary analysis for factor analysis were checked - see Table 4.3. The Kaiser Meyer Olkin (KMO) estimates were greater than .80, denoting that data were sufficiently correlated for factor analysis. Additionally, Bartlett's test of sphericity was significant for all scales, indicating that we can reject the null hypothesis that the correlation matrix is an identity matrix. Similarly, determinants are another indicator that the item correlation matrix is not a singular or identity matrix; all were greater than .00, indicating that some multicollinearity exists among the variables and a latent factor can potentially be revealed through factor analysis. The communalities varied, but were primarily above .40, indicating a good amount of correlation in the original data. Further, the percent of variance captured by factors was evaluated. All scales captured more than 50% of the variance, indicating that the items within the factors represented broad aspects of the constructs. All analyses were conducted using principal axis factor analyses. The assumption of unidimensionality was tested based on the rule of extracting only one factor with an eigenvalues greater than 1, or the Kaiser criterion (Costello & Osborne, 2003).

Table 4.3.

Estimates evaluating the appropriateness and results of Exploratory Factor Analysis

	KMO	Determinant	Communalities	% Variance	Factor Loadings	N
TC Class	.84	.106	.32 - .61	64.1	*	254
TC Child	.92	.004	.56 - .75	71.6	*	329
TC Fam	.79	.119	.48 - .64	72.6	*	227
OC Class	.90	.010	.64 - .81	80.8	*	268
OC Child	.93	>.000	.62 - .91	78.7	*	286
OC Fam	.86	.014	.56 - .84	77.6	*	223

Note: *Factor loadings were not reported when only one factor was extracted.

As the Kaiser criterion is criticized due to its tendency to overextract, it can be considered a conservative rule when applied to test for unidimensionality, as it is here (Zwick & Velicer, 1986, as cited by Reise, Waller, & Comrey, 2000). That is, as the Kaiser criterion would tend to overextract rather than underextract, and we can consider this a conservative test. Thus, the assumption of unidimensionality holds for all scales, as only one factor was extracted for each.

A second way to test unidimensionality was to examine correlations greater than .50 between scales. Some scales were highly correlated, as shown in Table 4.4. The highest correlations were between TC Class and TC Child (r = .65), OC Class and OC Child (r = .70), and between TC Family and OC Family (r = .68). These are beyond the recommended level of r = .50 to consider constructs sufficiently unique (Ruel, Wagoner, & Gillespie, 2016). Other scales that are estimated to have correlations above .50 include OC Child and TC Child (r = .57), OC Fam and OC Child (r = .56).

Table 4.4.

Correlations Between Scales

		1	2	3	4	5	6
1	TC Class	1.00					
2	TC Child	0.65*	1.00				
3	TC Fam	0.41	0.42	1.00			
4	OC Class	0.41	0.41	0.43	1.00		
5	OC Child	0.42	0.57*	0.49	0.70*	1.00	
6	OC Fam	0.21	0.33	0.68*	0.41	0.56*	1.00

*Correlation is greater than .50, indicating possible overlap of constructs.

While the results of the factor analysis provide sufficient evidence that each scale is unidimensional, the correlations suggest that the scales may not be measuring unique constructs. Because the factor analysis suggests that the scales meet the criteria of unidimensionality, an important assumption has been met for the next step in analysis – item analysis using item response theory applications.

Evaluation of the Scales Using Item Response Theory Methods

Model Fit Statistics. Model fit statistics are shown in Table 4.5. The model fit statistics indicated that most of the scales were a good fit for the specified model – generalized partial credit model. Two of the scales – TC Child and OC Child did not have a good fit with the model. C2 statistics are goodness of fit statistics that, like a chi-square fit statistic, should be non-significant to indicate a good fit. C2 uses limited information to adjust for a small number of items and sparse data (Cai & Monroe, 2014). Significant p-values (< .05) for the two child scales indicate that these models were not a good fit. The C2 for TC Class is moderately significant and is not significant for OC Class, indicating good fit for OC Class.

The Tucker-Lewis Index (TLI) compares the fitted model to a null model with the same data and degrees of freedom (Tucker & Lewis, 1973). Higher numbers are considered more favorable, with a common rule of greater than .95 considered an adequate fit. While the TLI indicates that the models were a good fit, and were used as evidence for moving forward, the more rigorous fit statistic (C2) was non-significant for just one scale.

Table 4.5.

Model-Fit Statistics

	C2	df	p	TLI	N

TC Class	11.9	5	.04	.98	254
TC Child	42.3	14	<.00	.99	329
TC Fam	17.1	2	< .00	.92	227
OC Class	9.0	5	.11	.99	268
OC Child	145.6	27	.00	.97	286
OC Fam	19.1	5	.00	.98	223

Item Fit Statistics. Item fit statistics were estimated for all items in each scale, including S-X2 and RMSEA. S-X2 is a goodness of fit statistic. Thus, a model is interpreted as having a good fit if the S-X2 statistic is non-significant. As shown in Table 4.6., the item fit statistics for the TC Class were mostly a good fit. The S-X2 test for TC Child items indicated that all items were a good fit. Across the OC scales, the OC Class had poor fit for all items (note that this was the only scale that had model fit with C2). The Root Mean Square Error of Approximation (RMSEA) for these items varied from .01 to .12. RMSEAs of less than .05 are commonly considered adequate (Cai & Monroe, 2013).

Table 4.6.

Item Fit Statistics

	TC Class (N=254)		TC Child (N=329)		TC Fam (N=227)	
	S-X2	RMSEA	SX2	RMSEA	SX2	RMSEA

Item 1	19.6*	.11	12.7	.04	15.8*	.10
Item 2	7.6	.06	13.9	.04	15.7*	.11
Item 3	4.7	.00	8.8	.00	20.4*	.09
Item 4	16.0*	.09	11.1	.03	31.1*	.12
Item 5	9.3	.06	14.4	.04		
Item 6			9.4	.02		
Item 7			9.1	.01		
Item 8						
Item 9						

	OC Class (N=268)		OC Child (N=286)		OC Fam (N=223)	
	SX2	RMSEA	SX2	RMSEA	SX2	RMSEA
Item 1	73.6*	.14	48.1*	.06	15.2	.06
Item 2	66.9*	.14	24.0	.02	11.6	.05
Item 3	43.4*	.12	28.3	.03	24.9*	.09
Item 4	142.5*	.32	25.4	.03	47.2*	.12
Item 5	40.5*	.11	28.0	.04	19.1*	.07
Item 6			44.9*	.06		

Item 7		64.9*	.08		
Item 8		122.2*	.10		
Item 9		60.1*	.07		

*p-value < .05

RQ4: Are items discriminating at an acceptable level?

H5: Item discrimination estimates are acceptable

Discrimination. Hypothesis 5 was tested by examining the estimates of discrimination (α), shown in Table 4.7. Acceptable discrimination estimates of .80 or greater are expected for generalized partial credit models (De Ayala, 2009). Across all scales, only one item did not meet the criteria of .80 for an acceptable discrimination level – Item 8 on the OC Child scale. This item "I formed small groups based on data about children's learning and development" may need to be dropped. The low discrimination will mean that answers to the question will not contribute to distinguishing well between people who have lower and higher data use practices. Another item that was somewhat low (.94) was on the OC Fam scale – "I used data to plan family engagement activities with other staff." It is possible that this item was not discriminating well because the behavior is not under the choice of the data user. Some programs require family engagement staff to partner with teaching staff on a regular basis (i.e. once a month). Other programs may support the partnership and still others may give very little attention to partnerships between teaching staff and family support staff. In the case of little attention being given to the behavior by support networks, the teacher may have very little say in the frequency of this activity.

Table 4.7

Estimates of Discrimination

	TC Class (N=254)	TC Child (N=329)	TC Fam (N=227)	OC Class (N=268)	OC Child N=286)	OC Fam (N=223)
	α	α	α	α	α	α
Item 1	1.30	1.75	2.06	1.48	1.57	2.16
Item 2	2.12	1.69	3.35	1.56	1.85	3.42
Item 3	2.55	2.22	1.00	2.31	2.34	1.94
Item 4	2.07	2.39	1.12	12.07	2.54	.94
Item 5	2.07	2.48		2.48	2.10	1.72
Item 6		2.12			1.80	
Item 7		1.87			1.38	
Item 8					.78	
Item 9					1.10	

RQ5: Are items measuring across the construct’s latent continuum?

H6: Item locations measure across the construct level.

Item Location. The location of items was expected to cross over two units of measure in the latent continuum - for example, from -1 to 1. Additionally, the gaps between category thresholds was expected to be .25 or more. Table 4.8 shows the location parameters between the two lowest and two highest categories. The purpose of this table is to show the spread of locations that are captured by the items, sometimes called bandwidth, indicating that the categories function to cover a breadth of the latent spectrum. The wide bandwidth of a scale, when joined by items that are sufficiently discriminating (see description of Table 4.7), allows for measurements to be taken on a wide variety of participants – both those whose responses manifest high data use and those whose responses indicate low data use practices. As shown, almost all of the items have a bandwidth of at least two theta units. Exceptions include TC Class: Item 2, with a bandwidth of 1.77 and OC Class: Item 5, with a bandwidth of 1.86. While these items are not at or above the criterion of 2, their bandwidth was fairly close, indicating a respectable range of locations could potentially be captured by these items.

Table 4.8.

Item Location Parameters of the First and Last Category Thresholds

	TC Class (N=254)		TC Child (N=329)		TC Fam (N=227)	
	$\delta_{.1\text{-}2}$	$\delta_{.2\text{-}3}$	$\delta_{.1\text{-}2}$	$\delta_{.2\text{-}3}$	$\delta_{.1\text{-}2}$	$\delta_{.3\text{-}4}$
Item 1	-1.91	1.12	-1.18	1.08	-2.37	.97
Item 2	-1.11	.66	-1.53	1.19	-2.24	.78
Item 3	-1.74	.37	-1.34	1.14	-1.06	2.50

Item 4	-1.93	.31	-1.47	.83	-.76	1.85
Item 5	-1.63	.87	-1.25	1.09		
Item 6			-1.00	1.51		
Item 7			-1.03	1.49		
Item 8						
Item 9						

	OC Class (N=268)		OC Child (N=286)		OC Fam (N=223)	
	$\delta_{.1\text{-}2}$	$\delta_{.3\text{-}4}$	$\delta_{.1\text{-}2}$	$\delta_{.4\text{-}5}$	$\delta_{.1\text{-}2}$	$\delta_{.3\text{-}4}$
Item 1	-1.36	1.56	-3.42	2.69	-2.36	1.11
Item 2	-1.15	1.18	-2.84	2.71	-2.17	.90
Item 3	-1.14	.81	-2.76	2.41	-2.33	.82
Item 4	-1.83	.19	-2.85	2.45	-1.29	2.06
Item 5	-.91	.95	-3.04	2.57	-1.86	1.28
Item 6			-3.18	2.92		
Item 7			-2.83	2.70		
Item 8			-2.77	1.33		
Item 9			-3.23	2.15		

A second purpose of examining item location parameters is to evaluate the ordering of the scales. A strength of the GPCM is that it does not assume the order of thresholds, thus allowing for threshold estimation to be informed by the observed data and testing the possibility that the expected order is not manifested in the estimates. Beyond bandwidth and ordering, location of category parameters should also be evaluated for redundancy of categories within an item. That is, category parameters should not be measuring at the same location, but rather be spread out so that each successive category covers a slightly different section of the latent continuum than does the previous category. While Table 4.8 allows us to examine bandwidth by showing the lower and upper ends of the latent continuum, Table 4.9 allows us to evaluate potential misordering among categories. Table 4.9 lists differences between thresholds, starting with the difference (or gap) between threshold 1 and 2 and threshold 2 and 3. For three categories, there are only two thresholds, and therefore only one gap to evaluate – between the first and second threshold. Likewise, for the scale with the most categories (TC Child which has five categories), there are four category thresholds and thus three gaps to evaluate. Only the gap between adjacent categories is evaluated. That is, the table (4.9) only shows calculations of the difference between one threshold and the successive threshold. Gaps that are below a criterion of .25 are highlighted, as they indicated thresholds that are too close or, if negative values, they indicate misordered categories. When category thresholds are very close, then the categories' may be redundant and in need of being collapsed. Because categories have already been collapsed, this shouldn't be an issue unless the items has other issues, such as discrimination that is too high.

As shown in Table 4.9, most items do follow the expected order, with the first threshold having the lowest theta value and the last threshold having the highest theta value. Exceptions to this are gaps with a negative value – Items 3 and 4 of the TC Family scale and Item 4 of the OC Family scale. While not misordered, Item 4 of the OC Class scale has less than a .25 gap, indicating that it has some redundancy in the third and fourth category. This small gap is also due to the extremely high discrimination of that particular item (see Table 4.7).

Table 4.9.

Gaps Between Category Thresholds

	TC Class	TC Child	TC Family		OC Class		OC Child			OC Family	
	δ.1- δ.2	δ.1- δ.2	δ.1- δ.2	δ.2- δ.3	δ.1- δ.2	δ.2- δ.3	δ.1- δ.2	δ.2- δ.3	δ.3- δ.4	δ.1- δ.2	δ.2- δ.3
Item 1	3.03	2.26	1.59	1.75	1.56	1.36	2.71	2.17	1.23	1.26	2.21
Item 2	1.77	2.72	1.49	1.53	1.24	1.09	1.66	2.4	1.49	1.31	1.76
Item 3	2.11	2.48	-0.34	3.9	1.52	0.67	1.75	1.97	1.45	0.92	2.23
Item 4	2.24	2.3	-0.23	2.84	1.81	0.21	1.8	1.89	1.61	-0.02	3.37
Item 5	2.5	2.34			0.77	1.09	2.07	1.81	1.73	0.8	2.34
Item 6		2.51					2.4	1.75	1.95		
Item 7		2.52					1.92	2.08	1.53		
Item 8							1.43	2.35	0.32		
Item 9							2.19	2.51	0.68		

Note. Numbers that are underlined denote threshold gaps less than .25.

Evaluation of the DUCES Discriminant Validity

RQ6: Do scales discriminate between higher- and lower-level data users?

H7: Scales discriminate between high- and low-level data users.

Discriminant validity was assessed by comparing the mean DUCES scores of two groups that were expected to be contrasting. The contrasting groups were created by gathering ratings from key insiders. The rating – a brief 10-item survey – was provided only to those who had experience working with the teacher in the context of using data. Raters provided ratings on 112 teachers.

The ratings were scored by summing all items. The scores were used to create high and low groups based on an approximate mean split. The mean was 89.3, with a confidence interval of 87.6 – 90.9. The low group was made from teachers who were rated at 87.6 and below; the high group was made from teachers who were ranked at 90.9 and above. However, only 54% (n=61) of teachers who were rated by insiders also selected to participate in completing the DUCES survey. Of the 61 teachers who completed the DUCES survey, 15 had been placed in the low category based on their rating and 37 had been placed in the high category based on their rating; 9 were excluded from analysis to create two differentiated groups. Thus, the validation ratings may be biased due to fewer low-rated teachers completing the DUCES survey than did high-rated teachers.

Because the number of teachers rated by insiders was much larger than the number of teachers who completed the DUCES survey, it was suspected that the teacher sample who selected to complete the DUCES survey may be biased. Thus, a sensitivity check was conducted to assess the appropriateness of the insiders' rating data to discriminate between high and low data users, as indicated by the DUCES.

When checking to learn whether the validation ratings were biased due to fewer low-rated teachers completing a DUCES survey, a significant mean difference on the insider rating was found between teachers who completed the DUCES survey and those who did not – t (110) = -2.65, p < .01. The mean insider rating score for those who did not complete the DUCES was 86.8 (n = 51) and the mean insider rating score of those who did completed the DUCES survey was 91.3 (n=61). Thus, the sample of matched insider ratings and DUCES surveys should be considered biased, as a group of teachers who were considered lower-capacity data users by insiders selected not to complete the DUCES survey.

Despite the compressed scores of the matched insider ratings and DUCES surveys, analyses were conducted to assess whether any discrimination could be detected amongst the biased data. Mean differences between high and low groups were tested on each of the six scales with an independent t-test – reported in Table 4.10.

Table 4.10.

DUCES Scales Mean Differences in High and Low-Rated Teachers

(N = 52)

Scale	Group	X	n	T	df	N
TC Class	High	28.6	29	-1.72*	38	40
	Low	24.9	11			
TC Child	High	36.8	37	-1.04	49	51
	Low	33.6	14			

TC Fam	High	13.4	28	-0.22	34	36
	Low	13	8			
OC Class	High	23.4	36	-0.03	48	50
	Low	23.3	14			
OC Child	High	35.2	36	0.82	47	49
	Low	38.9	13			
OC Fam	High	16.6	31	0.59	37	39
	Low	18	8			

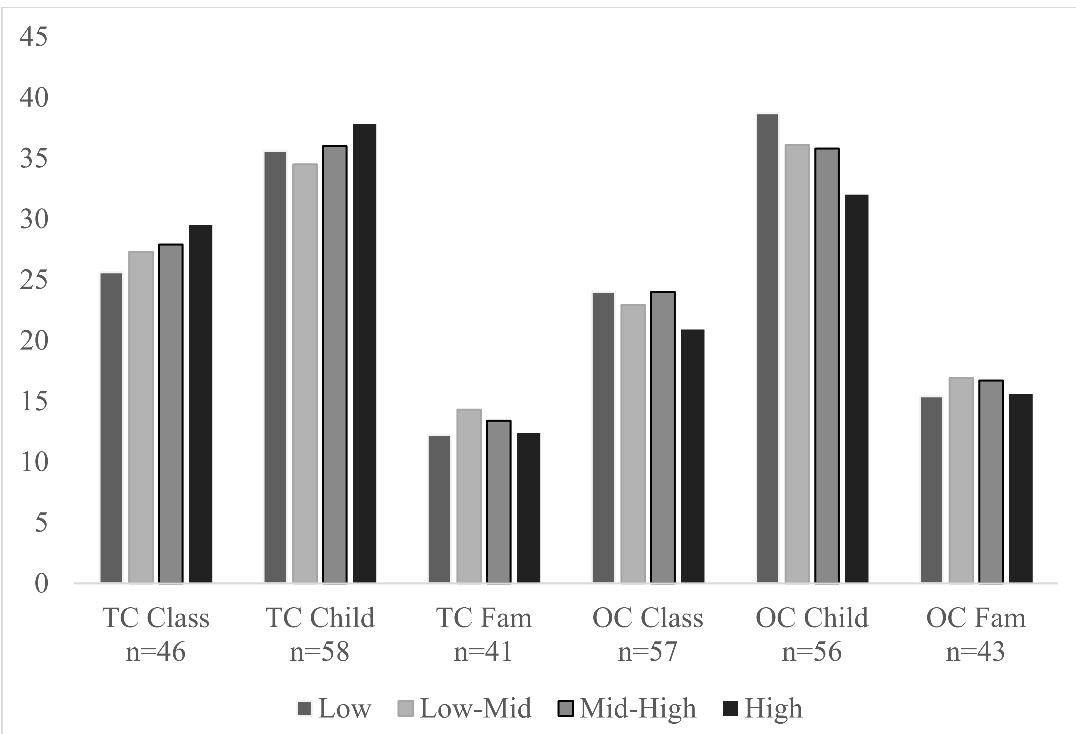
Note. $p < .10$

Differences were expected in the means of the six scales. There were no statistically significant differences between the high and low matched samples. The TC Class score means had a marginally significant difference ($p < .10$), with a higher mean score for the higher ranked group.

This pattern is somewhat visible in Figure 4.2. While only a marginal difference, this may reflect that classroom quality feedback (classroom data) provided by the insiders (sometimes called Master Teachers or Instructional Coaches) were used by the teachers who completed the DUCES. Thus, classroom teachers are using classroom data provided by the insiders, and the insiders were most aware of this use – lending strength to the statistical relationship between the insider ratings and the teachers' use of teacher-collected classroom data.

Interestingly, while not significant, there seems to be a negative correlational trend between the insider ratings and the OC Child scale. The high-rated data users self-report using other-collected child data less frequently than do low-rated data users. Unfortunately, the biased sample, as denoted by the sensitivity check, likely has constrained correlations reducing the interpretability of these relationships. Another possibility is that the use of teacher-collected data is considered most effective by teachers and the insiders who completed data use ratings; the time taken to use teacher-collected data may displace the need for or time available to devote to the use of other-collected data.

Insider ratings were used to create four groups, rather than two, in order to visually explore patterns. The mean scores of the scales across the four groups are shown in Figure 4.2.



Figures 4.2. Mean Differences Between Higher- and Lower-Users of Data

Synopsis and Synthesis

Each scale was evaluated with the seven hypotheses. These evaluations have led to the following conclusions about how the measures should be used in the field. The first hypothesis was that items would be interpreted as intended – an important aspect of the validity of items. Overall, while the cognitive interviewing and behavioral coding indicated more categories were needed to provide respondents with choices that matched their experience, this was problematic to the normality of items and scales as well as to the fit statistics that relied on expected cell distribution to make comparisons to the observed data. Thus, the first evaluation led to the slight adaptations in wording of the items and additions to the response categories – however these response categories should be collapsed when scoring and using the measure.

The second hypothesis was that items were calibrated to a theoretical level. This was expected to be seen through the agreement of panelists rating the items at the intended level. However, while some panelists rated the items at the same level as the developer intended the item to measure – use of data to 1) learn, 2) build knowledge, or 3) respond – the raters often chose multiple levels for each item due to the progressive nature of data use, according to one rater. Another way to look at calibration was to examine the variation of location in the items. These were also not widely varying for any scales. Thus, while the calibration of items to theory added to validity, these items were not empirically different. That is, there was little evidence that items developed to assess ‘using data to respond’ were more difficult to endorse than items developed to assess ‘using data to learn’. For these reasons, further analysis indicating that some items need to be trimmed will not take into consideration whether loss of the item will be a loss of information at one of the three levels of data use. The third through sixth hypotheses will be examined per scale. The seventh hypothesis was only confirmed for the TC Class scale and will be addressed only in the synopsis of that scale.

TC Class. The TC Class scale was had a slight negative skew, but skewness and kurtosis were within an acceptable range and the W value of the Shapiro-Wilk test was high (.94), although significant. The EFA indicated a single factor that was correlated with one other scale – TC Child (r=.65). The model fit was marginally acceptable with, C2 = 11.9, p=.04, TLI = .98. The items that did not fit included "I used data to identify ways to strengthen my teaching. I used data to create strategies that support children's learning and development." When removing item 1 from the model, model fit statistics improved, but there were still two items with poor item fit. However, removing item 4 improved both model fit and item fit. Thus, the final recommended use of this measure will be without item 4. This scale was also the only scale that showed concurrent validity with the ratings given by teachers' supervisors of their data use abilities.

Figure 4.3 shows the category characteristic curves for the recommended set of items.

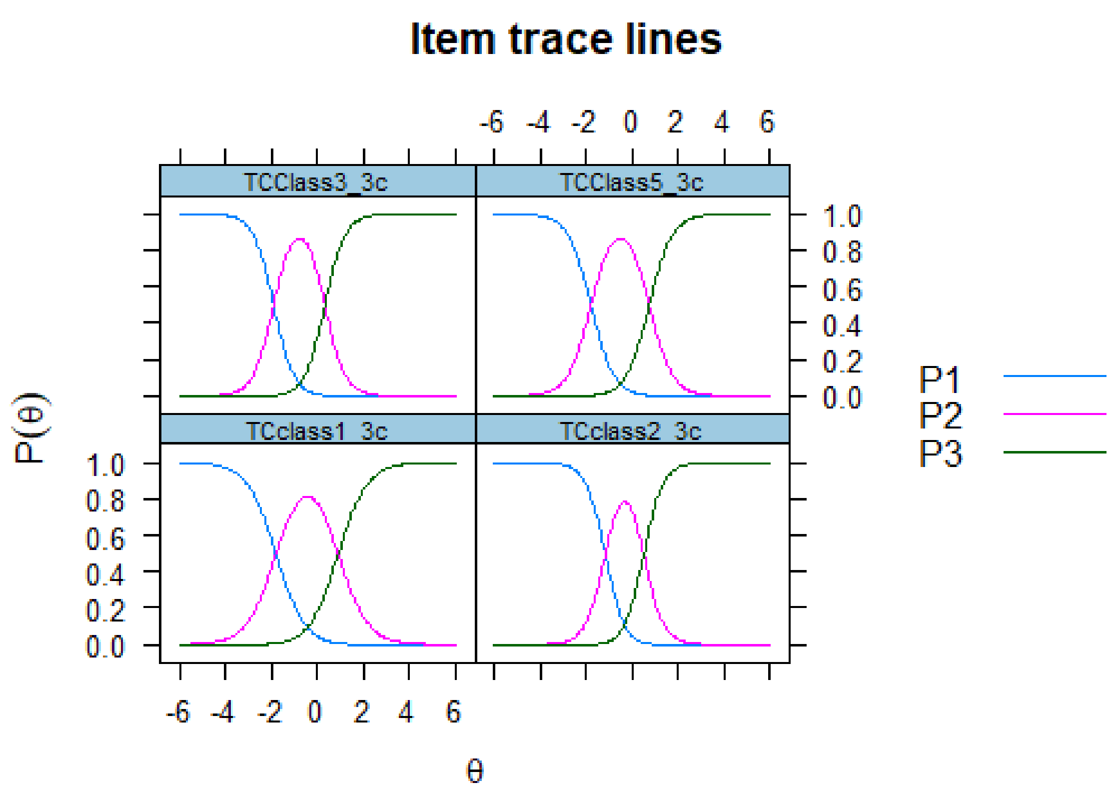


Figure 4.3. Category characteristic curves of the final set of TC Class items

OC Class. The OC Class scale, while having good model fit statistics, had poor item fit statistics for all items. Additionally, the item that was problematic in the TC Class set had an unusually high discrimination in the OC Class set. Removing the item did not improve the model fit or the item fit statistics to acceptable levels. While the removal of item 4 did not improve the item fit statistics, the high discrimination caused an unacceptably small gap between category thresholds. Thus, it is recommended that this item be removed. A plot of the items with item 4 removed is shown in Figure 4.4.

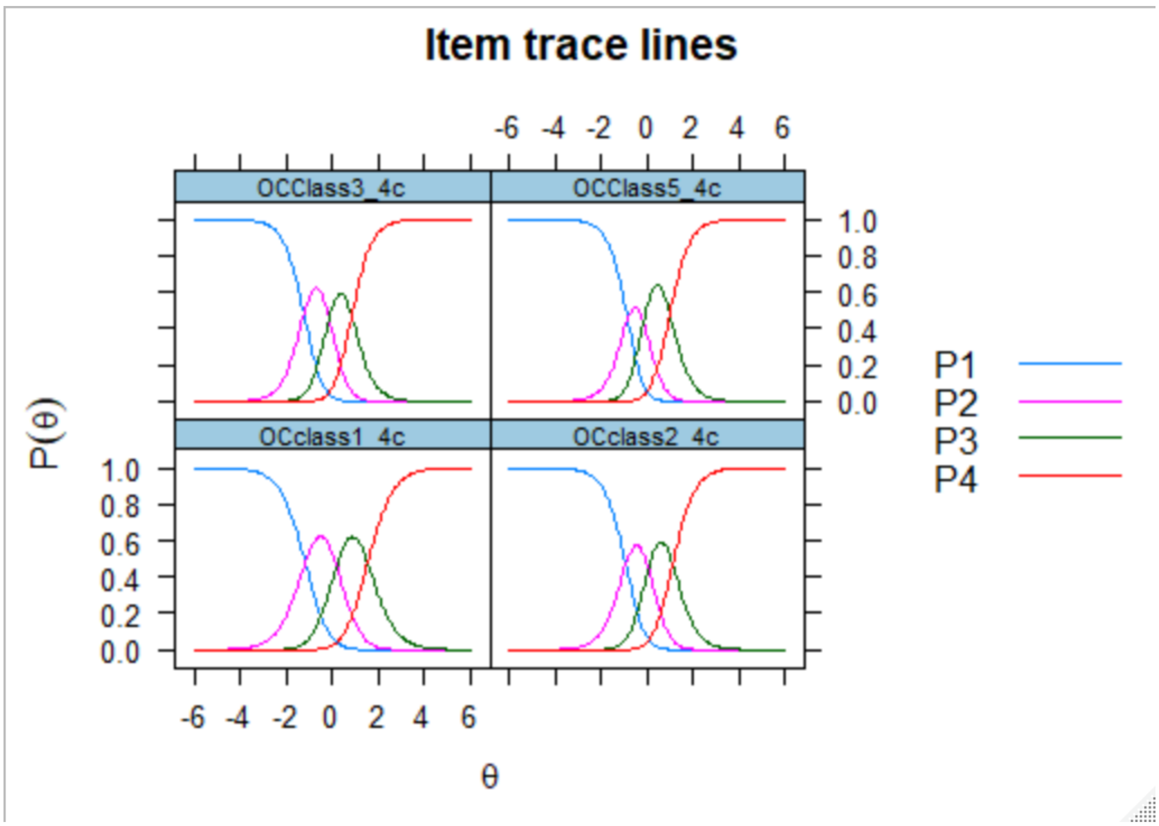


Figure 4.4 Category characteristic curves of the final set of OC Class items

TC Child. The TC child scale had slight negative skew. The factor analysis had a high KMO and high communalities, supporting one factor. The model fit for the gpcm was poor based on the C2, but the TLI was acceptable at .99. Item fit statistics were acceptable for all items using both S-X2 and RMSEA < .05 criteria. All discrimination parameters were above .80, with the lowest being at 1.69. The location parameters measured across the bandwidth of the latent continuum and did not have unacceptably small gaps or misordered categories. There are no recommended changes for this scale based on these analysis. The plot of the final set is shown in Figure 4.5.

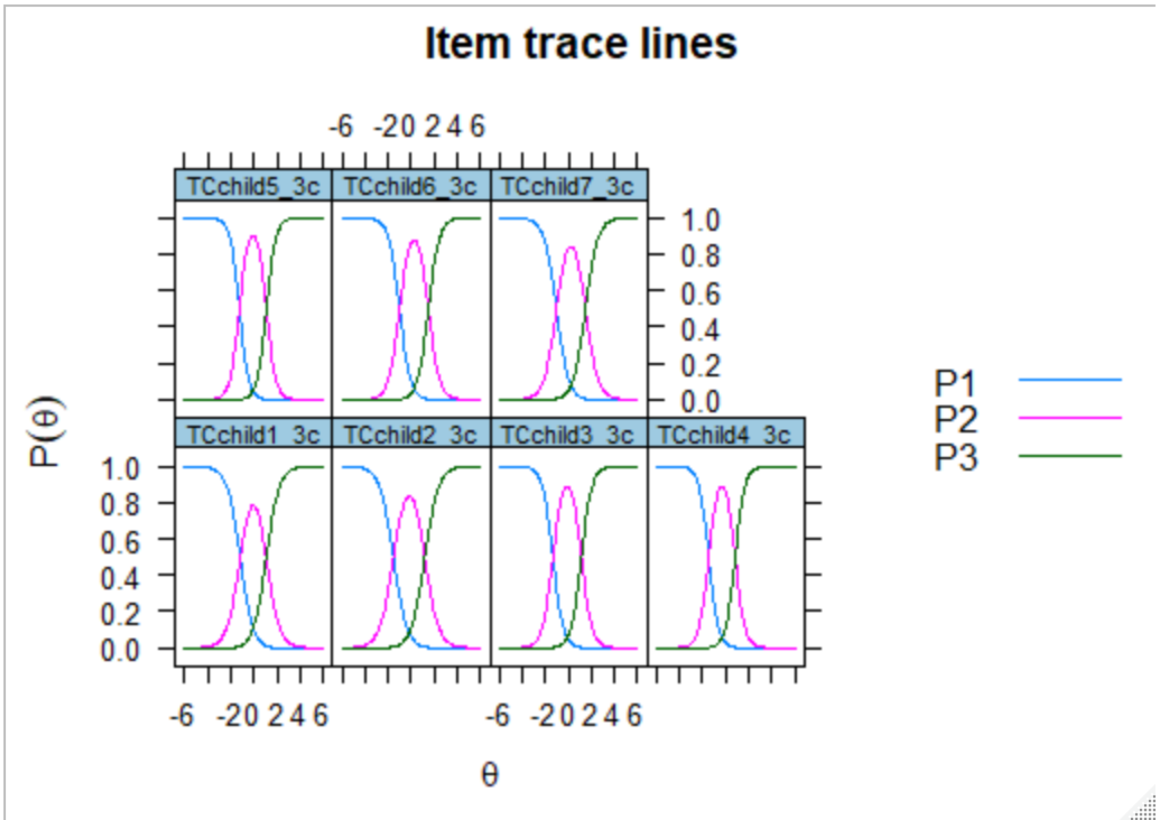


Figure 4.5 – Category characteristic curves of the final set of TC Child items

OC Child. The OC Child measure had fair normality, with a W value of .95, but a significant Shapiro Wilk test. Skewness and kurtosis were within acceptable levels. The EFA supported a unidimensional construct as a precursor to IRT analysis. The model fit statistics for the GPCM were not acceptable using the C2 and TLI was .97. Item fit statistics were acceptable for four items, but the S-X2 was significant for the other five items. The set of items were re-assessed for model and item fit, starting with the item with the highest RMSEA. The last three items were trimmed, which improved model fit somewhat. Both model fit and item fit statistics are still not at acceptable levels. The plot of items is shown in Figure 4.6.

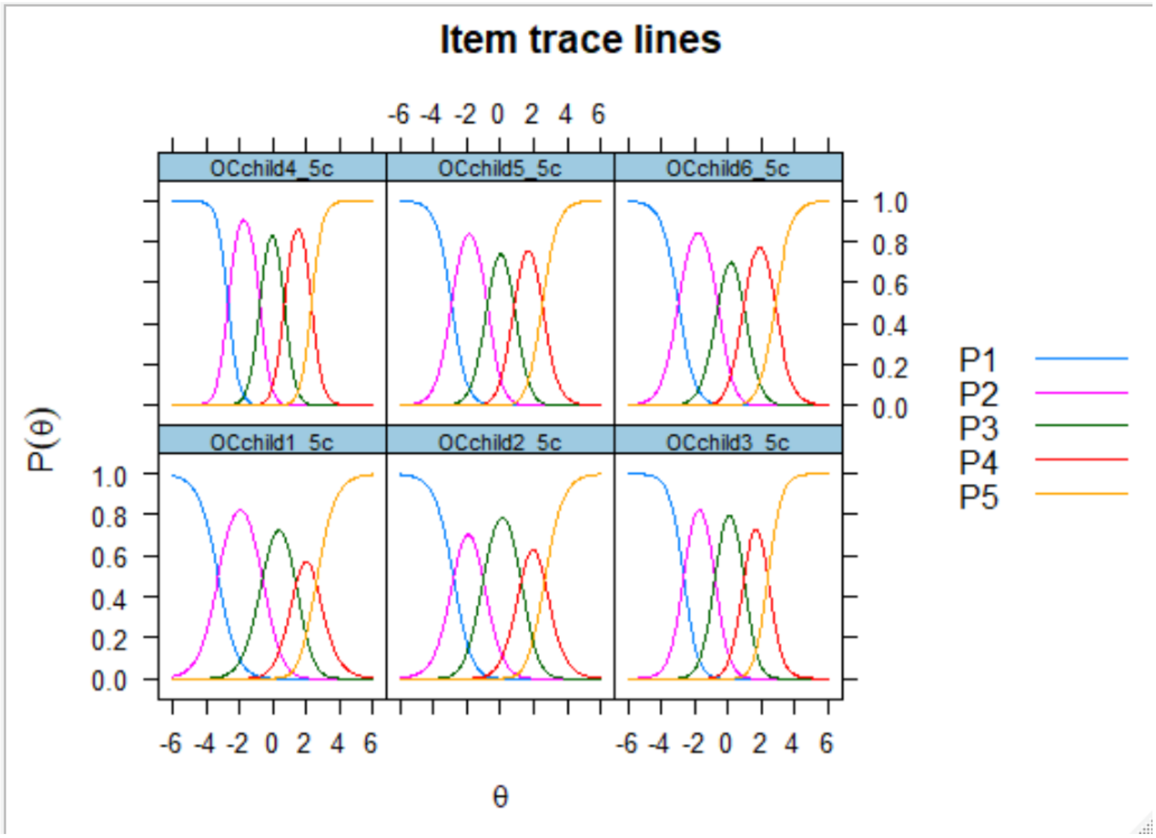


Figure 4.6 - Category characteristic curves of the final set of OC Child items

Family Scales. The clearest result of the IIOC process was that the scale that received the lowest ratings across all items also had poor model fit and item fit and reverse-ordered categories (TC Family and OC Family). The family items, because of the plethora of types of family that can exist and the multitude of ways that teachers could use these data, are not a strong set of items for use in evaluation, or professional development. The TC Family scale, in particular, does not seem useful as the model fit was very low (TLI = .92) and none of the items had acceptable fit statistics.

The OC Family had a TLI of .98 and only one item had an RMSEA greater than .09. It is likely that the high sampling from Head Start programs, in which a separate staff member has the role of working directly with families, resulted in family items working somewhat better in the context of others collecting the data. The item with poor fit in the OC Family set – "I use data to plan family engagement activities with other staff" also had low discrimination and misordered categories in the IRT analysis, both for OC and TC Family items. Thus, this item should be removed from both the TC Family and OC Family set of items. Removing this item improved model fit (C2 = 1.52, p = .48; TLI = 1.00), but only two items had acceptable item fit ("I looked at data as a way to learn about the family; I reflected on data as information about the strengths and needs of families.").

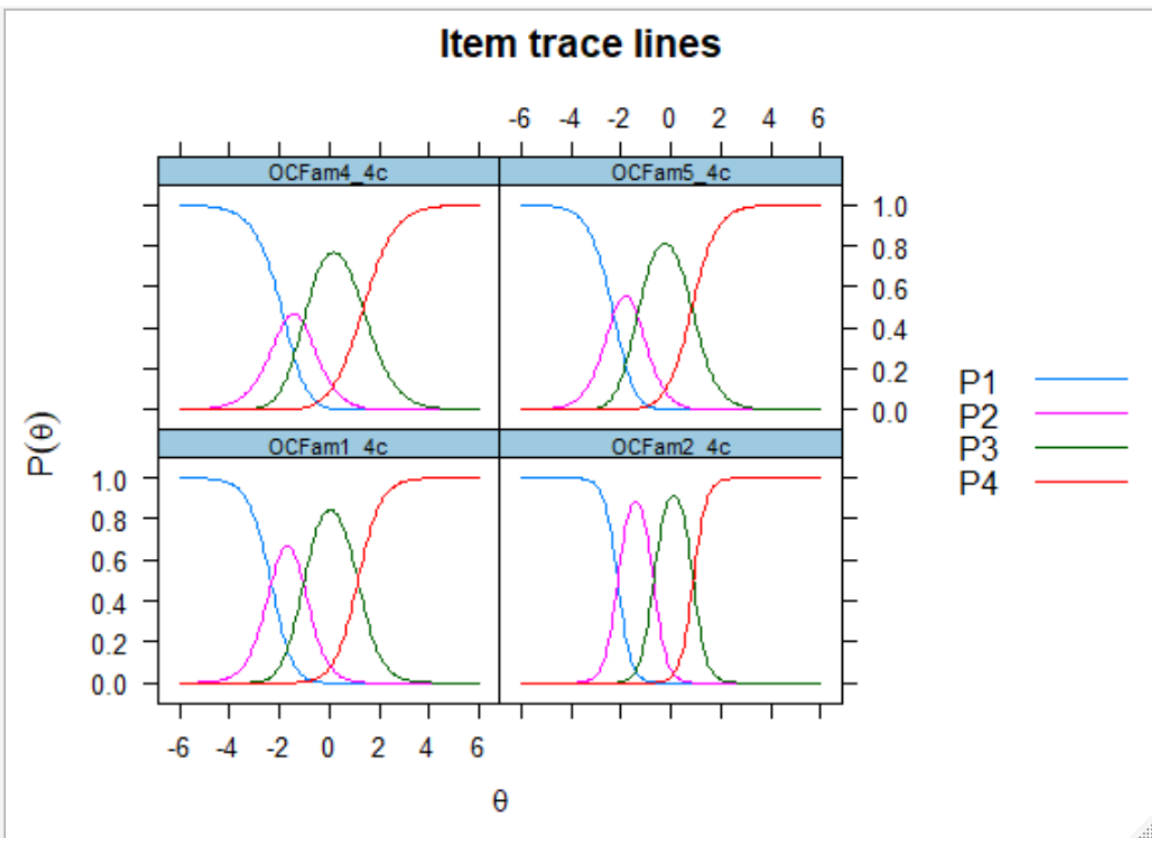


Figure 4.7 - Category characteristic curves of the final set of OC Family items

Removing the poorest fitting item from the TC Family set did not leave enough degrees of freedom to run the model fit test. Because this left few acceptable items remaining in the family sets, additional items were tested with a small subsample – TC Family $n = 68$; OC Family $n = 56$. They have internal consistency TC Family $\alpha = .89$; OC Family $\alpha = .93$, so may be sufficient as a descriptive scale.

This refined set of measures are graphed in Figures 4.3 through 4.7. The measures cover three types of data – child assessment, classroom quality, and family data. The measures also cover two contexts of data collection – by teachers and by non-teachers. However, the teacher-collected use of family data scale had poor model fit, item fit, misordered threshold locations and is not recommended for use. The OC Class scale had poor item fit, but adequate discrimination and location. This left four scales that have adequate model- and item-fit, acceptable discrimination, and cover a breadth of the latent spectrum and one scale (OC Class) that needs further research to clarify why its results were mixed. Additionally, the scale – TC Class – had evidence for validity based on the insider perspective ratings.

CHAPTER V

CONCLUSION

The six scales of the Informational Use section of the Data Use Contexts and Experiences Scales (DUCES) offer measures uniquely developed for early childhood contexts, tested with early childhood practitioners in multiple ways – through cognitive interviews, behavioral coding, and rigorous analysis of survey responses. The DUCES set of Informational Use scales will provide the early childhood field with a brief measure of teachers' data use practices. These measures capture data use across three commonly used types of data – classroom quality data, child assessment data, and data about families. Additionally, the measures can be used in different contexts – whether data are collected by teachers themselves or by others, such as evaluators, researchers, or special education staff.

The Informational Use Scales

The results of this study found evidence to drop one item from the TC Class scale. The final evaluation of the measure indicates that its psychometric qualities are sufficient for use in research and professional development.

Its marginal association with the validation ratings, despite the compressed score variability, indicate that higher data users will score higher on this measure. As the insiders who rated teachers were primarily responsible for assisting teachers with improving their classroom quality, the use of this measure to assess teachers' use of data for continuous quality improvement has some early evidence. More evidence that will verify the usefulness of this measure for professional development and research purposes include studies that include an intervention intended to change classroom quality with the use of classroom quality data as moderator of change.

The OC Class scale is recommended for use as a descriptive measure for exploratory research. Due to problematic item fit indices, the use of this scale for purposes beyond descriptive are not currently recommended. One area for potential research is how this measure differs between teachers who are employed in a school that emphasizes a learning culture versus one that emphasizes an accountability culture (Firestone & Gonzalez, 2007; Gannon-Slater et al., 2017). It is possible that measurement invariance will not be verified between teachers in these two cultures, as a learning culture may primarily focus on regularly-collected internal data and an accountability culture may focus primarily on annually-collected external data.

The TC Child scale is recommended for research or professional development without any changes to items. Like the TC Class scale, these items are most likely related to regularly-collected internal data. However, in the context of early childhood, a concern with the use of this scale is the dearth of useful formative measures of children's learning and development. For example, one of the most widely used curriculums in early childhood – Creative Curriculum -has an aligned assessment system – Teaching Strategies Gold. This system has been found to have low variability (Ehrlich, et al., 2019) and may be as more of a burden to teachers than a tool when used in accountability contexts (Kim, 2018). Thus, it will be important to use this scale as part of research that also gathers rigorous evidence about the type of data being used to interpret why teachers may or may not frequently use the data.

Likewise, the OC Child scale will need additional information about the data being used and the context of its use to understand the patterns of frequency of use in the early childhood field. For example, gathering data that may be provided from the Motivational Use scales – indicating whether the organizational culture leans more toward an accountability or learning culture – may provide important information about why other-collected data is or is not used. For example, in an accountability use culture that measures teacher performance and pay raises on end of year assessments, other-collected data from the previous year may be used all year long. In contrast, a learning culture may rarely look at other-collected data, opting to use regularly collected formative assessment to flexibly differentiate children's instruction based on their changing needs. One of the most interesting findings of the insider ratings data was that higher-rated teachers had visibly lower scores on the OC Child scale – refer to Figure 4.2. One possible interpretation of this findings is that teachers who are highly skilled at using data to make decisions are opting to use their own data rather than data collected by others. This echoes previous qualitative research on early educators (Ford, et al. 2016).

The two family data measures both had model fit and item fit problems. Due to this early finding, when additional data were collected late in the summer of 2019, some additional items were added. Those items had a minimal sample size, so could not be assessed at the level of this study, but basic internal consistency met frequently desired levels of .80. The measures, including their additional items, can be used for descriptive purposes, with the OC Family measure more highly recommended than the TC Family. The OC Family items had sufficient qualities to be used as shown in Figure 4.7. It is possible that the scale focused on family information collected by others had more reliability due to the heavy sampling from Head Start programs – which mandate a family support worker who can provide family information to the teacher. These two scales, like all the Informational Use scales, have been developed separately from the other types of data use so that programs can choose the set of scales that best meets their informational needs and organizational conditions.

Review of Development

The measures were developed across a four-year project, culminating in a set of questions with evidence for validity and reliability. Validity evidence assessed in this dissertation study was gathered from cognitive interviews and behavioral coding of pre-testing participants. Reliability evidence was assessed with tests of unidimensionality, which indicated single factor structures for each of the scales. Additionally, the application of IRT methods allowed for examination of how well items could discriminate between respondents, as well as how well the items measured across the latent spectrum of data use practices.

While the calibration of items using panelists to assess item fit within three theoretically-based levels of data use did not produce the expected results, there are some themes across the panel ratings and the item analysis. However, the process of planning the IIOC (Turner & Carlson, 2002) was a check of each item against a theoretical level of data use, from data as information, to knowledge, and then to response (Marsh & Farrell, 2015). The use of a theoretical framework, informed by a wide literature review (see Table 1.2 and Appendix 1) supported the design of items around the intended construct. These aspects of validity were embedded into the design of the measure, while more common types of validity, such as concurrent validity, were difficult to obtain due to the uniqueness of the measure. That is, the measure cannot be compared or contrasted to similar or dissimilar constructs due to the dearth of data use measures available to the education field, and especially in early childhood.

In addition to this measure being unique to the field, its development has been rigorous, using modern methods of measurement evaluation to consider ways to improve the measure. A particularly new method was the application of C2, a model fit statistic developed in 2014, to assess a small sample of respondents using a polytomous scale and few items (Cai & Monroe, 2014). This fit index, while the most appropriate for the type of data in this study, is limited in the scope of literature testing its utility and feasibility with various models, as has been done with older tests such as S-X2 or G2 (see Kang & Chen, 2010). It is possible that this particular test (C2) needs additional evaluation regarding its limitations and applications.

Implications

Contribution to theory. An expected contribution of this study to the theory of data use in schools was to test whether the data use framework articulated by Marsh & Farrell (2015) would be supported with empirical evidence. Specifically, there was an expectation that in both item paneling and estimates of item location, items would be aligned to a low-, mid-, and high-level of data use as they were designed to represent learning, knowledge-building, and responding. The process of data utilization in the Marsh & Farrell framework informed the design of items that represented each of these three levels. While there was some evidence that items designed to represent the highest level of data use (response) had higher item location estimates than those at lower levels, items designed for learning and knowledge-building did not consistently have respective low- and mid-level location estimates. For example, for both the TC Child and OC Child scales, the item "I used data to identify whether student learning needs were met using data" had slightly higher item location estimate in the last category than other items. This item represents a higher level in the data use framework because a teacher would have needed to use data to identify a child's needs through *learning*, consider how to meet those learning needs in the context of wider curricula expectations (*knowledge*), implement a plan (*response*), and then identify whether the child's learning needs were met through additional data, technically returning to learning within a cycle. Similar patterns were found for items in the TC Fam and OC Fam location estimates for the item "I used data to assess progress in the well-being of families."

However, while the TC Class item "I used data to evaluate the effectiveness of my plans" was at a high level, the same item was at a low-level within the OC Class scale. This further suggests that the OC Class scale has differences that need to be explored by examining it in the context of other DUCES scales.

Considerations for future data use models include the interchangeability of the first two levels – learning and knowledge – within the data use framework. This supports other frameworks for differentiating instruction for children with disabilities, including *universal design for learning*, which begins with learning about children, and *retrofitting* which begins with expectations set by the curriculum or standards (Thousand, Villa, & Nevin, 2015). Similarly, the Learning Trajectories framework developed by Clements and Sarama (2009) outlines a process of beginning with a learning goal, then using formative assessment or observation to learn about the child, followed by principled decision to implement learning activities that support children's development to the next level of learning. This debate was echoed at an AERA presentation in which special interest groups (sigs) from data use and curriculum were represented by researchers making a case for one or the other as the foundation for excellent pedagogy. (AERA, 2015). The answer for which one should be considered more important – data use or curriculum - is not clear and leaves empirical questions that can be explored in early childhood settings with these scales and the other scales of the DUCES. Rather than reordering the framework, it seems likely that the effectiveness of using data before or after considering curricular and pedagogical expectations will depend on the situation. A new theoretical framework may be more clearly stated as 1) synthesis of learning from data with curricular and pedagogical expectations, and 2) response – with a continuous cycle of these steps leading to more positive outcomes for teaching and learning. For this reason, the validity of the scale not threatened because, even with poor items removed, the new sets of items still covers these two levels – synthesis of learning and response. Thus, this study contributed to the theory of data use in schools by empirically testing the validity of a theoretical framework, finding that the order of learning from data and synthesizing information to create new knowledge are unlikely to need a specific order, as long as they each occur and precede the data-informed response. The study produced a research tool that covers knowledge-building as a synthesis of learning and response based on data-informed learning.

Contribution to research. With clarity on the theoretical cycle of effective data use, the measure contributes to research by providing a flexible set of measures that include both items related to learning and synthesis with other knowledge, as well as items designed to capture various types of data-informed responses by teachers. These two steps – synthesizing data with previous knowledge and data-informed responding – are reported as behavior frequencies that capture the latent trait of data use practices for early childhood teachers. With this tool of data use practices, the early childhood field can capture the presence and importance of data use.

In addition to the measures' contribution as tools with which to study a piece of the teaching and learning process, the measures may help to study the early childhood workforce – informing the amount of time that teachers spend in data use and in what contexts this is important . For example, will teachers' frequency of data use increase in the context of extrinsically motivating experiences? One of the scales that is not included in this study asks about teachers' experience of pressure related to using data. While this may be important to increase the amount of data use, will it moderate the relationship between data use and child outcomes? Research from the self-determination theory perspective would suggest that teachers' creativity and well-being would be diminished in a highly controlling context (Deci & Ryan, 1987, Deci, et al., 1982, Ryan & Deci, 2003). A study that included DUCES measures of data use practices, controlling significance, and classroom and child outcomes could contribute to questions of whether accountability cultures in schools help to improve teaching and learning. It may be that, in some contexts, a higher frequency of data use would not contribute to more optimal outcomes. If a study using this set of measures suggested that school cultures focused on accountability were not associated with improved teaching and learning outcomes, this would be an important contribution to educational research, particularly in terms of policy decisions.

In early childhood, we need to understand the conditions under which teachers will use data in ways that contribute to children's learning and development. These measures fill a gap in the literature by providing a measure to the field to further understand the presence and importance of data use to teaching and learning in early childhood settings.

Contribution to practice. These results can inform the field of early childhood to consider how data use is defined, to capture aspects of data use that are unique to early childhood, and to inform professional development. When data use is defined as synthesizing learning about children with knowledge of teaching, the expectations of data use become clearer for practice. Both practices of teachers and administrators can focus on collecting quality data that is aligned to learning goals and pursuing professional development experiences that ensure teachers understand how to scaffold children's progress toward such goals. This type of horizontal coherence across curriculum, teaching, assessment, and professional development will ensure that data use practice expectations support high quality support for teachers and students (Shephard, Penuel, Davidson, 2017).

The Informational Use scales as stand-alone measures can be aligned with program goals and used to evaluate progress toward those goals that are specific to early childhood contexts. While most schools use data about child learning to inform practice, early childhood programs more frequently use classroom quality data to reduce assessment burden on young children. Another contribution to practice would be to inform professional development. For example, if a program expects teachers to use data that they have collected about children, the TC Child measure provides valid and reliable information about this type of use. However, if the program goals are to use classroom quality data collected by others to inform efforts toward improvement, the OC

Class scale can provide information about the frequency of this type of data use. Tailored professional development that trains teachers who have lower frequency of data use to increase their efforts can be informed by results from these scales. Thus, professional development and school improvement efforts may be influenced by the use of these measures.

Limitations

Despite the TLI and RMSEA being indicative of reasonable fit for the models, the lack of fit indicated by the C2 and S-X2 statistics are a limitation. Thus, the results of this study will need further confirmation. The sample selection bias may be a major limitation, given the results of the discriminant validity sensitivity check – recall that teachers rated as low-level data users selected not to complete the DUCES. Using the scales with a larger and more diverse groups of teachers, specifically focusing on gathering data from teachers who are considered low-level data users, and gathering data from teachers who work with wider ages of children or who have varying support systems for data use will improve the ability to generalize the scales to those contexts.

Another limitation is that the data did not have a normal distribution according to a Shapiro-Wilk test. However, levels of skewness and kurtosis were acceptable. While the skewness, kurtosis, and boxplots gave some indication of normality for the data, the classic test of normality – the Shapiro-Wilk – did not. While this is most likely due to the Shapiro-Wilk's sensitivity to sample size, the data from the gpcm should be interpreted with some caution, as the maximum likelihood estimation process relies on the normality assumption for a starting guess prior to running iterations to estimate parameters. The normality issue or sample diversity issue may also have contributed to the issue of model fit.

Future Research

In addition to assessing model and item fit with a larger sample size, future research should include a confirmatory factor analysis, evaluation of convergent validity, and use of the tool in an intervention. Unidimensionality was tested with exploratory factor analysis. A confirmatory factor analysis is the next appropriate step to follow the exploratory factor analysis conducted in this analysis. Again, with a larger sample size, a confirmatory factor analysis could be used with respondents who are answering across multiple scales, thus ensuring that the scales are not capturing too similar a construct. While all items in the exploratory factor analysis were fit to a single factor based on the Kaiser criterion, with only one eigenvalue greater than one being generated from the analysis, correlations between scale scores indicated a correlation greater than .5 between many scales. This is evidence for the possibility of these measures are not a single construct.

A study conducted with another measure of data use, such as the EDIT, could also be used to assess convergent validity. If the survey measure were to be highly correlated with a measure conducted through observation and interviews, its value would be confirmed for use as a tool that would provide excellent information at an accessible cost. Finally, an important use of this measure to test its sensitivity in the future will be to use it in an intervention. If the measure were used as a pre- and post-test, with an intervention on data use conducted in between, the expected behavior of the measure would be for it to change for the group of teachers who received the treatment. Further, if this change slope were also related to child outcomes, the hypothesis that data use is an important component in early childhood education and care could be tested.

A hypothesis that was not confirmed was that the scales would distinguish between high- and low-level data users. The t-test test found only marginal differences between groups that had been rated as higher and lower level data users. Further the marginal difference was only on one scale – the TC Class. This scale asked questions about the way that teachers used data that they or other teachers collected. In many of the schools where this data was collected, the teachers' supervisors are called Master Teachers. Their role is to provide coaching and modeling of best practices to teachers, as well as observe them and give them feedback. It is likely that the ratings of teachers' data use proficiency was based on Master Teacher's experience with teachers receiving classroom observation feedback from them. If the Master Teacher's data was frequently used, this would be the evidence the Master Teacher used to rate the teacher. Further, this would be the data about which the teacher described her own experience using data.

Another interesting pattern in the ratings data is that teachers who were in the group rated high by Master Teachers and reported frequently using teacher-collected data, also reported using other-collected data less frequently. The TC Class and OC Class had a correlation of .51, while the TC Child and OC Child had a correlation of .64. These seem to have shared variance, but the patterns in the ratings group data indicate a possible tradeoff. Perhaps those who are frequently using teacher-collected data do not have time to frequently use other-collected data, and vice versa. This could be an interesting next step for testing the measure in a mixed-methods study – with the addition of teachers explaining when they use data, how it is prioritized among their other tasks, and if it ever displaces other work that needs to be done. Penuel (Penuel, Shepard, Hamilton, Miller, Gummer, & Mandinach, 2018) argued that data use routines are time intensive and that a focus on curriculum may be a more effective use of teachers' time.

Finally, the set of measures can contribute to a variety of questions, particularly when combined in studies with organizational, interpersonal, intrapersonal and student learning goals. These are some lines of inquiry that could be pursued. Are student learning goals met with more frequency when more data use is present? Is this relationship mediated by higher teacher attitudes or confidence? Is the relationship moderated by the type of organizational culture that dominates the staff experiences? Is the relationship stronger for certain sub-groups? What would sub-group differences mean for practice – considering teachers and children of color, children with various types of disability, and children of different ages? Does data use in early childhood matter more once teachers have specific learning goals? Are teacher's more likely to use data to achieve their goals when children are in certain age groups? Is that a function of education, professional development, or different needs of children? This is would further lead to differences in how much policy needs to push teachers to use data in certain ages. While the infant to three year old brain is growing rapidly, the job of the teacher may be to respond within the data of the moment, rather than the systematically collected data. Would teachers of infants who are and are not using data to inform their teaching have differences in the development of their students?

In addition to child goals, teacher goals could also be pursued. Interventions developed to improve teachers' use of data may seek to change attitudes toward data. Program evaluations focused on improving child outcomes may need to include these measures, as well as focus groups, to have a full picture of whether teachers' use of data is a function of teachers' trust in data. It is logical that, when teachers consider data of low quality or low usefulness in informing current goals, they would not want to use that data at a high frequency. It is also likely that if data are not of high quality or coherently aligned with teaching and learning goals, then the more frequent use of them would not lead to more positive outcomes. Thus, it is extremely important that users of these measures are informed that DUCES data should be interpreted in light of the possibility that more frequent data use may not be helpful due to contextual factors beyond this measure, such as what teachers think about the data.

In conclusion, this study has produced five acceptable scales. Two of these scales – TC Child and TC Class – seem acceptable for both research and professional development, as they have excellent discrimination and excellent bandwidth. The TLI model fit test and S-X2 item fit statistics support the conclusions of the IRT analysis and the scales should be considered ready for use. The OC Child and Class scales, while having many strengths such as acceptable discrimination and bandwidth, were not supported by the model and item fit statistics. Thus, there may be additional issues that need uncovering with these scales. They are acceptable for use, but would not be recommended for use to inform an individual teacher. Rather, the aggregate data from these measures may be informative in a larger context of other data points. Finally, the OC

Family scale can also be used in a similar way as the OC Child and OC Class, but needs additional cautions. While no scale is expected to capture all aspects of what is being measured, a sufficient sampling of behaviors are expected to capture the latent construct. I believe the Family scales are currently not broad enough to capture the types of data that teachers use about families. Despite the limitations of the scales related to data collected by those other than teachers and the family scales, this set of measures will add to the limited set of tools early childhood researchers currently have to understand data utilization in classrooms serving young children. Scales related to child and classroom quality data collected by teachers met all expectations of this study for factor structure, model fit, discrimination, and location – leading to the conclusion that these two scales, in particular, will be beneficial to early education research and professional development.

REFERENCES

Adams, C., Forsyth, P., Ware, J., & Mwavita, M. (2016). The Informational Significance of A-F School Accountability Grades. *Teachers College Record, 118*(7), Teachers College Record, 2016, Vol.118(7).

Akers, L., Grosso, P. D., Atkins-Burnett, S., Boller, K., Carta, J., & Wasik, B. (2014). *Tailoring teaching: Teachers' use of ongoing assessment to individualize instruction (Volume II)*. Washington, DC: U.S Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

American Educational Research Association [AERA], American Psychological Association [APA] National Council on Measurement in Education [NCME] (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Ayoub, O'Connor, Rappolt-Schlictmann, Vallotton, Raikes, & Chazan-Cohen. (2009). Cognitive skill performance among young children living in poverty: Risk, change, and the promotive effects of Early Head Start. *Early Childhood Research Quarterly, 24*(3), 289-305.

Bambrick-Santoyo, P. (2010). *Driven by data: A practical guide to improve instruction*. San Francisco, CA: Jossey-Bass.

Beecher, C., Abbott, M., Petersen, S., & Greenwood, C. (2017). Using the quality of literacy implementation checklist to improve preschool literacy instruction. *Early Childhood Education Journal, 45*(5), 595-602. doi:10.1007/s10643-016-0816-8

Bickman, L. & Rog, D.J. (2009). *The Sage handbook of applied social research methods*. Los Angeles: SAGE.

Boyd, J., Barnett, W. S., Bodrova, E., Leong, D.J., & Gomby, D. (2005). *Promoting children's social and emotional development through preschool education. Preschool Policy Brief*. New Brunswick, NJ: NIEER.

Buysse, V., & Peisner-Feinberg, E. (2010). Recognition & response: Response to intervention for PreK. *Young Exceptional Children, 13*(4), 2-13. doi:10.1177/1096250610373586

Cai, L., & Monroe, S. (2013). IRT model fit evaluation from theory to practice: Progress and some unanswered questions. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 102-106. doi:10.1080/15366367.2013.835172

Campbell, F. A., Wasik, B. H., Pungello, E., Burchinal, M., Barbarin, O., Kainz, K., . . . Ramey, C. T. (2008). Young adult outcomes of the Abecedarian and CARE early childhood educational interventions. *Early Childhood Research Quarterly, 23*(4), 452-466. doi:http://dx.doi.org/10.1016/j.ecresq.2008.03.003

Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). Making the most of interim assessment data: Lessons from Philadelphia. In R. f. Action (Ed.), *Learning from Philadelphia's School Reform*. Philadelphia, PA.

Clements, D., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach.* (Studies in mathematical thinking and learning). New York: Routledge.

Cosner, S. (2011). Teacher learning, instructional considerations and principal communication: Lessons from a longitudinal study of collaborative data use by teachers. *Educational Management Administration & Leadership, 39*(5), 568-589. doi: 10.1177/1741143211408453

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-9.

Crocker, L. M. & Algina. (1986). Intro to classical and modern test theory. New York: Holt, Rinehart, & Winston.

De Ayala (2009). *The theory and practice of Item Response Theory.* New York, Guilford Press.

Deci, E.L., & Ryan, R.M. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology, 53*(6) 1024-1037.

Deci, E.L., Spiegel, N.H., Ryan, R.M., Koester, R., & Kauffman, M. (1982). Effects of performance standards on teaching styles: Behavior of controlling teachers. *Journal of Educational Psychology, 74*(6), 852-859.

Derrick-Mills, T. Sandstrom, H., Pettijohn, S., Fyffe, S., Koulish, J. (2014). Data use for continuous quality improvement: What the Head Start field can learn from other disciplines: A literature review and conceptual framework. OPRE report #2014-77. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Derrick-Mills, T., Winkler, M. K., Healy, O., & Greenberg, E. (2015). *A resource guide for Head Start programs: Moving beyond a culture of compliance to a culture of continuous improvement*.

Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Dunn, Airola, Lo, & Garrison. (2013). What teachers think about what they can do with data: Development and validation of the data driven decision-making efficacy and anxiety inventory. *Contemporary Educational Psychology, 38*(1), 87-98.

Ehrlich, S., Pacchiano, D., Stein, A., Wagner, M., Park, S., Frank, E., Luppescu, S., Young, C. (2019). Early Education Essentials: Validation of Surveys Measuring Early Education Organizational Conditions. *Early Education and Development, 30*(4), 540-567.

Firestone, W.A., Gonzalez, R.A. (2007). Culture and processes affecting data use in school districts. *Yearbook of the National Society for the Study of Education, 106th*. Chicago, IL: Malden, MA: National Society for the Study of Education; Distributed by Blackwell Publishing.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics : And sex and drugs and rock 'n' roll / Andy Field.* (4th ed.).

Fitzpatrick, J.L., Sanders, J.R., & Worthen, B.R. (2011). *Program evaluation: Alternative approaches and practical guidelines – Fourth edition*. Boston, MA. Pearson.

Ford, T. G., Van Sickle, M. E., & Fazio Brunson, M. (2016). The role of "informational significance" in shaping Louisiana elementary teachers' use of high-stakes teacher evaluation data for instructional decision making. In K. K. Hewitt & A. Amrein-Beardsley (Eds.), *Student growth measures in policy and practice: Intended and unintended consequences of high-stakes teacher evaluations.* New York: Palgrave Macmillan.

Gannon-Slater, N., La Londe, P.,G., Crenshaw, H. L., Evans, M. E., Greene, J. C., & Schwandt, T. A. (2017). Advancing equity in accountability and organizational cultures of data use. *Journal of Educational Administration, 55*(4), 361-375. doi:10.1108/JEA-09-2016-0108

Gerzon, N., & Guckenberg, S. (2015). *Toolkit for a workshop on building a culture of data use. (REL 2015-063)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Gottfried, M. A., Ikemoto, G. S., Orr, N., & Lemke, C. (2011). What four states are doing to support local data-driven decision making: policies, practices, and programs. *(Issues & Answers Report, REL 2012-No.118)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs.

Henry, C., Morris, A.S., & Harrist, A. (2015). Family Resilience: Moving into the Third Wave. *Family Relations, 64*(1), 22-43.

Ho, A.D., Yu, C.C. (2014). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement.* 75(3), 365-388/ doi:10.1177/0013164414548576.

Institute of Medicine (IOM) and National Research Council (NRC). 2015. *Transforming the workforce for children birth through age 8: A unifying foundation.* Washington, DC: The National Academies Press.

Jimerson, J. (2016). How are we approaching data-informed practice? Development of the Survey of Data Use and Professional Learning. *Educational Assessment, Evaluation and Accountability, 28*(1), 61-87.

Johnson, R. S. (2002). *Using data to close the achievement gap: How to measure equity in our schools*. Thousand Oaks, CA: Corwin Press.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X 2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391-406. doi:10.1111/j.1745-3984.2008.00071.x

Kim, K. (2018). Early childhood teachers' work and technology in an era of assessment. *European Early Childhood Education Research Journal, 26*(6), 927-939.

Kippers, W. B., Wolterinck, C. H. D., Schildkamp, K., Poortman, C. L., & Visscher, A. J. (2018). Teachers' views on the use of assessment for learning and data-based decision making in classroom practice. *Teaching and Teacher Education, 75*, 199-213. doi:10.1016/j.tate.2018.06.015

Krugly, A., Stein, A., & Centeno, M. G. (2014). Demystifying data: Designing and implementing data-driven systems and practices for continuous quality improvement. *Zero to Three, 34*(3), 11-19.

Lee, J., & Reeves, T. D. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State 1990-2009 NAEP reading and math achievement gaps and trends. Educational Evaluation and Policy Analysis, 34(2), 209- 231.

Liang, T., & Wells, C. S. (2009). A model fit statistic for the generalized partial credit model. Educational and Psychological Measurement, 69, 913–928.

Love, N. B., Stiles, K., Mundry, S., & DiRanna, K. (2008). *The data coach's guide to improving learning for all students*. Thousand Oaks, CA: Corwin Press.

Mandinach, E. (2012). A perfect time for data use: Using data-driven decision making to inform practice. Educational Psychologist. 47. 71-85. 10.1080/00461520.2012.667064.

Marsh, J. A., & Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making: A framework for understanding capacity building. *Educational Management Administration & Leadership, 43*(2), 269-289. doi: 10.1177/1741143214537229

McCubbin, M.A. (1993). Family stress theory and the development of nursing knowledge about family adaptation. In S.L. Feetham, S.B. Meister, J.M.Bell, & C.L. Gillis (Eds.) The Nursing Family. New Bury Park: Sage Publications, 46-58.

Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). *Implementing data-informed decision making in schools: Teachers access, supports and use*. Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development.

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level from accountability to instructional improvement*. Washington, D.C.: U.S. Department of Education, Evaluation and Policy Development.

Monahan, S., S. Atkins-Burnett, B. A. Wasik, L. Akers, F. Hurwitz, and J. Carta. (2016) "Developing a Tool to Examine Teachers' Use of Ongoing Child Assessment to Individualize Instruction." Mathematica Policy Research Reports 92534e43eaeb4ec39398b9fc8, Mathematica Policy Research..

National Center on Program Management and Fiscal Operations, Center for (NCPMFO) (2013). *Data in Head Start and Early Head Start: Creating a culture that embraces data.* Washington, D.C.: Administration for Children and Families (ACF), Office of Head Start (OHS) Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/operations/data

NICHD Early Childcare Research Network. (2002). Early child care and children's development prior to school entry: Results from the NICHD study of early child care. *American Educational Research Journal, 39(1)*, 133-164.

Pierce, R. & Chick, H. (2012). *Improving teachers' statistical literacy. The 8th Australian Conference on Teaching Statistics OZCOTS 2012*.http://www.stat.auckland.ac.nz/~iase/publications.php

Penuel, W., Briggs, D., Davidson, K., Herlihy, C., Sherer, D., Hill, H., Farrell, C., Allen, A. (2017). How School and District Leaders Access, Perceive, and Use Research. *AERA Open, 3*(2), 23.

Penuel, W.A., Shepard, L.A., Hamilton, L.S., Miller, S.R., Gummer, E., & Mandinach, E.B. (2018). Data-driven decision making: Does it lack a theory of learning to inform research and practice. AERA Conference 2018, New York City, New York.

Raykov, T., & Marcoulides, G. (2011). *Introduction to psychometric theory Tenko Raykov, George Marcoulides.* New York: Routledge.

Reeves, T. D., Summers, K. H., & Grove, E. (2016). Examining the landscape of teacher learning for data use: The case of Illinois. *Cogent Education*, *3*(1), 1211476. https://doi.org/10.1080/2331186X.2016.1211476

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297. doi:10.1037/1040-3590.12.3.287

Ruel, E.E., Wagner, W.E., Gillespie, B.J. (2016). *The practice of survey research: Theory and applications.* Thousand Oaks, CA: Sage Publications, Inc.

Ryan, R. M., & Deci, E. L. (2003). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. In L. W. Porter, G. A. Bigley & R. M. Steers (Eds.), *Motivation and work behavior* (7th ed.). Boston: McGraw Hill.

Sandstrom, H., Isaacs, J. & Rohacek, M. (2015). How Head Start grantees set and use school readiness goals: Highlights from a research study, OPRE Report # 2015-12b, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test for normality. *Biometrika, 52,* 591-611.

Shepard, L., Penuel, W., & Davidson, K. (2017). Design principles for new systems of assessment. Phi Delta Kappan, 98(6), 47-52.

Spillane, J. P. (2012). Data in Practice: Conceptualizing the Data-Based Decision-Making Phenomena. *American Journal of Education, 118*(2), 113-141. doi: 10.1086/663283

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievment: Review of research. *Psychology in Schools, 42*(8), 795-819. doi:10.1002/pits.20113

Stevenson-Garcia J., Horm, D. Atkins-Burnett, S. (2019). Improving the accuracy and feasibility of infant and toddler assessment in applied settings. Child Care and Early Education Policy Research Consortium: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Sue, V.M., Ritter, L.A. (2012). *Conducting online surveys.* Los Angeles: SAGE.

Thousand, J., Villa, R., Nevin, A. (2015). Differentiating instruction: Planning for universal design, and teaching for college and career readiness / Jacqueline S. Thousand, Richard A. Villa, Ann I. Nevin. (Second ed.).

Tucker, L., & Lewis, R. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10.

Turner, R.C., & Carlson, L. (2003). Indexes of item-objective congruence for multidimensional items. *International Journal of Testing,* 3(2), 163-171. https://doi.org/10.1207/S15327574iJT0302_5

Vandell, D., Burchinal, M., Pierce, K., Eccles, Jacquelynne S., & Dubow, E. F. (2016). Early Child Care and Adolescent Functioning at the End of High School: Results From the NICHD Study of Early Child Care and Youth Development. *Developmental Psychology, 52*(10), 1634-1645.

Virmani, E.A., & Mangione, P. L. (Eds.). (2013). Infant/toddler caregiving: A guide to culturally sensitive care (2nd ed.) (pp. 72–75). Sacramento: California Department of Education.

Walker, D., Carta, J. J., Greenwood, C. R., & Buzhardt, J. F. (2008). The Use of Individual Growth and Developmental Indicators for Progress Monitoring and Intervention Decision Making in Early Education. *Exceptionality, 16*(1), 33-47. doi: 10.1080/09362830701796784

Wayman, J., Shaw, S., & Cho, V. (2017). Longitudinal effects of teacher use of a computer data system on student achievement. *AERA Open, 3*(1), 18.

Whitebook, McLean, Austin, & Edwards (2018). Early childhood workforce index. Center for the Study of Childcare Employment. Retrieved from: https://cscce.berkeley.edu/topic/early-childhood-workforce-index/

Yang, J., & Zheng, X. (2018). Item Response Data Analysis Using Stata Item Response Theory Package. *Journal of Educational and Behavioral Statistics, 43*(1), 116-129.

Zweig, J., Irwin, C. W., Kook, J. F., & Cox, J. (2015). *Data collection and use in early childhood education programs: Evidence from the Northeast Region* (REL 2015–084). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

APPENDICES

Appendix 1: Literature Review

https://drive.google.com/file/d/1iZ8Bi97BZChMSQ07Ecy5IR0NsT5UW1iF/view?usp=sharing

Appendix 2: Pre-testing Scripts

https://drive.google.com/file/d/1O6DrRqRwxTtN0LYw9xmmpGFO-5JIwGBz/view?usp=sharing

Appendix 3: DUCES 2.0

https://drive.google.com/file/d/1T9X_30kpbt2xz49_nkvt4Mc33neTEWn3/view?usp=sharing.

Appendix 4: DUCES 3.1 All Items

I have read the consent and agree to participate in this survey.

○ Yes (5)

○ No (6)

Skip To: End of Survey If I have read the consent and agree to participate in this survey. = 6

Page Break

For the purpose of this survey, we define "Data" as 'any information that is collected in a purposeful/systematic way.'

Data is...

○ information that is collected purposefully. (1)

Fantastic! You're ready to take our survey!

Page Break

Sometimes, you collect data (e.g. observing child behavior) and sometimes other people collect data (e.g. speech therapists assess a child's language).

This survey will ask you about Teacher-Collected data (that you or other teachers collect) and will repeat some of the same questions about data other people collect.

Page Break

Below are types of teacher-collected data.
In the last 12 months, which type of data did you or other teachers collect?

- ☐ Screening Tools (e.g. Ages and Stages Questionnaire, ESI-R, Denver-PDQ) (1)
- ☐ Informal Observations (e.g. child behavior, anecdotes of development, classroom activities) (2)
- ☐ School Records (e.g. attendance, daily sheets, parent communication) (3)
- ☐ Child Rating Tools (e.g. DECA) (4)
- ☐ Direct Child Assessments (e.g. IGDIs collected by teachers) (5)
- ☐ Child Observation Tools (e.g. DRDP, TS Gold) (6)
- ☐ Teacher-created formative assessments (e.g. exit tickets) (7)
- ☐ Classroom observations (e.g. observation of teachers' classroom practice) (8)
- ☐ Information about families (e.g. family needs, home culture, family preferences/interests) (9)
- ☐ None of these. (11)

Skip To: TypeOT If Below are types of teacher-collected data. In the last 12 months, which type of data did you or... = 11

CLASSROOM QUALITY DATA
(observations of teachers' classroom practice)

Please answer the following questions based on information about Classroom Quality **collected by you or other teachers.**
(e.g. classroom climate, instructional support, observations of teachers in the classroom)

In the last 12 months, how often did the following happen with teacher-collected Classroom Quality observation data?

How often did the following happen?	Not applicable/ Not in the last 12 months (1)	Once in last 12 months (2)	Once in last 6 months (3)	Every 2 to 3 months (4)	Once a month (5)	Two to three times a month (6)	Once a week (7)
I used data to identify ways to strengthen my teaching. (TCClass_1)	○	○	○	○	○	○	○
I used data to make plans (e.g.professional goals, work objectives). (TCClass_2)	○	○	○	○	○	○	○
I used data to change my activities with students. (TCClass_3)	○	○	○	○	○	○	○
I used data to create strategies that support children's learning/development. (TCClass_4)	○	○	○	○	○	○	○
I used data to evaluate the effectiveness of my plans. (TCClass_5)	○	○	○	○	○	○	○

CHILD ASSESSMENT DATA
(Data about children's learning and development)

Please answer the following questions based on Child Assessment data that teachers collect.

(e.g. ASQs, informal observations, daily sheets, DECA, direct child assessments, teacher-created assessments)

In the last 6 months, how often did the following happen with teacher-collected Child Assessment data?

How often did the following happen?	NA/Not in the last 6 months (1)	Once in the last 6 months (2)	Every 2 to 3 months (3)	Once a month (4)	Two to three times a month (5)	Once a week (6)	Daily (7)
I used data to evaluate progress toward student learning goals. (TCchild1)	○	○	○	○	○	○	○
I used data to evaluate the effectiveness of my instruction (e.g. lessons and/or projects). (TCchild2)	○	○	○	○	○	○	○
I used data to identify a group of children who weren't doing	○	○	○	○	○	○	○

well. (TCchild3)							
I used data to identify students' strengths and areas for improvement. (TCchild4)	○	○	○	○	○	○	○
I used data to identify students who needed more learning challenges. (TCchild5)	○	○	○	○	○	○	○
I used data to identify whether student learning needs were met using data. (TCchild6)	○	○	○	○	○	○	○
I used data to identify reasons a child struggled to learn. (TCchild7)	○	○	○	○	○	○	○

FAMILY DATA
(Information collected about families)

Please answer the following questions based on Family Data that teachers collect.
(e.g. family needs, home culture, family preferences/interests)

In the last 6 months, how often did the following happen with teacher-collected Family Data?

	Not applicable/ Not in the last 6 months (1)	Once in the last 6 months (2)	Every 2 to 3 months (3)	Once a month (4)	Two to four times a month (5)
I looked at data as a way to learn about the family. (TCFam_1)	○	○	○	○	○
I reflected on data as information about strengths/needs of families. (TCFam_2)	○	○	○	○	○
I used data to plan family engagement activities with other staff. (TCFam_3)	○	○	○	○	○
I used data to assess progress in the well-being of	○	○	○	○	○

families.
(TCFam_4)

The previous set of questions were about data collected by teachers.

We are going to repeat these questions now, but please answer them based on data collected by people other than teachers (e.g. researchers, therapists, evaluators, special services).

Below are types of data often collected by people other than teachers.

Which of these data did other people collect, but you had access to in last 12 months?

▢ Classroom Observations (e.g. ITERS, ECERS, CLASS, TPOT, Master Teacher Observations) (1)

▢ Family Data (e.g. parent surveys, enrollment records, family needs assessment) (2)

▢ Child Health Data (e.g. dental records, health needs) (3)

▢ Child Rating Tools (e.g. PSRA) (4)

▢ Direct Child Assessments (e.g. IGDIs/ECI, PLS, PPVT, Bracken) (5)

▢ Child Observation Tools (e.g. DRDP, TS Gold) (6)

▢ Developmental Records (e.g. IEP status, assessment records) (7)

▢ Other (8) ____

▢ None of these (9)

CLASSROOM QUALITY DATA
(observations of teachers' classroom practice)

Please answer the following questions based on data collected by others about Classroom Quality.
(e.g. ITERS, ECERS, CLASS)

In the last 12 months, how often did the following happen with Classroom Quality data collected by others?

How often did the following happen?	Not applicable/ Not in the last 12 months (1)	Once in the last 12 months (2)	Once in the last 6 months (3)	Every 2 to 3 months (4)	Once a month (5)	Two to three times a month (6)	Once a week (7)
I used data to identify ways to strengthen my teaching. (1)	○	○	○	○	○	○	○
I used data to make plans (e.g.professional goals, work objectives). (4)	○	○	○	○	○	○	○
I used data to change my activities with students. (5)	○	○	○	○	○	○	○
I used data to create strategies that support children's learning/development. (6)	○	○	○	○	○	○	○
I used data to evaluate the effectiveness of my plans. (7)	○	○	○	○	○	○	○

CHILD ASSESSMENT DATA
(Data about children's learning and development)

Please answer the following questions based on Child Assessment data collected by others.

(e.g. child health data, child rating tools, direct child assessments, child observation tools, developmental records)

In the last 6 months, how often did the following happen with Child Assessment data collected by others?

How often did the following happen?	Not applicable/ Not in the last 6 months (1)	Once in the last 6 months (2)	Every 2 to 3 months (3)	Once a month (4)	Two to three times a month (5)	Once a week (6)	Daily (7)
I used data to evaluate progress toward student learning goals. (1)	○	○	○	○	○	○	○
I used data to evaluate the effectiveness of my instruction (e.g. lessons and/or projects). (4)	○	○	○	○	○	○	○
I used data to identify a group of children who weren't doing well. (5)	○	○	○	○	○	○	○

I used data to identify students' strengths and areas for improvement. (6)	○	○	○	○	○	○	○
I used data to identify students who needed more learning challenges. (7)	○	○	○	○	○	○	○
I used data to identify whether student learning needs were met using data. (8)	○	○	○	○	○	○	○
I used data to determine children's knowledge or skills before teaching. (9)	○	○	○	○	○	○	○
I formed small groups based on data about children's learning and development. (10)	○	○	○	○	○	○	○

I used data to identify reasons a child struggled to learn. (11)	○	○	○	○	○	○	○

FAMILY DATA
(Information collected about families)

Please answer the following questions based on Family Data collected by others.
(e.g. parent surveys, enrollment records, family needs assessment)

In the last 6 months, how often did the following happen with data collected by others about families?

How often did the following happen?	Not applicable/ Not in the last 6 months (1)	Once in the last 6 months (2)	Every 2 to 3 months (3)	Once a month (4)	Two to four times a month (5)
I looked at data as a way to learn about the family. (OTFam_1)	○	○	○	○	○
I reflected on data as information about the strengths/needs of families. (OTFam_4)	○	○	○	○	○
I noticed that family data explained changes (positive or negative) in a child. (OTFam_5)	○	○	○	○	○
I used data to plan family engagement activities with	○	○	○	○	○

other staff. (OTFam_6)					
I used data to assess progress in the well-being of families. (OTFam_7)	○	○	○	○	○

How effective were these resources?	Not at all effective/Did not have (1)	Slightly effective (2)	Moderately effective (3)	Very effective (4)	Extremely effective (5)
Leaders who facilitated understanding of data. (tools1)	○	○	○	○	○
Interdisciplinary teams who brought multiple perspectives to interpret data. (tools2)	○	○	○	○	○
Access to staff who gave advice on how to apply data to work. (tools3)	○	○	○	○	○
Access to technology that generated reports about data. (tools4)	○	○	○	○	○

Conversations that guided you to reflect on data. (tools5)	○	○	○	○	○
Tools that helped focus the interpretation and use of data. (e.g. paper forms, reflection guides) (tools6)	○	○	○	○	○
Procedures that helped prioritize teaching practice using data. (tools7)	○	○	○	○	○
Processes to create strategies that address concerns revealed by data. (tools8)	○	○	○	○	○

How effective were these resources?	Not at all effective/Did not have (1)	Slightly effective (2)	Moderately effective (3)	Very effective (4)	Extremely effective (5)
Training on using a cycle of inquiry (e.g. prepare data, interpret meaning, implement plans, evaluate plans). (time1)	○	○	○	○	○
Protected time (e.g. plan time, meetings) to <u>evaluate progress</u> on data-informed plans. (time2)	○	○	○	○	○
Protected time (e.g. plan time, meetings) <u>to set goals</u> with data. (time3)	○	○	○	○	○
Necessary training on interpreting data appropriately. (time4)	○	○	○	○	○
Timely access to data. (e.g. results available when need, quick to get a score) (time5)	○	○	○	○	○

To what extent do you agree?	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
I have people at work who would help me if I did not understand the data. (supp1)	○	○	○	○	○	○	○
My co-workers would be supportive if I made a mistake when interpreting data. (supp2)	○	○	○	○	○	○	○
I have opportunities to talk with people I trust about data. (supp3)	○	○	○	○	○	○	○
My co-workers help me use my skills and knowledge to make meaning	○	○	○	○	○	○	○

from data. (supp4)							
We work as a team to understand how data connects to classroom practice. (supp5)	○	○	○	○	○	○	○
My ideas about applying data are respected by my colleagues. (supp6)	○	○	○	○	○	○	○
We would share new strategies if data indicated a need. (supp7)	○	○	○	○	○	○	○
I have received support to develop data use skills. (supp8)	○	○	○	○	○	○	○

To what extent do you agree?	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
Data can help inform interactions and instruction. (att1)							
Data can help me identify my own strengths and weaknesses. (att2)							
Data can offer valuable insight into work with children. (att3)							
Teachers can make better decisions with data than without it. (att4)							
Data can make my job easier. (att5)							
Data can motivate continuous improvement. (att6)							
Using data is an important part of teaching. (att7)							

How confident are you in your ability to…	Not at all confident (1)	Slightly confident (2)	Moderately confident (3)	Very confident (4)	Extremely confident (5)
Use data to decide whether new strategies are needed. (Conf1)	○	○	○	○	○
Use data to design lesson plans for the whole class. (Conf2)	○	○	○	○	○
Use data to help parents understand how their child is doing. (Conf3)	○	○	○	○	○
Use data to create a plan for an individual student. (Conf4)	○	○	○	○	○
Use data to generate data-informed plans. (Conf5)	○	○	○	○	○
Participate in a cycle of inquiry (i.e. prepare data, interpret data, implement data-informed plans, evaluate with data). (Conf6)	○	○	○	○	○
Use data to understand whether children are progressing. (Conf7)	○	○	○	○	○
Apply your early childhood expertise to create plans with data. (Conf8)	○	○	○	○	○

To what extent do you agree?	NA/ Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
I worried that data would show that my work was less successful than my colleagues. (motiv1)	○	○	○	○	○	○	○
I saw data represented in ways that could create competition among co-workers. (motiv2)	○	○	○	○	○	○	○
I wished I could do things differently in my classroom, but I feel pressure based on data. (motiv3)	○	○	○	○	○	○	○
I worried that I would lose my job if my scores were low. (motiv4)	○	○	○	○	○	○	○
I felt pride or shame when my scores were compared to others. (motiv5)	○	○	○	○	○	○	○
I enjoyed using data to inform my classroom practice. (motiv6)	○	○	○	○	○	○	○

I felt that using data was valuable/beneficial. (motiv7)

People I know well (e.g. work friends, close colleagues) recommended using data to inform my classroom practice. (motiv8)

I thought it was fun to use data to inform my classroom practice. (motiv9)

I thought it was important to make the effort to use data. (motiv10)

I felt ashamed because I did not want to use data. (motiv11)

I felt under pressure to use data to inform my classroom practice. (motiv12)

I felt guilty because I did not want to use data. (motiv13)

I felt pressure to have data showing I am doing my job. (motiv14)

I felt pulled in multiple directions by data. (motiv15)	○	○	○	○	○	○	○
I felt that data was used to help me become more self-aware in my practice. (motiv16)	○	○	○	○	○	○	○

Appendix 5: Insider Rating Questionnaire

How likely is this teacher to...?`

	Extremely likely	Somewhat likely	Neither likely nor unlikely	Somewhat unlikely	Extremely unlikely
Try something new in the classroom because of data?	○	○	○	○	○
Share data about families with colleagues?	○	○	○	○	○
Use data to learn more about a child?	○	○	○	○	○
Change how they interact with certain children based on data?	○	○	○	○	○
Use data to identify areas of classroom practice they want to improve?	○	○	○	○	○
Consider whether a child needs new challenges based on data?	○	○	○	○	○
Use data to learn whether their classroom was considered to be high quality?	○	○	○	○	○
Seek to learn about families based on data from surveys or interviews?	○	○	○	○	○
Engage with families based on needs identified through surveys or interviews?	○	○	○	○	○
Give additional supports to a child based on data?	○	○	○	○	○

Appendix 6: R Code

```
library(mirt)

library(readxl)
```

Model fit

```
#TCclass_3c

TCclass <- read_excel("Dissertation/TCclass.xlsx")
View(TCclass)

TCclassMod <- mirt(TCclass, 1, itemtype = "gpcm", method = "EM", optimizer = "NR", calcNull
= TRUE, quadpts = 7, verbose = TRUE)

M2(TCclassMod, type = 'C2')

#TCchild

TCchild_3c <- read_excel("Dissertation/TCchild_3c.xlsx")
View(TCchild_3c)

TCchildMod <- mirt(TCchild_3c, 1, itemtype = "gpcm", method = "EM", optimizer = "NR",
calcNull = TRUE, quadpts = 7, verbose = TRUE)

M2(TCchildMod, type = 'C2')
```

```
#TCfam

TCfam_4c <- read_excel("Dissertation/TCfam_4c.xlsx")

View(TCfam_4c)

TCfamMod <- mirt(TCfam_4c, 1, itemtype = "gpcm", method = "EM", optimizer = "NR",
calcNull = TRUE, quadpts = 7, verbose = TRUE)

M2(TCfamMod, type = 'C2')

#OCclass

OCclass_4c <- read_excel("Dissertation/OCclass_4c.xlsx")

> View(OCclass_4c)

OCclassMod <- mirt(OCclass_4c, 1, itemtype = "gpcm", method = "EM", optimizer = "NR",
calcNull = TRUE, quadpts = 7, verbose = TRUE)

M2(OCclassMod, type = 'C2')
```

```
#OCchild
```

```
OCchild_5c <- read_excel("Dissertation/OCchild_5c.xlsx")
View(OCchild_5c)
```

```
OCchildMod <- mirt(OCchild_5c, 1, itemtype = "gpcm", method = "EM", optimizer = "NR",
calcNull = TRUE, quadpts = 7, verbose = TRUE)
```

```
M2(OCchildMod, type = 'C2')
```

```
#OCfam
```

```
OCfam_4c <- read_excel("Dissertation/OCfam_4c.xlsx")
> View(OCfam_4c)
```

```
OCfamMod <- mirt(OCfam_4c, 1, itemtype = "gpcm", method = "EM", optimizer = "NR",
calcNull = TRUE, quadpts = 7, verbose = TRUE, SE = TRUE)
```

```
M2(OCfamMod, type = 'C2')
```

Item fit

```
itemfit(TCclassMod)
```

```
itemfit(TCchildMod)

itemfit(TCfamMod)

itemfit(OCclassMod)

itemfit(OCchildMod)

itemfit(OCfamMod)
```

Item parameters

```
TCclassItems <-coef(TCclassMod, IRTpars=TRUE, as.data.frame=TRUE)

TCchilditems <-coef(TCchildMod, IRTpars=TRUE, as.data.frame=TRUE)

TCfamitems <-coef(TCfamMod, IRTpars=TRUE, as.data.frame=TRUE)

OCclassItems <-coef(OCclassMod, IRTpars=TRUE, as.data.frame=TRUE)

OCchilditems <-coef(OCchildMod, IRTpars=TRUE, as.data.frame=TRUE)

OCfamItems <-coef(OCfamMod, IRTpars=TRUE, as.data.frame=TRUE)

TCclassItems
```

TCchilditems

TCfamitems

OCclassItems

OCchilditems

OCfamItems

VITA

Shannon Hazel Stark Guss

Candidate for the Degree of

Doctor of Philosophy

Dissertation: REFINEMENT OF A MEASURE OF DATA USE PRACTICES

Major Field: Educational Psychology: option REMS

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Educational Psychology: Research, Evaluation, Measurement, and Statistics at Oklahoma State University, Stillwater, Oklahoma in December, 2019.

Completed the requirements for the Master of Arts in Liberal Studies: Administrative Leadership at University of Oklahoma, Norman, OK in 2012.

Completed the requirements for the Bachelor of Science in Sociology at Langston University, Langston, OK in 2003.

Experience:

Early Childhood Education Institute, University of Oklahoma-Tulsa
Project Director: (6/2011 to present) Data analysis, presentation, supervision.
Site Coordinator: (8/2010 to 6/2011) Data sharing, IRB, engagement.
Research Associate: (9/2008 to 8/2010) Data collection & cleaning

Professional Memberships:
American Educational Research Association 2017-18
Society for Research in Child Development 2016-17
American Evaluation Association 2012-2016