January 2019

Reliability And Validity Of An Evaluation Capacity Assessment Instrument For Principals Of Public Schools In Michigan And The Current Distribution Of Evaluation Capacity By Poverty Concentration

Christine M. Lewis
Wayne State University

RELIABILITY AND VALIDITY OF AN EVALUATION CAPACITY ASSESSMENT INSTRUMENT FOR PRINCIPALS OF PUBLIC SCHOOLS IN MICHIGAN AND THE CURRENT DISTRIBUTION OF EVALUATION CAPACITY BY POVERTY CONCENTRATION

by

CHRISTINE LEWIS

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: EDUCATION EVALUATION
AND RESEARCH

Approved By:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

For my husband, who promised I would be able to “finish school” and followed-through.

Thank you.

For my children, who remind me of the importance of every moment and the promise of every child.

Thank you.

For my grandmother and mother, who always sacrificed what they needed to and did what they had to do for the education and well-being of the family.

Thank you.

For all teachers, principals and every one of my colleagues in education, who accept the challenge and responsibility to positively impact ***all*** students each day.

Thank you.

ACKNOWLEDGEMENTS

First, I must thank Dr. Shlomo Sawilowsky, my major advisor. I am very fortunate to have had his support and guidance throughout my doctoral experience. Of many, many lessons, I will always remember one of my earliest, when it comes to statistics and measurement, the questions matter, as Lord (1953) pointed out. Though more of a statistics and measurement lesson, it has evolved to become more of a beacon as I move through my work.

I must also thank Dr. Carolyn Shields. I am also very fortunate to have had her challenging me to ask the questions that matter. The questions that matter can be used to move people. The right questions at the right times can deconstruct and reconstruct knowledge frameworks that stand in the way of social justice, that stand in the way of the promise of education to empower and elevate.

I am also very thankful for those individuals who really helped me navigate through the weeds, Dr. Elizabeth McQuillen and Dr. Jennifer Heisler. They were more than willing to get in there with me and point the way out. An additional debt of gratitude I owe to Dr. Whitney Moore who Zoom©ed right in with additional guidance just when I needed it most.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Today's educational landscape is rife with data. A climate of mandated testing and increased emphasis on accountability has left educators with an abundance of data and a multitude of accountability activities. Ideally, educators should be using data to the benefit of students, like everything else they do. They should be using data to continually ask and answer the question, "Did our program/strategy/initiative achieve the results we wanted for our students?" This is an evaluative question and if school communities are not continually focused on answering this question, they do not know if what they have been doing, and continue to do, is benefitting students. Without asking and answering those kinds of questions, tomorrow's educational work is most likely the same work as the day before, and so will be the results. Engaging in this questioning or engaging in evaluation activities to determine whether instructional practices and programs are having a positive impact on student outcomes is a critical school community practice, especially for communities where many students live in poverty. Having the capacity to evaluate is a prerequisite to engaging in these evaluation activities to plan for and verify improved outcomes for students, especially for students in poverty. Do school communities have the capacity to evaluate the impact of their programs on student outcomes? Having an assessment to determine whether they do, or at what level they do, is necessary, and has yet to be established. Building on previous evaluation literature, this study is an investigation of a previously designed assessment used in service organizations as a potential assessment instrument for evaluation capacity of school leaders and

communities. It is also an exploration into whether evaluation capacity is related to poverty.

Background

The passage of No Child Left Behind (NCLB) in 2001 impacted state accountability systems and focused the attention of educators on the use of multiple sources of data. Under the *Government Performance and Results Act* of 1993, government agencies develop annual performance plans and reports. The passage of NCLB in 2001 became part of a movement towards greater accountability in government overall and greater responsibility for results (General Accounting Office [GAO], 2002). The federal government began requiring states and school districts to test more, report more and to apply sanctions more quickly to schools not meeting improvement goals (Goertz & Duffy, 2003). For schools, congress mandated new data reporting requirements in two major areas: student academic performance and staff qualifications (Yeagley, 2003). Within this accountability movement, state, district, school, and classroom assessments led to more data than school officials can reasonably manage (Huffman et al., 2006). Heard in educational discussions since then is the notion educators are data rich but information poor (Hall, 2004; Mills, 2006). This concept has an acronym, DRIP (data rich information poor). It reflects a distinction between the terms data and information and highlights the notion that educators lack information.

What is meant by phrase *data rich and information poor*? What is the difference between the terms data and information? The difference can be examined in the context of a hierarchy outlined by Ackoff (1999) connecting the words data, information, knowledge and wisdom (DIKW). Data were defined as raw components of information

without significance; information as data that has meaning by relational connections; knowledge as a collection of information that is useful; and wisdom as the process by which judgments are made between good and bad alternatives. Educators, being data rich and information poor under the DIKW hierarchy, have data but lack the relational connections transforming data into information and by virtue of the hierarchy, lack knowledge and wisdom, borne of the data. Within the field of education, a raw data list of high school graduates and non-graduates and associated variables can provide educators with data in order to gain information from relational connections between the variables (numbers or percent of graduates, minority graduates, female non-graduates, etc.) which can add to the educator's knowledge surrounding graduation and can be applied as educators make judgments between good and bad alternatives (ascribe value to alternatives) and improve retention programs.

The process of evaluation is not simply wisdom in action. Rather, evaluation is defined in different ways. According to King and Volkov (2005), program evaluation is "a systematic, data-based process for judging the value of a program" (p. 10). Stufflebeam and Shinkfield (2007) defined "evaluation is the systematic process of delineating, obtaining, reporting, and applying descriptive and judgmental information about some object's merit, worth, probity, feasibility, safety, significance, and/or equity" (p.16). Fitzpatrick, Sanders, and Worthen (2011) offered another definition of evaluation, "the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit) in relation to those criteria", (p. 7). Each of these definitions included the root word of evaluation, which is value and identified evaluation as a process.

The process of evaluation is then linked to wisdom following Ackoff's (1999) statement that wisdom is the process by which judgments are made between good and bad alternatives. Judgments are made between that which has more or less value (good or bad alternatives, good being imbued with greater value) in a given context. Wisdom then is related to value judging, to evaluation.

Within the evaluation process, the results are intended to be used in some manner. For evaluation to be worth the resources the evaluation process itself consumes, it must leave decision makers with an ability to improve what was being evaluated (Guerra-Lopez, 2007). In an educational context, the results could be used to inform changes in curriculum, instruction or supplemental programs. After the connections between data, information, and knowledge are made, and, after value is assigned, productive action then has a chance of being taken. According to Hanwright and Makinson (2008), the process of evaluation contributes to both working knowledge and organizational action. Once judgments are made between good and bad alternatives, questions about how to improve outcomes arise. Under such paradigm, data, leads to information, leads to knowledge, leads to value assignment (wisdom) and, finally, leads to productive action. Productive action is then rooted in data under such a paradigm.

Evaluation Capacity

Currently in the State of Michigan, members of school communities are involved in an evaluation process which connects data, information, knowledge, value assignment and action to improve outcomes for students. This involvement comes through the members of a school community completing the Michigan Department of Education's (MDE) Program Evaluation Tool (PET) (see Appendix A). This tool is required by the MDE

to be completed annually by all public schools in the state. The introduction of the PET marked the first time a prescribed report regarding the evaluation of programs was directly mandated for all public schools to complete, placing program evaluation, in a specific format, front and center as a State reporting requirement. Though legislation had contained the requirement to evaluate programs prior to the introduction of the PET, school communities were free to define and engage in evaluation activities as best they could, and according to whatever capacity they had at the time. The State set an evaluation bar for school communities to clear, no matter how close or how far they were from this bar prior to the PET.

This evaluation process is intended to produce improved outcomes for students. The MDE PET states evaluation is good practice and will likely improve outcomes (Appendix A). Owen (2007) indicated, “evaluation can and should enhance the quality of interventions (policies and programs) designed to solve or ameliorate problems in social and corporate settings”, (p. 1). Additionally, Owen (2007) indicated “Evaluation should be seen as a process of knowledge production which rests on the use of rigorous empirical inquiry,” (p. 1). When applied to programs specifically funded and designed to improve the achievement of the disadvantaged, or students of poverty, outcomes for the disadvantaged should be improved through robust evaluations, and organizations should become more knowledgeable regarding problems that impede the success of students. The goal of the evaluation process is to plan for, and subsequently be able to take productive action to improve programs overall but, particularly programs undertaken to realize the goal of improving the achievement of the disadvantaged.

The construct evaluation capacity is directly related to increasing student achievement. According to Boyle and Lemaire (1999), evaluation capacity refers to “the human capital (skills, knowledge, experience, etc.) and financial/material resources” (p. 5). It is also conceptualized as an organization’s ability to design, implement and manage effective evaluations; access, build and use evaluative knowledge and skills; and create awareness and support of evaluation as an improvement strategy (King & Volkov, 2005). According to Mulford and Silins (2011), “Principals who promote both capacity building and systems of accountability and evaluation, to the extent that their teachers perceive these two factors as characterising their schools, advance student empowerment, social development and academic achievement”, (p. 61). Therefore, students in those schools are more empowered, socially developed and have greater academic achievement than students in schools where principals do not promote capacity building and systems of accountability and evaluation. This advantages students attending those schools over those students who do not attend schools with systems of accountability and evaluation promoted by the principal. Evaluation capacity could have a direct impact on student achievement.

There are many unanswered questions related to the MDE PET because it was implemented statewide just several years ago and because of the required focus it placed on evaluation activities in schools. A key question is whether school communities have the capacity to evaluate programs? Do they have what they need to engage in the DIKW hierarchy at all levels and take productive action for improving student outcomes, particularly outcomes for disadvantaged students? Do some school communities have greater capacity than others? Does this advantage some students over others?

Preceding the question of whether local school communities have the capacity to improve student outcomes by engaging in robust evaluation and all questions following that, are two critical questions (1) if and (2) how evaluation capacity can be assessed in local school settings? Without answers to those questions, there cannot be an answer to the question of whether schools have the capacity to evaluate and to use that evaluation to improve student outcomes, particularly for disadvantaged students.

Constructs, or ideas not directly measurable against a clear metric such as a yardstick or liter, have generally been problematic for educators to measure (Shadish, Cook & Campbell, 2002). Evaluation capacity is a construct comprised of multiple factors and is of growing interest in the field of education. There has been a relatively recent increase in published literature within the educational field as well as the requirement of program evaluation included in NCLB (2001).

A ProQuest® Multi-Search conducted in June 2016 of peer reviewed publications containing the exact phrase “evaluation capacity” returned 196 entries beginning in 1975. Narrowing the search to focus only on evaluation capacity as it pertains to education both pares down the number of total results as well as publication timeframes. Adding the exact phrase “Schools” to “Evaluation Capacity” returned 24 results, beginning in 2001, the year NCLB was signed into law. Altering the search to combine “Evaluation Capacity” with “K-12” produced a list of 5 results from 2001 to the search date and searching “Evaluation Capacity” and “Measurement” returned 11 results beginning in 2008. Thus, evaluation capacity as it pertains to public schools is not as researched as evaluation capacity itself as evidenced by fewer entries returned when school related terms were added to the evaluation capacity search. It is also a more recent concept in public school

research as evidenced by the 26-year gap between results returned for a search of evaluation capacity alone and results returned by the addition of school related terms to the search.

Statement of the Problem

This study is being conducted to investigate the reliability and validity of an evaluation capacity assessment instrument for school principals, those responsible for leading a school community through the program evaluation process. This will allow examination of the key questions (1) if and (2) how evaluation capacity can be assessed in local school settings. The capacity of principals to systematize and use evaluation as an improvement strategy is similar in nature to the capacity of program staff in the community service organization (CSO) setting. Within each setting, there is a host of individual and organizational factors impacting the evaluation capacity of individuals responsible for conducting evaluation activities. Synthesizing these individual and organizational factors into a single model,

Taylor-Ritzler et al. (2013) found support for the use of the Evaluation Capacity Assessment Instrument (ECAI) as a generalizable tool for assessing the evaluation capacity of staff members in the CSO setting. The ECAI includes awareness of the benefits of evaluation, motivation to conduct evaluation, competence (knowledge and skills), leadership, climate and resources (Taylor-Ritzler et al., 2013). These factors are not confined to a CSO. They are also factors which could help or hinder program evaluation efforts in local schools. There is insufficient data to determine if the ECAI is a reliable instrument and valid for assessing the evaluation capacity of K-12 school principals. The first purpose therefore, is to establish reliability and fit evidence for the

model outlined within the ECAI developed by Taylor-Ritzler et al. (2013) for program evaluators in the CSO setting for assessing principal evaluation capacity in the public K-12 setting.

A second purpose of this study is to determine if differences in capacity are connected to the poverty level of students in a school. On the surface, communities of poverty are often seen as having less of just about everything, a focus on deficits. A natural extension of a deficit focus to evaluation capacity might lead some to the conclusion that poverty matters when examining evaluation capacity. That notion deserves investigating as varying levels of capacity will likely impact the process and product of evaluation. With numerous components comprising evaluation, any one of which can be changed by "having less" of it, evaluation capacity has the potential to be differentially distributed among local school communities. This distribution, in turn, has the potential to vary in such a way that the qualities of conclusions drawn and actions planned within the evaluation process itself have the potential to be distinctly different in some schools. The distribution of assessed evaluation capacity among school principals will be examined to determine if principals of schools with higher poverty rates are overrepresented at the lower end of ranked capacity assessment scores.

Synthesizing each outlined purpose to distinct research questions, the two research questions addressed in this study are as follows:

1. *Can reliability and fit evidence be established for the ECAI to be used to assess evaluation capacity of K-12 principals?*
2. *If so, is evaluation capacity differentially distributed between high and low poverty schools?*

Importance of the Study

Should results indicate the ECAI is a generalizable tool for assessing evaluation capacity of K-12 principals, it could be used to identify areas of focus for evaluation capacity building (ECB) efforts, assess progress along the way and allow school communities to implement robust systems of evaluation that improve student outcomes overall and specifically outcomes for disadvantaged students. Beyond that, the critical question regarding poverty and program evaluation can be explored in order to determine whether the process of evaluation in some settings has the potential to further disadvantage students of poverty that the process and product of evaluation was intended to benefit? If that is a finding, there is a problem that requires solving and once again the ECAI could be used to identify areas of focus for ECB efforts, assess progress during the ECB process and develop robust evaluation systems within school communities, specifically those with high numbers of disadvantaged students.

Title I is a federally authorized grant funding source provided to some schools and is intended to provide supplemental educational programs to improve the achievement of the disadvantaged. As Title I was re-authorized under the banner, “improving the academic achievement of the disadvantaged” (No Child Left Behind Act of 2001, Sec. 101), the program evaluation activities required to be performed under the act should improve achievement, particularly for the most disadvantaged. The act’s funding is allocated to local school districts by formula which includes available funding and geographic census poverty counts. Local school districts receive a funding allocation from the State and allocate to their Title I schools by free and reduced lunch rates. That rate is

used as the local identifier of poverty concentrations by school attendance areas across the district.

Title I is intended to raise student achievement, particularly for students of poverty, improving the achievement of the disadvantaged. Disadvantaged students can be further disadvantaged if evaluation systems within their schools required under the act are not as strong or developed as others. The evaluation capacity of the school community, centered within the role of the principal, has a direct impact on the strength and development of the evaluation system and on the subsequent achievement of all students, most importantly, students of poverty.

Limitations

Evaluation capacity, as a multi factored construct, including individual and organizational factors, poses a unique challenge in terms of assessment. Including individual and organizational factors, it is difficult to separate the evaluation capacity of the leader from the evaluation capacity of the community. By virtue of the ability to assign and leverage resources (human, fiscal, material) and shape organizational culture and climate, the evaluation capacity of the principal and the capacity of the school community are inherently linked and problematic to separate. Evaluation capacity is contextual in terms of the community. The ECAI is designed for a single respondent, the principal who currently leads within a given community context, but also reflects organizational factors (some within and others outside of a principal's immediate sphere of control). It is the larger evaluation capacity of the community which is assessed through the principal respondent at that point in time.

This study is based on self-reported data of randomly selected principals of K-12 schools responding to a survey. The results of research which requires survey submission after a random sample of participants is generated can be impacted by non-response of those selected, the characteristics of the participants who choose, or do not choose to participate. In a comparison of sample respondents during the initial phase of a survey, and those who participated in the second, responsive design, phase, Axxin et al. (2011) found the two groups of respondents differed in terms of age, job status, education and racial/ethnic group membership; and coefficients estimated from the two groups were significantly different.

Functional Definitions of Key Terms

High Poverty School: schools with free and reduced lunch participation percentages between 70 and 100, inclusive.

Low Poverty School: schools with free and reduced lunch participation percentages between 0 and 30 percent, inclusive.

Principal: the highest-ranking leader of a school.

Public Schools: schools receiving a state per-pupil foundation allowance rather than charge students themselves (including charter schools).

Title I-Participating School: a school receiving an allocation of Title I program funds and operating either a Title I targeted or school-wide program.

CHAPTER 2

Evaluation Literature and Education

In a climate emphasizing data and accountability for educational results without reassurance school communities have the prerequisite capacity to evaluate, the literature addressing reasons to evaluate, approaches to evaluation and individual roles within conducting evaluations undergirds the construct of evaluation capacity. Beyond those elements is situated the notion of how to build evaluation capacity, as necessary, to conduct evaluations and ascertain verifiable improvement in student outcomes.

Reasons to Evaluate

Owen (2007) indicated evaluation can enhance the quality of programs designed to solve problems in social settings. The results of robust evaluations in schools can be used to improve outcomes for students. Developing the capacity of school communities to carry-out such evaluations increases the likelihood school programs and processes will improve outcomes for students. Improving outcomes for students not only benefits the student but benefits society. Labaree (1997) outlined three alternative goals of education: democratic equality (preparation of citizens); social efficiency (training workers); and social mobility (preparation for competition for social position). The first two reflect the notion education is a public good (social betterment), the last, education is a private good (individual betterment). Better preparing students to become citizens and workers to fulfill their roles in a democracy and an economy supports the health of each and benefits all members of the democracy and economy, contributing to the public good. Fierro (2012) outlined a linear model connecting Evaluation Capacity Building (ECB), Evaluation Capacity, Organizational Evaluation Practice, Improved Educational Programming and

Social Betterment. Through an educational lens, the model can be adjusted as depicted in figure 1.

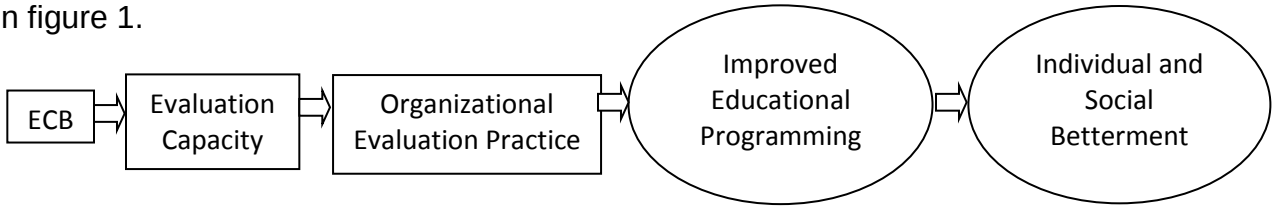

Figure 1*. Linear depiction of the outcomes of ECB.*

The ECAI will be an assessment to establish baseline evaluation capacity and to use to target professional learning designed to build evaluation capacity of staff in local schools. This will improve the evaluation practice of the organization which will allow the school community to improve the educational programming it provides. The ECAI will also be a tool to evaluate capacity building activities themselves. Ultimately, the tool is connected to providing all students a school community capable of using the process of evaluation to advance student learning and development in order to fulfill their roles as a successful citizen and contributor to the economy.

Despite these more improvement-oriented reasons for engaging in the process of program evaluation, it remains a compliance activity as well under federal legislation. Janssens and van Amelsvoort (2008) stated that accountability demands imposed on school self-evaluations generate an accountability-oriented evaluation, while improvement demands generate improvement-oriented evaluations. Ebrahim (2005) argued organizations involved in activities designed to alleviate complex social problems can be hindered by too much accountability. Ebrahim (2005) further contended that mechanisms for holding organizations accountable to funders can marginalize other mechanisms holding organizations accountable to the communities they serve and to pursuing their own missions. This is referenced as a type of myopia in which the attention

of the evaluators is highly focused on satisfying funder demands to the point of eclipsing attention to the mission of the organization and the responsibility of the organization to the stakeholders the program was designed to benefit. With competing reasons to engage in evaluation, accountability/compliance vs. mission/improvement, both the process and product of program evaluation can be shaped differently. The reasons for the individual(s) completing the evaluation determine how the results of the evaluation will be used and, ultimately, the usefulness of the evaluation towards improving outcomes for the individuals the program was intended to benefit.

Evaluation Approaches

In a wider perspective, evaluation is a family of research methods. Kahan (2008) stated most evaluations rarely include only one approach. Aspects of one approach can be combined with others. A mix of methods can be applied to evaluate a given program as the questions prompt the choice of methods (Owen, 2007). Given a variety of approaches, evaluations begin with an intentional plan. It is difficult to determine the value of a given program if the intended outcome is undefined, the implementation attributes are undefined and the mechanisms for data collection throughout are undefined. An evaluation plan is a written document that describes how the program is to be monitored and ultimately evaluated. It clarifies the “what,” the “how,” and the “why it matters” (Lavinghouse & Snyder, 2013). The “what” describes the program, the “how” defines the process for implementation with fidelity and the “why” provides the rationale for the program. All of these components can be viewed within the MDE PET (Appendix A).

According to the glossary published by the GAO (2011), four types of program evaluation are identified: process (or implementation) evaluation; outcome evaluation; impact evaluation and cost-benefit/cost-effectiveness evaluation. Employing process evaluation, evaluators assess the degree to which a program is operating as the creators intended. Conducting an outcome evaluation, the extent to which the program's objectives were achieved is examined. Impact evaluations result in the assessment of the net effect of the program, after estimating what would have happened in the absence of the program and comparing the two outcomes. A cost-benefit or cost-effectiveness evaluation can be conducted to determine if the benefit achieved was worth the cost of the program. Except for the question of comparisons to outcomes which would have otherwise occurred (impact evaluation), these approach categories of evaluation appear within the MDE PET (Appendix A).

Kahan (2008) classified a variety of evaluation models, primarily based on how the evaluation was conducted. Some of the classifications and models Kahan (2008) outlined were: results focused (goals based, goals free, theory based); collaborative; external; appreciative inquiry; context, input, process, product (CIPP); and utilization focused. Results focused evaluations can be sub-classified into three types. A goal-based evaluation identifies whether the objectives of the program were met. A goal free evaluation identifies all results of the program (anticipated or not). Kahan (2008) stated it is fairly common to combine these two approaches. While determining whether a goal was attained, investigations of unintended consequences can also be examined. This can be seen in the MDE PET (Appendix A). The third results focused evaluation model is theory-based evaluation (often referred to as a logic model). A theory-based evaluation

is conducted by defining the logical relationships between all the parts of the program. Actions defined within the implementation of the program theoretically lead to a given outcome. Action 1, predicted outcome 1, action 2, predicted outcome 2, etc. culminating in the last action and successful prediction of the final outcome, provide support for the theory of change outlined in the logic model and is used as evidence that the outcome was achieved based on the *a priori* theory.

A collaborative evaluation includes various stakeholders engaged in the process of evaluation. Kahan (2008) stated many types of evaluations are classified as collaborative based on the overview provided by Butterfoss, Francisco and Capwell (2001). Practical participatory evaluation is one type in which the evaluation is oriented toward decision making. Stakeholder participation is designed to enhance relevance, ownership and utilization of results. Though not specifically defined in the MDE PET (Appendix A), the involvement of a variety of stakeholders in the evaluation of Title I programs is a requirement. Additional collaborative models defined by Butterfoss et al. (2001) and referenced by Kahan (2008) include: transformative participatory evaluation; stakeholder-based evaluation; democratic evaluation; and action research. Transformative participatory evaluation aims to empower the individuals engaged in the process by including their awareness of their reality (based on Friere's concept of conscientization) and acknowledging the connection between knowledge, power and control. Learning, or evaluating, depends on uncovering problems and needs.

Stakeholder based evaluations are similar to practical participatory approaches but are more often used when groups may not generally agree on goals and require more guidance from an external evaluator (evaluations solely conducted by persons or

organizations outside of the program being evaluated are called external evaluations). The term stakeholder based does not equate to entirely stakeholder conducted. Depending on the level of control between an external evaluator and the internal stakeholders, some types of evaluations, those balanced between external evaluators and internal stakeholders, can be classified as democratic. Stakeholders conducting action research participate in the evaluation as both participants of the program and researchers investigating the outcomes of the program.

Although some methods of evaluation are used to determine the value of a program undertaken to solve some problem, Kahan (2008) noted the appreciative inquiry approach begins with evaluators carefully examining what is working well within organizations or programs. Instead of possible causes to a problem, and subsequent possible solutions, being the basis for the evaluation, the question of "what if the best that occurred" in one situation occurred more frequently, across other contexts and programs? Programs where there is already a substantiated level of success are examined to create replication plans across other programs. Participants engage in dialogue regarding needed tasks and resources to bring about the change in the given program based on what was learned by studying the success of other programs.

Stufflebeam was credited with creating the CIPP evaluation model in the 1960s. (Kahan, 2008). This model requires evaluators to study the context that gave rise to the program. The influence of priorities, personalities and politics may be considered within the context component of the program evaluation process, not simply "what was done" and "what was the impact." A utilization focused evaluation is designed based on who will use the results and how the results will be used. The MDE PET (Appendix A) specifically

includes reference to the “next steps” required. The intent is that the evaluation will be used to improve or somehow otherwise change the program.

Role of the School Leader in Program Evaluation

Huber and Harvey (2016) concluded the program leader’s perception and conceptualization of evaluation can distance evaluation theory from practice at the program level. Boyd, et al. (2007) stated funders often insist on quantitative evaluations which collapse the evaluation effort down to simple monitoring and accountability activities and do not lead to service providers being able to improve services.

Data can be used by principals and school staff for a variety of purposes. Shen et al. (2010) found the use of student achievement data by principals was limited to three primary purposes: (1) accountability, (2) comparing students in terms of norm-referenced growth, and (3) grouping and placing students (identifying student weaknesses and strengths) to receive particular instructional experiences. Fewer than half of the principals mentioned school improvement as an area of data use. Apart from limiting purposes, the types of data principals focused on were limited. Other data sources were often neglected in favor of achievement data (Shen et al., 2010). Community data and school process data were rarely used by principals in the sample. When used, principals reported using such data for decisions regarding which particular instructional strategy should be used, or understanding reasons for student’s behavior (Shet et al., 2010). Use of this type of data in this way focusses on defining student attributes, and activities the organization can implement or change in order to improve the outcome, rather than how this data reflects community, organizational or structural attributes, which could be used for organizational improvement. For example, concluding student tardiness data is an

indicator that the tardiness policy or process is not severe enough for students, before asking if attendance is somehow related to other factors, for instance, lack of childcare options which cause the student to be late caring for a younger sibling, an opportunity for organizational and community improvement is lost. In cases where community or process data was used, principals did not mention evaluation of academic or non-academic programs as areas being informed by the use of that data (Shen et al., 2010).

When leaders perceive evaluation, and subsequent data use, in a singular context, the applications to other contexts can be lost. The likelihood evaluations will be followed by appropriate actions to improve student outcomes is diminished when the evaluation is merely viewed as a compliance activity. When leaders are able to frame evaluation activities in schools as useful for multiple reasons, not just to comply with legislation, the process and product of evaluation has a greater potential to improve student outcomes. There is a need to support principals and teachers in the interpretation and use of data (Schildkamp & Teddlie, 2008; Schildkamp, Visscher & Luyten, 2009). Misinterpretation and misapplication of data decreases the likelihood that the data will be used to improve outcomes for students. More support and training in this area would support the drive to improve student outcomes through completing quality, actionable evaluations.

Engaging in ongoing inquiry and reflection is an important factor which separates schools with deep impact from schools with less significant impact (Lee, 1999). Having these activities as established norms of practice in school communities offers greater opportunity that outcomes will improve for students. Norms of practice do not necessarily just happen or evolve. Embedding inquiry and evaluation as an expected practice does not happen without specifically planning for evaluative activities. Boyd et al. (2007)

indicated the strategic planning of evaluation is often absent, particularly in voluntary organizations and community groups and the consequence of a lack of strategic processes to evaluate is the tendency to produce uncritical evaluations. Uncritical evaluations fail to produce actionable results. Without action or change, outcomes for students remain the same.

Conducting a program evaluation can be foreign to schools and embedding the process of evaluation takes time, patience and multiple attempts for targeted skill development (Lee, 1999). Rather than relegating the evaluation process to an uncritical compliance obligation without the potential of yielding actionable results and improving student outcomes, organizations can experience positive outcomes for both teachers and students using an evaluation capacity building model (Haeffele, Hood & Feldman, 2011). If the capacity to produce robust, actionable evaluations is something a school community lacks, it can be built.

Evaluation Capacity Building

Baizerman, Compton and Stockdill (2002) defined evaluation capacity building (ECB) as: "The intentional work to continuously create and sustain overall organizational processes that make quality evaluation and its uses routine" (p. 109). Two themes emerged in the literature regarding where evaluation capacity is situated, (1) individually and (2) organizationally. Farley-Ripple and Butram (2015) offered a third, socially. They defined individual capacity as the knowledge and skills individuals bring to the process of evaluation; organizational capacity as a function of leadership as the allocation and coordination of resources are leadership functions, as well as the actions of leaders which, in part, define organizational attributes; and social capacity as the social networks

which allow for interpretation, collaboration and action by groups. In that way evaluation capacity, and evaluation capacity building, occur across multiple contexts: individual knowledge and skills (of teachers and leaders); organizational attributes (actions of leaders); and social networks (actions and interactions of groups).

The model outlined by Taylor-Ritzler et al. (2013) posited organizational factors affect relationships between individual factors and the evaluation capacity of the organization. ECB practices develop individual knowledge, skills and attitudes but, organizational factors such as leadership, culture, systems and structures, as well as overall communication within the organization also either facilitate or hinder the individual learning into organizational learning (Taylor-Ritzler et al., 2013). The building of evaluation capacity begins with individual knowledge and skills before addressing issues of organizational factors,

Evaluation is a process that is inherently complex (Oliver et al., 2002). Regarding evaluation skills, considering the funding limitations of many K-12 schools, the use of an external evaluator trained in the complex process of program evaluation is unlikely. Without funds for an external evaluator, evaluation skills, a component of evaluation capacity, need to be found from within the program team (Huber & Harvey, 2016). The teachers and school principals, who form the school program team, comprise the skill pool for conducting quality program evaluations. One avenue for developing principal and teacher evaluation skills is within principal and teacher preparation programs.

Tucker and Codding (2002) suggested that principal preparation programs should highlight the “the crucial role of data in the drive for results, from the careful setting of targets to the collection, display, and analysis of implementation and outcome data to the

use of data for setting goals, monitoring progress, allocating and reallocating resources, and managing the school program" (p 37). Setting targets, collecting data, analyzing implementation data and outcome data are all components of an evaluation plan. These are substantial components of principal preparation programs, although school principals are usually not specifically trained in conducting research, data collection and data interpretation activities (Vanhoof et al., 2011) and the issues of data and research as a whole receive very limited attention in principal preparation programs (Hess & Kelly, 2007).

Teacher preparation programs are an avenue for developing evaluation skills of school community members. Though data literacy and evaluation are not synonymous, the skills of data literate educators are directly applicable to the process of program evaluation. Skills of data literacy include problem-focused skills (formulating questions, identifying problems and making decisions) and data-focused skills (accessing, generating and interpreting data) (Mandinach & Gummer, 2013). In an analysis of teacher preparation programs, Mandinach, Friedman and Gummer (2015) found a disconnect between what faculty in schools of education believe they are teaching in relation to data literacy and what was actually being taught. They also found conflation between the terms data literacy and assessment literacy. The Data Quality Campaign (2014) described data literate educators as those who "continuously, effectively, and ethically access, interpret, act on, and communicate multiple types of data from state, local, classroom, and other sources to improve outcomes for students in a manner appropriate to educators' professional roles and responsibilities" (p. 1). Mandinach, Friedman and Gummer (2015) offered a comprehensive definition of data literacy:

> Data literacy for teaching is the ability to transform information into actionable instructional knowledge and practices by collecting, analyzing, and interpreting all types of data (assessment, school climate, behavioral, snapshot, longitudinal, moment-to-moment, etc.) to help determine instructional steps. It combines an understanding of data with standards, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge, and an understanding of how children learn. (p. 3)

Data literacy for program evaluation and improvement can be extrapolated from this definition to a focus on the school program rather than classroom. The replacement of "teaching" and "instructional" with "improvement" yields the following paraphrased statement: Data literacy for improvement is the ability to transform information into actionable improvement knowledge and practices by collecting, analyzing…to help determine improvement steps. Reframing in terms of Ackoff's (1999) DIKW hierarchy, data leads to information, leads to knowledge, leads to wisdom (decisions of good vs. bad, assigning value or, evaluating). Though the analysis of assessment data is one component within this process, there remain many more components to data literacy when compared to assessment literacy, and many more components to evaluation skills or capacity than assessment literacy alone.

State licensure standards reflect skills and knowledge that should be addressed by preparation programs, not necessarily what schools of education are including within their required coursework. In this way, perhaps policy makers are ahead of institutions of teacher preparation as Mandinach, Friedman and Gummer (2015) stated: "There is an obvious disjunture (sp.) between what the schools of education report and think they are teaching, and what content actually appears in the syllabi" (p. 38). In regard to state licensure standards, Mandinach, Friedman and Gummer (2015) found more than 20 states, in a sample of 49 states (Wyoming not included due to lack of identifiable

regulations), reference 25 elements or skills related to data literacy. Among these skills, with the corresponding number of the 49 states including those skills in parentheses, were: Plan (39, 80%); Use multiple measures (39, 80%); Use data (37, 76%); Involve stakeholders (36, 73%); Monitor (36, 73%); Communicate (34, 69%); Evaluate (33, 67%); Analyze (33, 67%); Collect/gather (31, 63%); Document/review (31, 63%); and Make decisions (28, 57%) (Mandinach, Friedman & Gummer, 2015). Though a clear majority of institutions indicate they either offer a stand-alone data course or integrate data literacy into existing courses, the review of syllabi indicated these courses emphasize assessment analysis to the detriment of other data types (Mandinach, Friedman & Gummer, 2015).

Another avenue for developing evaluation skills is professional learning for practitioners who have already completed preparation programs. The need to build evaluation capacity within current leaders engaged in program evaluation was made apparent as many leaders could not clearly separate the term evaluation from program. Huber and Harvey (2016) found numerous instances where leaders expressed a lack of understanding between the terms evaluation and program. When asked if the evaluation was conducted according to plan, the responses of program leaders focused on the program not going as planned and did not address evaluation as separate from the program. The terms were interpreted synonymously. With additional questions regarding the use of evaluation results, most project leaders discussed the usability of the program results towards ameliorating the conditions the program was intended to address, rather than the usability of the evaluation itself to improve the program (Huber & Harvey, 2016). For teachers having already completed preparation programs, Jimerson and Wayman

(2015) found data-related professional learning was not a specified process within their district, nor part of a longer-term district plan. Professional learning around data use was instead gained in isolated occurrences and more often used in the moment than preserved for further learning and use within the organization. Articulated plans to guide this type of work are important to fully develop capacity surrounding data use (Wayman, Jimerson & Cho, 2012). Jimerson and Wayman (2015) offered specific recommendations regarding professional learning in school districts surrounding data use: (1) District leaders should have a fully integrated plan; (2) District leaders should implement additional context-relevant platforms to enable knowledge sharing and preservation; and (3) District leaders should include data systems training using timely contexts of application (when teachers were solving actual problems).

CHAPTER 3

Evaluation Model

Taylor-Ritzler et al. (2013) used a synthesis model of evaluation capacity created through a systematic review of published models, evaluation capacity building principles, and factors believed to support evaluation in organizations. The model includes individual factors (awareness, motivation and competence) and organizational factors (leadership, climate, resources) as well as mainstreaming (evaluation practices as part of regular work processes) and use (of evaluation findings). Mainstreaming and use were initially components of a first order outcomes factor but, the data was not a good fit. Mainstreaming and use were maintained as separate first order factors and an acceptable fit was established by Taylor-Ritzler et al. (2013). Under that model, they found support for the ECAI as a generalizable tool for assessing evaluation capacity within the CSO setting.

The Taylor-Ritzler et al. (2013) model will be adjusted and reexamined in terms of K-12 structural attributes and will reflect mainstreaming and use as separate first order factors. The model requires adjustment as funders, program managers and leaders within community service organizations do not always have a direct counterpart in K-12 due to variances in organizational structures such that a one-to-one substitution cannot easily be made between the two settings.

Free and reduced lunch participation as a percent of the number of students attending the school will be used as a variable indicating the school's poverty level.

Procedures

From the “2016-2017 District and School Accountability Master Results File” (MDE, 2018), a new blank column will be added on the left side of the sheet to receive a generated random number. The Excel® data analysis add-in will be used to assign the schools a random number between 0 and 1. The number of variables will be 1; the number of random numbers will be 3,433; the selected distribution will be uniform; the number will be between 0 and 1; the output range will be placed in a new column; and the 1,150 schools with the highest random values will be included in the sample. The initial seed number will be set by using the randbetween function. The number 769 was selected via this method.

Principals of sampled schools will be sent an e-mail outlining the study, providing all required components regarding participation, informed consent and a request for participation. E-mail addresses of sampled principals will be obtained from the reports publicly available via the Center for Educational Performance and Information (CEPI) (2017), a State of Michigan sponsored electronic depository for public education information. The Educational Entity Master (EEM) page within CEPI contains reports detailing lead administrator contacts for all institutions of public education, by school identifier code. A link will be included in the e-mail to complete the revised ECAI on-line via Survey Monkey®. The Survey Monkey® account will be held and administered only by the researcher. Responses will not be identified by individual name but by the random number generated for that school. Non-respondents will receive an additional e-mail three to four weeks after initial contact containing a second request for participation and a final request of remaining non-respondents will be sent three to four weeks after the second

request. Results will downloaded from Survey Monkey®. They will be compiled and maintained on a USB flash drive stored in a fire-proof lock box at the researcher's home.

Schools will be partitioned into high and low poverty schools using school free and reduced lunch student count data. This will be obtained by matching the respondent's random number to the unique state code for each school. This will be cross referenced with the publicly available data set, "Free and Reduced Lunch Counts" retrieved from MISchool Data (2017). MISchool Data is another State of Michigan public portal for education data. The school's free and reduced lunch percentage will be added to each principal's survey response set as a demographic variable.

Sample

A simple random sample of 1,150 public and charter school principals of the State's 3,433 K-12 schools will be selected from the publicly released MDE District and School Accountability (2018) information, specifically the "2016-2017 District and School Accountability Master Results File." This file contains all K-12 public and charter schools in the State.

Data analysis

Data will be downloaded from Survey Monkey® to a flash drive and uploaded to both SPSS version 25 and SPSS Amos version 24 using the same flash drive. Data integrity will be checked by verification that the number of response sets in Survey Monkey® matches the number of response sets in SPSS and the number of responses to each question appearing in Survey Monkey® is the same number of responses to each question in SPSS. Every 50th set of responses downloaded to SPSS will be compared, item by item, to the corresponding set of responses in the Survey Monkey® file. The

descriptive statistics to be included are the number of principals of high poverty schools responding, the number of principals of low poverty schools responding and the number of principals of non-high/low poverty schools responding. The median Likert values by ECAI subsection and median Likert values of the ECAI in total will be disaggregated by poverty level and included as part of the analysis.

Cronbach's α, a measure of internal consistency reliability, will be computed to determine the ECAI's reliability. Values of .8 or higher are considered adequate in most applications (Nunnally & Bernstein, 1994).

A confirmatory factor analysis will be conducted to determine the degree to which the questions reflect factors and fit the model of evaluation capacity outlined by Taylor-Ritzler et al. (2013) and to provide support of construct validity. Four fit indices will be computed in SPSS Amos. A chi-square (χ^2) value equal to or less than the critical value established by setting nominal α at .05 will be determined to be an acceptable fit. Root mean square error approximation (RMSEA) less than or equal to .08, as identified by Hu and Bentler (1999) to establish relatively good fit will be reported as will the comparative fit index (CFI), using a suggested cutoff value of .95 (Hu & Bentler, 1999) as it adjusts for attenuation from sample size. The Tucker-Lewis index (TLI) value of at least 0.95 will be used as a marker of acceptable fit (Hu & Bentler, 1999). These fit indices, χ^2, RMSEA, CFI and TLI were used by Taylor-Ritzler et al. (2013) with the original ECAI as an indicator of model fit. A One Sample Kolmogorov-Smirnov test will be conducted to confirm a multivariate normal distribution, a baseline assumption of model fit indices (Kline, 2011), even though the formulas themselves are apparently non-parametric, the robustness of the formulas themselves has yet to be ascertained (Rose et al., 2017).

The Wilcoxon Mann-Whitney will be used to determine if principals of higher poverty schools are more likely to appear in the lower ranks of self-reported evaluation capacity in order to surface possible differential distribution of evaluation capacity between principals of high and low poverty schools (nominal α to establish result significance is $p \leq .05$). This nonparametric test will be used, because it is unknown if evaluation capacity responses are normally distributed. Zumbo and Zimmerman (1993) commented in applied settings the unique optimal or best test is unknown. Normal distributions in these settings are quite rare as Micceri (1989), in a review of 440 distributions from published studies, found virtually none were normal. An assumption of normality is required should the independent samples t test be used. Robust tests are not sensitive to violations of underlying assumptions. Sawilowsky (1990) stated, "the robustness is related not only to Type I error, but also to Type II error (p. 98)." Non-normal distributions can impact Type I and Type II error rates. Moreover, the Wilcoxon Mann-Whitney is more powerful when normality is violated than the independent samples t test (Sawilowsky & Blair, 1992).

In order to administer the survey, approval will be obtained from Wayne State University's Human Investigations Committee (HIC) prior to administration and follow all established principles for research with human subjects.

CHAPTER 4

Descriptive and Internal Consistency Statistics

The survey was completed and submitted by 121 (10.5%) respondents of the 1,150 sampled from the population of 3,433. This is equivalent to a 95% Confidence Interval with a margin of error of ±8.75%.

Descriptive statistics by group are compiled in Table 1. Thirty-seven respondents (31%) were from schools with 70% or greater free and reduced lunch rates, with 20 (17%) of the 121 from schools with 30% or less. A One Sample Kolmogorov-Smirnov test was conducted to confirm a multivariate normal distribution. The total score distribution departed from normality, p <. 05. The mean = 200.98 and standard deviation = 25.58 (Figure 1). The high poverty group had a slightly lower mean and median than mid and low poverty groups but had a standard deviation between the mid and low poverty groups. The group with the greatest standard deviation of total score was the low poverty group. The subsection with the greatest SD = .96 for Likert responses from 1 - 4 was Resources as shown in Table 2. Learning Climate SD = .69 was the lowest subsection deviation of the 8 in the assessment. Cronbach's α, a measure of internal consistency, were above .8 for the total score, and for seven of the eight subsections. The only subsection with α below .8 was Mainstreaming, with a value of .74. The highest Cronbach α was the assessment in total, .96, as well as the Competence subsection, .96 as shown in Table 2.

Table 1. *Total Score Mean, Median and Standard Deviation by Study Groups*

	N	Mean Total Score	Median Total Score	SD Total Score
High Poverty	37	199.35	200.00	26.70
Mid Poverty	64	201.69	202.00	24.49
Low Poverty	20	201.70	202.00	28.05
All	121	200.98	201.00	25.58

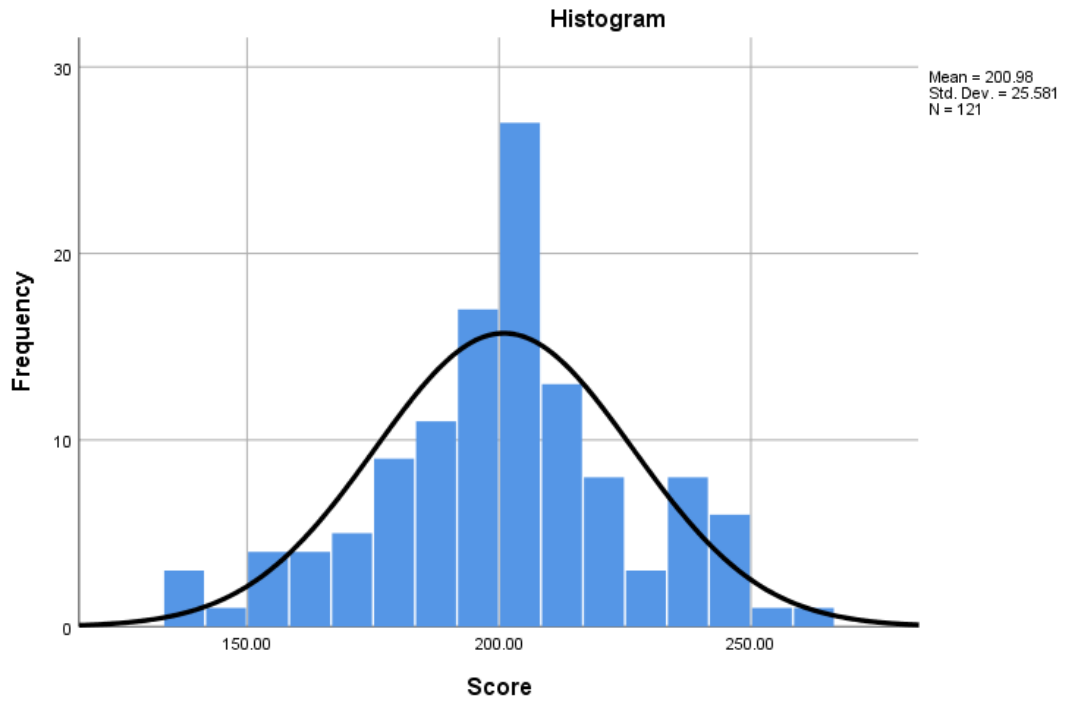

Figure 2. *Histogram of Score Distribution with Normal Curve*

Table 2. *Descriptive and Internal Consistency Statistics of Construct Questions by Subsection and Total*

	# Items	M	Med	Min	Max	SD	Cronbach α
Awareness	11	3.03	3	1	4	.77	.89
Motivation	4	3.10	3	1	4	.72	.86
Competence	14	3.09	3	1	4	.69	.96
Leadership	5	2.81	3	1	4	.74	.85
Learning Climate	9	3.21	3	1	4	.69	.85
Resources	9	2.80	3	1	4	.96	.84
Mainstreaming	6	2.98	3	1	4	.73	.74
Use	11	2.63	3	1	4	.78	.93
Total	68	2.96	3	1	4	.79	.96

Scores of each subsection were significantly correlated (p <.0 5) with all other subsections with one exception (Table 4). Learning Climate and Motivation were not statistically significantly correlated. The greatest correlation was between Mainstreaming and Use, .71; Awareness and Motivation, .68; Resources and Use, .61; and Mainstreaming and Resources, 57. The lowest correlation values were between Learning

Climate and Motivation (non-significant), .17; Awareness and Learning Climate, .19; Motivation and Resources, .22; and Awareness and Competence, .22.

Table 3. *Mean, Median and Standard Deviation of Subsection Scores by Study Groups*

	High Poverty			Mid Poverty			Low Poverty			All		
	M	Med	SD	M	Med	SD	M	Med	SD	M	Med	SD
Awareness	33.49	34	4.72	33.61	34	5	32.3	32.5	5.53	33.36	34	5.53
Motivation	11.97	12	2.48	12.67	12	2.39	12.2	12	2.31	12.38	12	2.41
Competence	42.38	42	7.59	43.31	43.5	7.06	44.9	43.5	5.68	43.29	43	7.56
Leadership	13.55	14	2.92	14.2	15	2.9	14.4	14.5	2.96	14.03	15	2.91
Learning Climate	28.46	27	4.28	28.97	29.5	4.19	29.7	29.5	3.77	28.93	29	4.14
Resources	25.68	27	4.35	24.91	25	4.92	25.15	26	6.75	25.18	25	5.07
Mainstreaming	14.62	14	2.36	15.03	15	2.23	14.95	15	2.91	14.89	15	2.38
Use	29.22	30	6.9	29.98	29.5	5.72	28.1	27.5	6.74	28.91	29	6.23

Table 4. *Correlations Matrix of Study Constructs*

	1	2	3	4	5	6	7	8
1. Awareness	-	-	-	-	-	-	-	-
2. Motivation	.68**	-	-	-	-	-	-	-
3. Competence	.22*	.25**	-	-	-	-	-	-
4. Leadership	.31**	.27**	.42**	-	-	-	-	-
5. Learning Climate	.19*	.17	.47**	.44**	-	-	-	-
6. Resources	.28**	.22*	.44**	.50**	.43**	-	-	-
7. Mainstreaming	.34**	.31**	.54**	.55**	.50**	.57**	-	-
8. Use	.35**	.24**	.51**	.53**	.52**	.61**	.71**	-

**p<.01. Correlation is significant at the 0.01 level.

*P<.05. Correlation is significant at the 0.05 level.

Factor Analyses

Three separate Confirmatory Factor Analyses were conducted using the eight first order and three second order latent constructs as specified in the original Taylor-Ritzler et al. ECAI model (2013) (Figure 3, Figure 4 and Figure 5). Fit indices for those models appear in Table 5. Chi-square gave significant results (754.78, *df*=374, *p*<.05; 449.27, *df*=206, *p*<.05; 233.72, *df*=103, *p*<.05) (Table 5), and all fit indices (RMSEA, CFI and TLI) were out of recommended ranges (Table 5). There is then cause for concern regarding model fit. It should be noted that chi-square is sensitive to sample size.

Table 5. *Indices of Fit*

	χ^2	*df*	χ^2/df	RMSEA	CFI	TLI
CFA 1: Individual Factors	754.78*	374	2.02	.09	.84	.83
CFA 2: Organizational Factors	449.27*	206	2.18	.10	.81	.76
CFA 3: Outcome Factors	233.72*	103	2.27	.10	.89	.87

Note CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker-Lewis Index.

**P*<.05. Chi-square is significant at the 0.05 level.

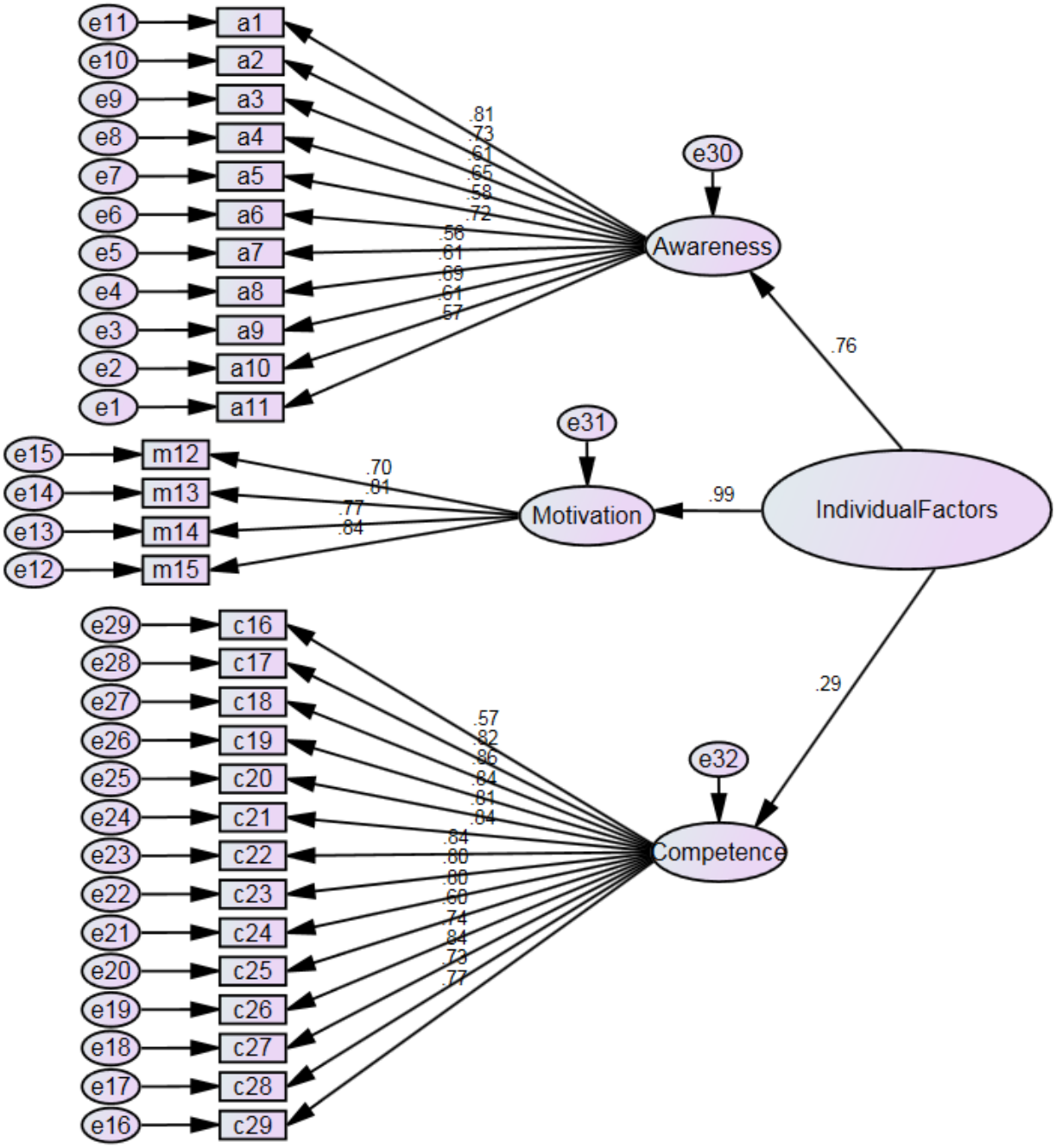

Figure 3. CFA Model 1, Individual Factors

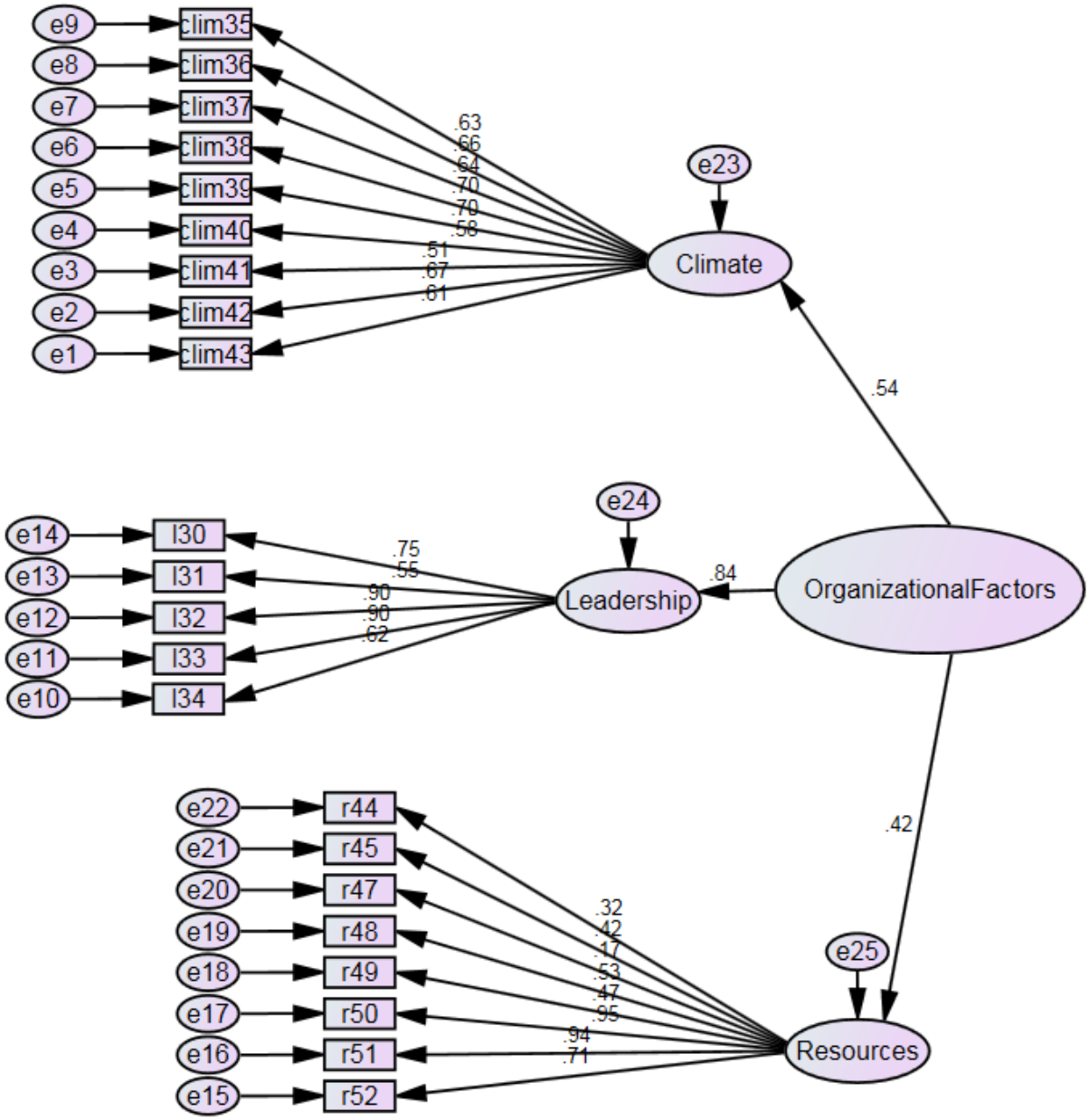

Figure 4. CFA Model 2, Organizational Factors

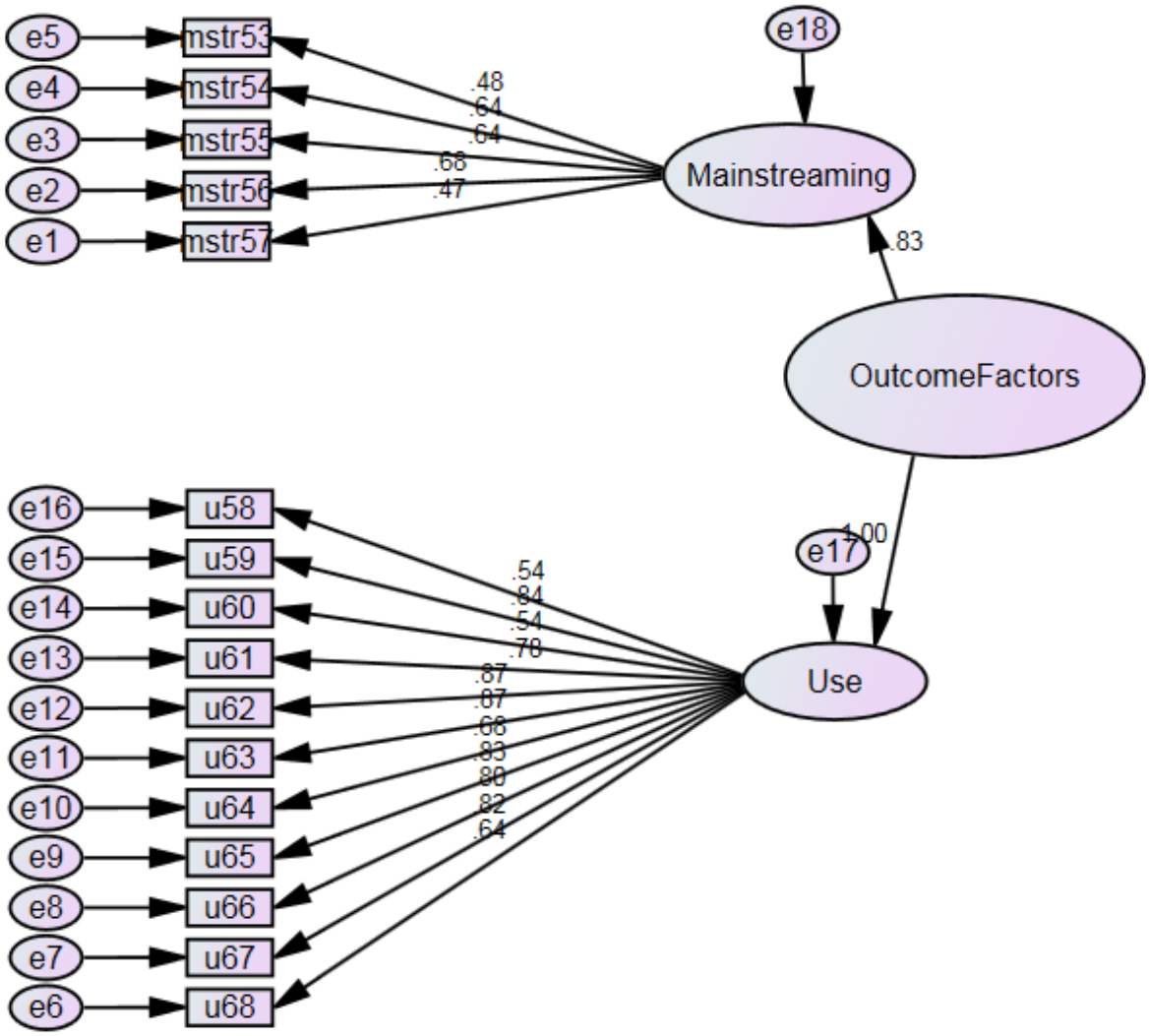

Figure 5. CFA Model 3, Outcome Factors

A preliminary exploratory factor analysis via principle component analysis (PCA), was conducted. Based on the scree plot, an inflection point indicated five potential factors from the sample data (Figure 6) and Kaiser's criterion for retaining factors, those with eigenvalues greater than 1, indicated 15 factors. Cumulatively, five factors explained 52.55% of the variance (Table 6). A solution with eight components, the number of first order latent factors explicated in the original Taylor-Ritzler et al. (2013) ECAI model,

explained 61.36% (Table 6) and following Kaiser's criterion, 15 factors explained 74.40% (Table 6).

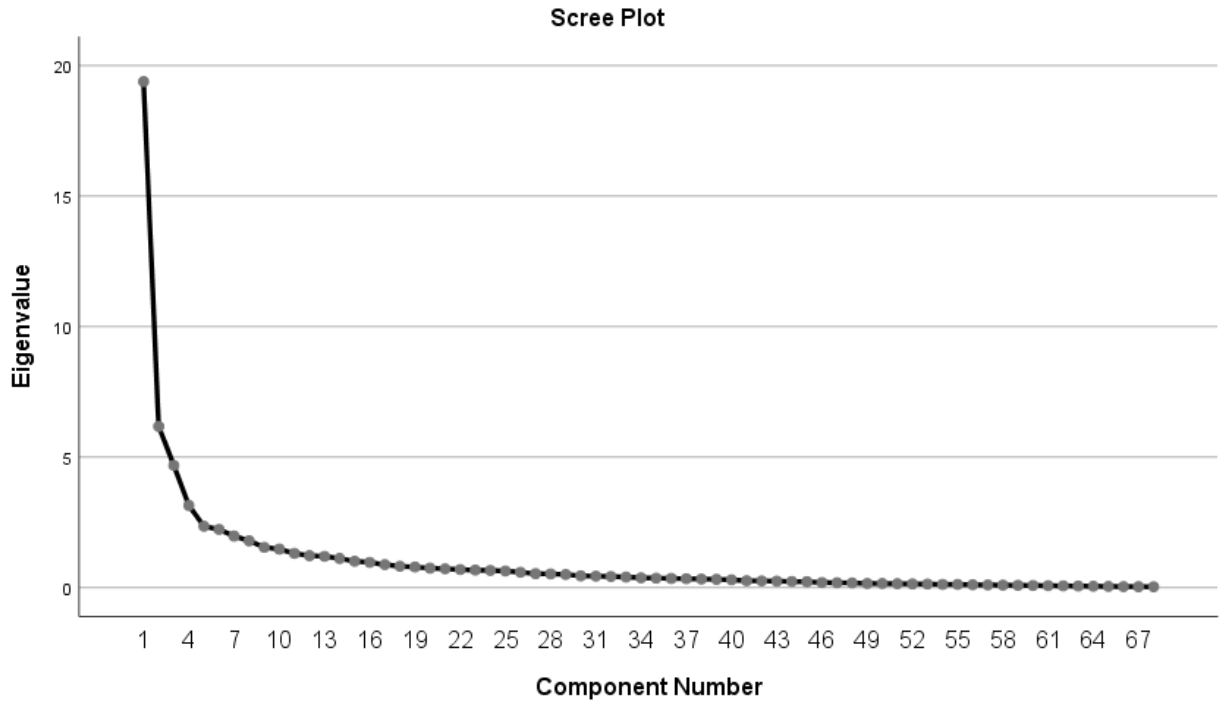

Figure 6. Scree Plot from PCA

Table 6. *PCA: Total Variance Explained*

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumu-lative %	Total	% of Variance	Cumu-lative %	Total	% of Variance	Cumu-lative %
1	19.38	28.51	28.51	19.38	28.51	28.51	10.18	14.97	14.97
2	6.17	9.08	37.58	6.17	9.08	37.58	7.66	11.26	26.23
3	4.68	6.88	44.47	4.68	6.88	44.47	5.04	7.42	33.65
4	3.15	4.63	49.09	3.15	4.63	49.09	3.49	5.13	38.77
5	2.35	3.45	52.55	2.35	3.45	52.55	3.45	5.07	43.85
6	2.23	3.28	55.83	2.23	3.28	55.83	3.26	4.80	48.65
7	1.97	2.90	58.73	1.97	2.90	58.73	3.24	4.77	53.42
8	1.79	2.63	61.36	1.79	2.63	61.36	2.60	3.83	57.24
9	1.54	2.27	63.63	1.54	2.27	63.63	2.30	3.38	60.62
10	1.47	2.17	65.80	1.47	2.17	65.80	1.95	2.86	63.49
11	1.31	1.93	67.72	1.31	1.93	67.72	1.85	2.72	66.20
12	1.22	1.80	69.52	1.22	1.80	69.52	1.64	2.41	68.62
13	1.20	1.76	71.28	1.20	1.76	71.28	1.42	2.09	70.71
14	1.12	1.64	72.92	1.12	1.64	72.92	1.30	1.92	72.62
15	1.01	1.48	74.40	1.01	1.48	74.40	1.21	1.78	74.40

Extraction Method: Principal Component Analysis.

Some of the ECAI items loaded onto several of the 15 factors in alignment with the *a priori* 8 factors explicated in the original ECAI model (Table 7, Table 8 and Table 9). Exact loading values are listed in Table 7. Values obtained by suppressing values less than .3 are listed in Table 8. Competence items (c16 to c29) loaded onto component 1 with values from .53 to .82. Comparing loading values across the components, each loading value for Competence items was more than double the item's loading value for any other potential component.

Use items (u59 to u68) loaded onto component 2 with values from .48 to .80. Comparing loading values across the components, each loading value, except that of u60, was more than double the item's loading value for any other potential component.

Motivation items (m12 to m15) loaded onto component 3 with values from .71 to .83 however, Awareness items (a9 to a11) also loaded to component 3 with values from .59 to .69 (Table 7). Learning Climate items (clim36 and clim40 to clim43) loaded to component 4 with values from .59 to .77 however, Learning Climate items (clim37 to clim39) loaded to component 8 with values from .55 to .72. The item clim35 from Learning Climate had similar factor loadings to both components 4 (.38) and 8 (.31). Leadership items (l30 and l32 to l34) loaded onto component 5 with values from .72 to .79 and the item l31 also loaded onto component 5 however, the loading value was .34 and similar loading values for l31 were apparent in components 12 (.31) and 4 (.30). Three Resource items (r50 to r52) had high loading values to component 7 (.76 to .89). The remaining Resource items (r44 to r49) displayed the highest loading values either alone or in pairs across several components (r46 and r47 to component 9 with values .77 and .79; r48 and r49 to component 13 with value .51 each; r44 to component 12 with value .66; and r45 to component 8 with value .47). Mainstreaming item mstr57 had a high loading value to component 11 (.75). Items mstr53 and mstr54 had loading values of .44 and .42 to component 2 but, mstr53 had a loading value of .53 to component 14 and mstr54 had a loading value of .52 to component 15. Item mstr55 had identical loading values (.41) to components 9 and 1.

Table 7. *PCA: Rotated Component Matrix*[a]

	PCA Component Number														
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
c27	.82	.29						.11							
c18	.82	.12	.11	.17			.10		.10				.11	.15	.18
c23	.81	.16							.15				-.11	.11	
c19	.80	.16						.20	.12						.17
c22	.80	.14	.18			.11					.23				
c20	.80	.12		.11	.12			.10		.14			.17	-.16	
c17	.80	.11	.20	.11											.11
c24	.79	.20			.15		.14					.11			
c21	.77	.17	.19	.18					.10		.22				
c29	.76	.11	.12		.14		.15							.26	
c28	.73				.14	.11		.17	.12				-.23		
c26	.73	.20			.15						.11	.10		-.16	-.25
c25	.61	.29							-.30	-.14	.10	.34	.14		.11
c16	.53	.23			.18		.14				-.13	-.17		-.50	
u62	.23	.80		.14		.18		.22	.10						
u67	.27	.80		.12	.18									.11	
u66	.20	.79	.11		.12					.16				.13	
u65	.26	.78		.10	.15				.14						
u63	.29	.76		.13		.16		.24			.10		.11	-.16	
u59	.31	.71	.17	.22		.11	.12		.14		.20				.12
u61	.21	.70		.13	.17	.14	.12	.12				.29	-.11		
u68		.63	.22			-.22	.17	.13		.23		.22		-.25	.15
u64	.27	.62	.12		.20		.13		-.14	.25	.13	.11	.18	.14	
u60		.48				.34	.39		-.20	.11	.13	.33			
mstr56	.30	.33	.30	-.12		.12	.13	.27	.29	.17	.21	.29			.27
m12			.83										.14	-.16	
m15			.77		.15	.18	.12	-.15	.14	.23					
m13	.16	.13	.76			.12									.19
m14	.23		.71	.22	.14	.17			.25						
a10			.69	-.14		.31		.29		-.10		-.12	.12		
a11	.19	.12	.64		.10	.14		.30	-.15		.11		-.19	.18	.12
a9	.19	.18	.59			.31	.10			.20			.21	.33	-.16
a6		.23	.42	-.12		.40	.15	.25		.37		.22	.18		-.11
clim40				.77				.13			.24				.18
clim41	.16	.25		.68						.20	.11		-.24		
clim42		.22		.67	.17			.23			.11		.24		-.18
clim36	.30			.66			-.13	.18	.12					.17	
clim43	.29	.22		.59					.18			.14	.11		-.11
clim35	.12	.29		.38	.23			.31		-.20		.30	.19	.22	.24
l33	.21	.23	.13	.10	.79				.10		.17	.17		.11	
l32	.28	.22			.78	.11	.14					.19	-.17		
l34		.18			.76	.13				.14		-.19	.15		.11
l30	.23	.21	.20		.72		.15			.13		.11	.12	-.17	
l31	.27	.20	.13	.30	.34		.18	.25	-.27	-.19	.21	.31			-.14
a2			.33			.78					.10				-.15
a3			.21			.72				.17				-.19	
a1			.43		.15	.72				.13		.11		.17	

a7		.14	.31		.23	.43	-.14	.21		.10	-.12	.17	-.26		.24
r51	.18	.10			.12		.89								
r50	.18	.15					.88						.11		
r52	.17	.24			.20		.76				.11				
clim37	.23	.19		.26				.72	.12			-.16	.12		
clim38	.36	.11		.31				.64	.13		.17	.18			-.10
clim39	.15	.25	.17	.42	.13			.55			-.15	.22			.22
r45		.32	.12			-.20	.32	.47	.12	.24	.14	.13	.22	.14	
r46	.20					.13	.11		.77		.12	.12			
r47	.26	.10			.22				.75				.14		
mstr55	.41	.35		.11	.17		.10	.12	.41			-.19		-.11	.17
a5		.12	.22	.11	.22	.32				.72					-.13
a8	.16		.33			.34	.11	.12		.53					
a4		.14	.26		.11	.49				.52	-.11		.23		.35
mstr57	.12	.16		.28	.19	.11			.16	.11	.75				
u58	.21	.37						.16			.74				
r44		.25			.19		.21		.25			.66			
r49	.29	.32	.21	.16	.15		.29	.16	.31			.10	.51		
r48		.44			.14		.40	.23					.51		
mstr53	.11	.44		.16	.29	-.10	.12	.10		.12				.53	
mstr54	.25	.42	.11		.12		.23	-.14	.13		.21				.52

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 20 iterations.

Table 8. *PCA: Rotated Component Matrix[a] (suppressed)*

	PCA Component Number														
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
c27	.82														
c18	.82														
c23	.81														
c19	.80														
c22	.80														
c20	.80														
c17	.80														
c24	.79														
c21	.77														
c29	.76														
c28	.73														
c26	.73														
c25	.61								-.30			.34			
c16	.53													-.50	
u62		.80													
u67		.80													
u66		.79													
u65		.78													
u63		.76													
u59	.31	.71													
u61		.70													
u68		.63													
u64		.62													
u60		.48				.34	.39					.33			
mstr56	.30	.33													
m12			.83												
m15			.77												
m13			.76												
m14			.71												
a10			.69			.31									
a11			.64					.30							
a9			.59			.31								.33	
a6			.42			.40				.37					
clim40				.77											
clim41				.68											
clim42				.67											
clim36				.66											
clim43				.59											
clim35				.38				.31				.30			
l33					.79										

l32					.78								.31			
l34					.76											
l30					.72											
l31					.34											
a2			.33			.78										
a3						.72										
a1			.43			.72										
a7			.31			.43										
r51							.89									
r50							.88									
r52							.76									
clim37								.72								
clim38	.36			.31				.64								
clim39				.42				.55								
r45		.32					.32	.47								
r46									.77							
r47									.75							
mstr55	.41	.35							.41							
a5						.32				.72						
a8			.33			.34				.53						
a4						.49				.52						.35
mstr57											.75					
u58		.37									.74					
r44												.66				
r49		.32							.31				.51			
r48		.44					.40						.51			
mstr53		.44												.53		
mstr54		.42														.52

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 20 iterations.

Table 9. *Greatest Loading Values from the PCA by Item and Actual Question*

PCA Component	Item		Question
1	c	16	Develop a program, strategy or initiative evaluation plan.
1	c	17	Clearly state measurable goals and objectives for my program, strategy or initiative.
1	c	18	Identify strategies to collect information from participants.
1	c	19	Define outcome indicators of my program, strategy or initiative.
1	c	20	Decide what questions to answer in an evaluation.
1	c	21	Decide from whom to collect the information.
1	c	22	Collect evaluation information.
1	c	23	Analyze evaluation information.
1	c	24	Develop recommendations based on evaluation results.
1	c	25	Examine the impact of my program, strategy or initiative on students from diverse ethnic/racial backgrounds and/or students with disabilities.
1	c	26	Write an evaluation report.
1	c	27	Conduct an evaluation of my program, strategy or initiative.
1	c	28	Conduct an evaluation of my program with support from others.
1	c	29	Present evaluation findings orally.
1	mstr	55	I have access to the information I need to make decisions regarding my work.
2	mstr	56	I am able to integrate program, strategy or initiative evaluation activities into my daily work practices.
2	u	59	To improve programs, strategies or initiatives.
2	u	60	To get additional funding.
2	u	61	To design ongoing monitoring processes.
2	u	62	To assess implementation of a program, strategy or initiative.
2	u	63	To assess quality of a program, strategy or initiative.
2	u	64	To improve community engagement.
2	u	65	To make informed decisions.
2	u	66	As a component of staff training.
2	u	67	To develop best practices.
2	u	68	To eliminate unneeded programs, strategies or initiatives.
3	a	6	Is absolutely necessary to improve my program, strategy or initiative.
3	a	9	Will help improve services to students from diverse ethnic/racial backgrounds and/or students with disabilities.
3	a	10	Is unnecessary because we already know what is best for our students.
3	a	11	Of a program, strategy or initiative is too complex to do.

3	m	12	Learn about program, strategy or initiative evaluation.
3	m	13	Evaluate my program, strategy or initiative.
3	m	14	Support staff to evaluate programs, strategies or initiatives.
3	m	15	Encourage others to buy into evaluating our program, strategy or initiative.
4	clim	35	Program, strategy or initiative evaluation information is shared in open forums.
4	clim	36	Staff is supported to introduce new approaches in the course of their work.
4	clim	40	Staff respects each other's perspectives and opinions.
4	clim	41	Staff errors lead to teachable moments rather than criticisms.
4	clim	42	Staff participates in making long-term plans for the program, strategy or initiative.
4	clim	43	Staff concerns are considered in most decisions regarding strategic planning and evaluation.
5	l	30	District leaders provide effective leadership.
5	l	31	Staff understands how everyone's duties fit together as part of the overall mission of the program, strategy or initiative.
5	l	32	District leaders communicate program, strategy or initiative goals and objectives clearly.
5	l	33	District leaders have a clear plan for accomplishing program, strategy or initiative goals.
5	l	34	District leaders have realistic expectations of what staff can accomplish given the resources they have available.
6	a	1	Will help me understand my program, strategy or initiative.
6	a	2	Will inform the decisions I make about my program, strategy or initiative.
6	a	3	Will justify funding for my program, strategy or initiative.
6	a	7	Should involve program, strategy or initiative participants in the evaluation process.
7	r	50	Grant funders provide resources (e.g. training, money, etc.) to conduct evaluation.
7	r	51	Grant funders provide leadership for conducting evaluation.
7	r	52	District leadership engages in ongoing dialogue with grant funders regarding evaluation.
8	clim	37	It is easy for staff to meet regularly to discuss issues.
8	clim	38	Staff is provided opportunities to assess how well they are doing, what they can do better, and what is working.
8	clim	39	Staff can encourage others to make use of evaluation findings.
8	r	45	Staff has time to conduct evaluation activities (e.g. identifying or developing a survey, collecting information from participants).
9	r	46	Staff has access to technology to compile information into computerized records.
9	r	47	Staff has access to adequate technology to produce summary reports of information collected from participants (e.g. computerized database).

10	a	4	Will help to convince others that changes are needed in my program, strategy or initiative.
10	a	5	Will inform changes in our documentation systems.
10	a	8	Will influence policy relevant to my program, strategy or initiative.
11	mstr	57	The evaluation activities I engage in are consistent with the State's expectations.
11	u	58	To report to the State.
12	r	44	Resources are allocated to provide accommodations for people from diverse ethnic backgrounds and for people with disabilities to collect evaluation information (e.g. interpreters, translated documents).
13	r	48	Resources are allocated for staff training (e.g. money, time, bringing in consultants).
13	r	49	Technical assistance is available to staff to address questions related to evaluation.
14	mstr	53	My school gathers information from diverse stakeholders to gauge how well the program, strategy or initiative is doing.
15	mstr	54	My school has adequate records of past program, strategy or initiative evaluation efforts and what happened as a result.

The greatest loading values for all Competence questions were associated with component 1 (Table 7). All but one question in the Use subsection loaded with the greatest value to component 2 and all Leadership items had the highest loading values to component 5. Four Awareness items and four Motivation items had the highest values associated with component 3. Other items from Awareness had high values for component 6 and 10. Learning Climate similarly had highest loading values in two components, 4 and 8. Mainstreaming and Resource items had their highest values over more than two potential components (Table 7).

Resource and Mainstreaming items (r48, r49, mstr53 and mstr54) had greatest loading values (.51 to .53) to components 13, 14 and 15 (Table 7) but, had the second highest loading values (.32 to .44) to component 2.

A Mann-Whitney U test was conducted to determine whether there was a difference in total scores between high and low poverty principals. Results indicated that there was not a difference (z = -.35, p = .73).

CHAPTER 5

Discussion

The first purpose of this study was to establish reliability and construct validity as fit evidence for the model outlined within the Evaluation Capacity Assessment Instrument (ECAI) developed by Taylor-Ritzler et al. (2013) for program evaluators in the CSO setting for assessing principal evaluation capacity in the public K-12 setting. Cronbach's α was computed to determine the ECAI's reliability. The assessment in total, and seven of the subsections, had adequate internal consistency reliability. The Mainstreaming subsection's Cronbach's α was the lowest at .74. The next lowest value was .84 and .96 was the highest. Reliability evidence was found for 7 of the 8 subsections and the assessment in total. Significant correlations between all pairs of subsections except for the Learning Climate and Motivation subsections were also found. Regarding model fit evidence, RMSEA, CFI and TLI did not meet suggested markers for good fit in any of the three CFAs and there are additional concerns with model fit given the significant results of the chi-square and the results of the fit indices. Chi-square is highly sensitive to sample size. Though some evidence of reliability was established, fit evidence was not established regarding the use of the ECAI as an assessment of the evaluation capacity of K-12 principals.

Digging deeper regarding the primary purpose and comparing these results with the original results from Taylor-Ritzler et al. (2013), Cronbach's α was above 0.8 in the Taylor-Ritzler et al. (2013) study for all subsections. Cronbach's α was not above 0.8 for the Mainstreaming subsection. Correlations between subsections in this study were all significant, except for Learning Climate and Motivation. In the Taylor-Ritzler et al. (2013)

study, that correlation was significant but, the correlations between Leadership and Awareness; Resources and Awareness; and Leadership and Motivation were not significant. For the first two CFAs, in this study, RMSEA was higher at .09 and .10 than the Taylor-Ritzler et al. (2013) study, which were .06 and .04. None of the CFI and TLI values in this study exceeded .84, whereas in the Taylor-Ritzler et al. (2013) study, all four of the values exceeded .90. There was some alignment of results from the Taylor-Ritzler et al. (2013) study, because the third CFA, Mainstreaming and Use, did not meet good fit criteria in either that or this study. Taylor-Ritzler et al. (2013) then adjusted the model regarding Mainstreaming and Use however, that was not done in this study as the first two CFAs did not meet good fit criteria.

There was also alignment of some results from the PCA to the original study however, there was also misalignment. The scree plot had an inflection point at 5 indicators which can be interpreted as potentially fitting a model with 5 components and the application of Kaiser's criterion of retaining factors with eigenvectors greater than 1, resulted in a potential 15 component model. Both the 5 and 15 component models are different than the original Taylor-Ritzler et al. (2013) model with 8 factors.

Commonalities and differences between the original 8 factor model and a potential 15 component model were displayed through a cursory examination of loading values in the PCA model with 15 components. The construct of Competence could be interpreted as component 1 from the PCA. All Competence items, and one Mainstreaming item (mstr55) had their highest loading values to component 1 (Table 8). Use could be interpreted as component 2. All Use items, and one Mainstreaming item (mstr56) had their highest loading values to component 2 (Table 8). Leadership could be component

5. All Leadership items had their highest loading values to component 5 (Table 8). The loading values for Competence, Use and Leadership as three of the original 8 factors in the model appeared distinctly as complete groups of assessment items potentially representing components 1, 2 and 5 and the two of the greatest Cronbach's α values were in the Competence and Use subsections (Table 2). The loadings for the other 5 Taylor-Ritzler et al. (2013) subsections (Awareness, Motivation, Learning Climate, Resources and Mainstreaming) were not as distinct as complete component sections within the PCA but, there were some similar loading values of combinations of sections and the remaining components.

Motivation items (m12 to m15) loaded with greatest values onto component 3 and so did Awareness items (a9 to a11) (Table 8). Items 12 through 15 were the entirety of the Motivation subsection (Table 8). Items a9 and a10 reference improving outcomes for students, "Will help improve services to students from diverse ethnic/racial backgrounds and/or students with disabilities," and "Is unnecessary because we already know what is best for our students" (Table 8). An individual's motivation to enter the field of education in many cases is to improve outcomes for students. Items a9 and a10 for these respondents were potentially assessing motivation rather than the hypothesized factor of Awareness or, the motivation items and these awareness items were assessing another unspecified factor.

From the remaining Awareness items, 4 items (a1, a2, a3 and a7) were connected to potential component number 6. Items a1, a2 and a3 had higher loading values to the component than did a7. The remaining three items (a4, a5 and a8) in the Awareness subsection were connected to potential component number 10. Items a4, a5 and a8

reflect a possible theme of change or improvement (changing programs, systems change, informing policy) (Table 8). The items intended to assess the hypothesized Awareness component were partitioned into three potential components: items that were intended to assess motivation; and two others, perhaps awareness, improvement, change or other unspecified factors entirely. Awareness items did not appear to distinctly assess a singular component.

A partitioning of Learning Climate items was also evident. Items clim36, and clim40 to clim43 loaded to component 4 however, Learning Climate items clim37-clim39 loaded with greater values to component 8. Learning Climate items clim36 and clim40-clim43 share a theme of a safe and collaborative work environment (Table 8). Items clim37-clim39, "It is easy for staff to meet regularly…Staff is provided opportunities to assess how well…Staff can encourage others…" and item r45 as a resource component (has time to conduct evaluation activities), centers around staff having time to participate in collaborative learning (Table 8). The items intended to assess learning climate were partitioned into two potential components: a safe and collaborative learning community, time to participate in collaborative learning, or other unspecified factors entirely.

Resource items and Mainstreaming items distributed their greatest loading values among 4 and 5 potential components, respectively, within the PCA. The 4 potential components that the Resources items were associated with share similarities and display differences. Resource items r50 to r52, as a potential component 7, include grant funders within the question, r44 and r47 include technology resource reference (potential component 9) and a potential component 13 includes questions related to resources for staff training and assisting staff with questions. The partitioning of these resource items

over 4 components could be indicative that the type of resources applied to evaluation capacity components matter. In another way, given reduced resources, or given an abundance of resources, the areas that school communities choose to allocate and/or access resources matters. Resource items did not appear to distinctly assess a singular component under the construct of evaluation capacity.

Considering Mainstreaming items loaded onto 5 different potential components; failed to meet .8 Cronbach's; posed difficulties in the original study, prompting a change to the original model by Taylor-Ritzler et al. (2013); and lacked good fit results as a factor in this study, Mainstreaming is not likely to be a distinct factor under the construct of evaluation capacity. Mainstreaming items mstr55 and mstr56 had high loading values with distinct components of Competence and Use, respectively. These questions may have been assessing these two other factors under the model. Potentially items mstr53 and mstr54 could also have loaded to Use as that was the component with the second highest loading values for these questions, rather than a potential component 14 or 15.

Mainstreaming, the notion that evaluation practices are part of regular work processes, and Resources appeared to be the least distinct as factors unto themselves. It is possible that these two components are much more integrated with other components. Some Mainstreaming and Resource items could gravitate to other components based on the context of the component. If Mainstreaming and Resources are not distinct factors themselves, but associated within other factors, that would reduce the original number of 8 *a priori* factors to a potential 6. This would be closer to the 5 potential factors reflected by the scree plot examined in this study.

A second purpose of this study was to determine if differences in capacity were connected to the poverty level of students in a school**.** Though good model fit evidence was not established, making any potential significant difference in capacity distribution inconclusive, the differences between total score distributions between high and low poverty groups were not significant as determined by the Wilcoxon Mann-Whitney. Evidence that evaluation capacity is differentially distributed between high and low poverty schools was not found.

Recommendations

Some promising elements to the ECAI exist. Internal consistency statistics provide support to the idea that some of the assessment items belong together and belong on the assessment overall. Lack of good fit evidence, however, prompts a need to re-conceptualize the model for K-12 principals.

Re-conceptualizing the model could begin with the number of potential factors. Given the relatively high loading values, Competence items (16-29) and Use items (59-68) could likely remain as items on the assessment and as 2 factors in the model. Seven items (9-15), some Awareness and all Motivation items, could be examined as potential items all assessing Motivation, a third factor. The Awareness items loading similarly to the entirety of the Motivation section involve some level of improving outcomes for students, the motivating factor for many educators. Potentially, items 9-15 could remain as items on the assessment for the factor of Motivation.

The partitioning of 8 Climate items into exactly 2 potential factors could also be addressed in a new model. What makes clim36 and clim40-43 different from clim37-39? The answer may be related to items clim35 and r45 (Table 8). Item clim35 had similar

loading values to both parts of that partition (Table 7), “Program, strategy or initiative evaluation information is shared in open forums.” For that type of sharing to happen in schools, two things must be present. The first is the forum for sharing itself and the second is the “open” nature of that forum. Items clim37-39 and r45 share a theme of opportunities for staff to work together, a forum for collaboration to occur. The other side of that partition is composed of items that share a theme closer to what may be more traditionally classified as climate, a respectful and supportive work atmosphere, the “openness” of the forum. Items clim36 and clim40-43 could remain as Climate questions, while clim37-39 and r45 may be identified as Collaboration. Three of the 4 Leadership items also had high loading values to a distinct component. As such, Competence, Use, Motivation, Climate, Collaboration and Leadership could frame a 6-component model for principals.

In any re-conceptualization of the model, Resources and Mainstreaming items should be closely examined in terms of how each item may impact the other factors under the model. Given the work of the principal and school community, the extent to which resources and daily evaluation activities are mainstreamed into the work are unlikely distinct components. For example, time to collaborate and examine data (collaboration) could be of greater importance to collaboration as a factor than of simply having time is to the importance of resources as a factor.

Additionally, the second order factor structure of the model could be reviewed and re-conceptualized. As referenced earlier, there are inherent difficulties separating the evaluation capacity of the principal from that of the community. As such, on one level, it makes some sense to attempt a partition between individual and organizational factors but, on another level, a view of the components as covarying first order factors may be

more beneficial. It could be likely they cannot be separated as second-order factors given the complex interaction between leader and community, between the individual and the organization.

Engaging in evaluation activities to determine whether instructional practices and programs are having a positive impact on student outcomes is a critical school community practice, especially for communities where many students live in poverty. Though this investigation did not conclude with evidence that the ECAI could be used as an assessment of that capacity, evaluation capacity can still be increased within school communities engaging collaboratively, purposefully and intentionally in evaluation activities, using a variety of evaluation approaches and formats best suited for their important questions, their community, and their students.

APPENDIX A

Michigan Department of Education Program Evaluation Tool

Evaluation of strategies, programs, and initiatives to accelerate achievement and close achievement gaps is a key step in the continuous school improvement process. In addition, all federal programs (Title I Part A, C, and D; Title II, and Title III) require annual evaluation, especially when federal and/or state funds are used to support such efforts. More importantly, evaluation represents good practice and will likely improve outcomes. The Program Evaluation Tool can be used both during implementation to make mid-course corrections as well as following implementation to identify why results turned out as they did and how to improve implementation that will lead to increased student achievement.

Strategy / Program/ Initiative Description
What is the name of the strategy/program/initiative being evaluated?
(In addition to the name, identify whether it is a strategy, program, or initiative)
Provide a detailed description of the strategy/program/initiative being evaluated.
(Include population being served – number of students, grade, demographics, etc.; who is implementing; delivery model; frequency of intervention; start date; assessments used to measure objectives, etc.)
What is the need being addressed by the strategy/program/initiative?
(Include the gaps identified using baseline/subgroup data)
What is the reason for selecting the strategy/program/initiative including intended results?
(Include the connection to the need cited above and the SMART objective(s) identifying intended results)
Cite the research supporting the strategy/program/initiative, including a brief summary of research findings and targeted population.
(Research should be current and evidence-based with a brief summary)

1. Readiness: What is the readiness for implementing the strategy/program/initiative?

IN AN IDEAL STRATEGY/PROGRAM/INITIATIVE, stakeholders are well-prepared to implement the program. They have read and can articulate the research foundation, and regularly use the terms in conversation with each other, students, and with parents. Staff, students and parents express a high level of interest in, support for and commitment to the program. Specific concerns have been identified and solutions have been planned/ implemented. Staff is able to seamlessly integrate the program within the context of other building/district initiatives.

a) What is the evidence regarding stakeholder (staff/students/parents) understanding of the need as well as stakeholder ability to articulate the reason for the choice of the strategy/program/initiative?

- ☐Meeting agendas/minutes
- ☐Books/papers about the program
- ☐Staff surveys
- ☐SI Plan elements
- ☐Professional development materials
- ☐Conference/workshop attendance
- ☐Data collection plan; data analysis work
- ☐Stakeholder survey results
- ☐Suggestion box ideas collected
- ☐SI team agendas
- ☐Focus group interviews
- ☐Other

What does the evidence show regarding stakeholder (staff/ students/ parents) understanding of the need as well as stakeholder ability to articulate the reason for the choice of the strategy/program/initiative?

(Include conclusion, aligned to evidence, regarding stakeholder understanding of the need & the reasons for selecting the strategy/ program/ initiative)

b) What is the evidence regarding stakeholders (staff/students/parents) having a shared vision and strong commitment to the strategy/program/initiative?

- ☐Meeting agendas/minutes
- ☐Books/papers about the program
- ☐Staff surveys
- ☐SI Plan elements
- ☐Professional development materials

☐Conference/workshop attendance
☐Data collection plan; data analysis work
☐Stakeholder survey results
☐Suggestion box ideas collected
☐SI team agendas
☐Focus group interviews
☐Other

What does the evidence show regarding stakeholders (staff, parents, students) having a shared vision and strong commitment to the strategy/program/initiative?

(Include a conclusion, aligned to evidence, regarding stakeholders having a shared vision and a strong commitment to the strategy/program/initiative)

c) What is the evidence regarding how stakeholder (staff, parents, students) concerns were identified and addressed?

☐Meeting agendas/minutes
☐Books/papers about the program
☐Staff surveys
☐SI Plan elements
☐Professional development materials
☐Conference/workshop attendance
☐Data collection plan; data analysis work
☐Stakeholder survey results
☐Suggestion box ideas collected
☐SI team agendas
☐Focus group interviews
☐Other

What is the evidence regarding how stakeholder (staff, parents, students) concerns were identified and addressed?

(Include concerns of each stakeholder group and how they were addressed)

d) What is the evidence regarding the ability of staff and administrators to integrate the strategy / program/ initiative with existing work?

☐Meeting agendas/minutes
☐Books/papers about the program
☐Staff surveys
☐SI Plan elements
☐Professional development materials

☐Conference/workshop attendance
☐Data collection plan; data analysis work
☐Stakeholder survey results
☐Suggestion box ideas collected
☐SI team agendas
☐Focus group interviews
☐Other

What does the evidence show regarding the ability of staff and administrators to integrate the strategy/ program/ initiative with existing work?

(Explain how strategy/program/initiative fits into current work)

Given the evidence you've assembled, choose one overall self-assessment of the readiness for implementing the strategy/program/initiative.

(Align rating to evidence)

☐Stakeholders were fully prepared to implement
☐Support and commitment were generally high, but some concern or work remains
☐Some promising elements exist, but were mixed with major gaps in knowledge or confidence.
☐Interest and/or commitment were low

What action steps are needed to increase readiness to implement the strategy/program/initiative?

(Deduce action steps for READINESS from the evidence and rating)

2. Knowledge and Skills: Did the staff and administrators have the knowledge and skills to implement the strategy/program/initiative?

IN AN IDEAL STRATEGY/PROGRAM/INITIATIVE, personnel are able to clearly articulate what successful implementation looks and sounds like and how specific practices will change as a result of its implementation. Staff and administrators can articulate specific outcomes and specific criteria for evaluation. Personnel can demonstrate their ability to apply the knowledge and skills required to successfully implement with fidelity, and professional learning opportunities are provided to address gaps in knowledge and skills.

a) What is the evidence regarding staff and administrators' plan for how practice would change as a result of the strategy/program/initiative?

☐Minutes of professional conversations
☐Self-assessment checklists
☐Staff surveys
☐Superintendent or administrator observations/walkthroughs
☐Professional learning agendas, sign-in sheets
☐Program simulations, administrator observations
☐Other

What does the evidence show regarding staff and administrators' plan for how practice would change as a result of the strategy/ program/ initiative?

(Provide examples of how practice would change)

b) What is the evidence regarding administrator knowledge of and ability to monitor and assess the effectiveness of the strategy / program/ initiative?

☐Minutes of professional conversations
☐Self-assessment checklists
☐Staff surveys
☐Superintendent or administrator observations/walkthroughs
☐Professional learning agendas, sign-in sheets
☐Program simulations, administrator observations
☐Other

What does the evidence show regarding administrator knowledge of and ability to monitor and assess the effectiveness of the program/strategy/initiative?

(Cite how administrator's professional learning supported the monitoring and assessment of effectiveness)

c) What is the evidence regarding the sufficiency of opportunities for staff to learn the knowledge and skills identified as essential (the non-negotiable or acceptable variations of the elements) to the strategy/program/initiative?

☐Minutes of professional conversations
☐Self-assessment checklists
☐Staff surveys
☐Superintendent or administrator observations/walkthroughs
☐Professional learning agendas, sign-in sheets
☐Program simulations, administrator observations
☐Other

What does the evidence show regarding the sufficiency of opportunities for staff to learn the knowledge and skills identified as essential (the non-negotiable or acceptable variations of the elements) to the strategy/program/ initiative?

(Include evidence of initial professional learning. Address sufficiency and effectiveness of professional learning, including meeting identified learning outcomes)

d) What is the evidence regarding staff ability to apply the acquired knowledge and skills?

☐Minutes of professional conversations
☐Self-assessment checklists
☐Staff surveys
☐Superintendent or administrator observations/walkthroughs
☐Professional learning agendas, sign-in sheets
☐Program simulations, administrator observations
☐Other

What does the evidence show regarding staff ability to apply the acquired knowledge and skills?

(Include results drawn from quantifiable evidence of staff's knowledge/ability to implement strategy/program/initiative)

Given the evidence you've assembled, choose one overall self-assessment of the participants' knowledge and skills to implement the strategy/ program/ initiative.

(Align rating to evidence)

☐ Participants had sufficient knowledge and skills to succeed.

☐ Much knowledge and skill were evident, but few skills (or some knowledge bases) still need work. ☐ A solid start was documented, but many skill levels and much knowledge need to be acquired. ☐ Participants were beginning to acquire the necessary knowledge and skills.
What action steps are needed to improve participants' knowledge and skills?
(Deduce action steps for KNOWLEDGE AND SKILLS from the evidence and rating)

3. Opportunity: Was there opportunity for high quality implementation of the strategy/program/initiative?

IN AN IDEAL STRATEGY/PROGRAM/INITIATIVE, building and district administrators provide significant support for project implementation. Sufficient funds have been allocated and continue to be managed by building principal and or program director. Adequate resources are available for full implementation including time for staff collaboration in various forms. Clearly defined structures/protocols are in place to collect and review formative implementation data.

a) What is the evidence regarding the sufficiency of administrative support to achieve the intended results?

- ☐Agendas/minutes
- ☐Action plans
- ☐Email correspondence
- ☐Focus group and/or anonymous surveys
- ☐Budget sheets
- ☐Logs, school schedules
- ☐Inventories
- ☐Curriculum pacing guides
- ☐Collaboration models(such as PLCs, Collaborative Action Research, Lesson Study Teams)
- ☐Staff meeting results
- ☐Protocols for reviewing formative assessments
- ☐Other

What does the evidence show regarding the sufficiency of administrative support to achieve the intended results?

(Include specific examples of administrative support/lack of support and draw conclusions from examples you cited)

b) What is the evidence regarding the sufficiency of opportunities for on-going professional learning, including modeling and coaching?

- ☐Agendas/minutes
- ☐Action plans
- ☐Email correspondence
- ☐Focus group and/or anonymous surveys
- ☐Budget sheets
- ☐Logs, school schedules
- ☐Inventories
- ☐Curriculum pacing guides

☐Collaboration models (such as PLCs, Collaborative Action Research, Lesson Study Teams)
☐Staff meeting results
☐Protocols for reviewing formative assessments
☐Other

What does the evidence show regarding the sufficiency of opportunities for on-going professional learning, including modeling and coaching?

(*Include examples of opportunities/lack of opportunities for on-going professional learning, including modeling and coaching; draw conclusions from examples you cited)*

c) What is the evidence regarding the sufficiency of resources – including financial, time and personnel - to achieve the intended results?

☐Agendas/minutes
☐Action plans
☐Email correspondence
☐Focus group and/or anonymous surveys
☐Budget sheets
☐Logs, school schedules
☐Inventories
☐Curriculum pacing guides
☐Collaboration models (such as PLCs, Collaborative Action Research, Lesson Study Teams)
☐Staff meeting results
☐Protocols for reviewing formative assessments
☐Other

What does the evidence show regarding the sufficiency of resources – including financial, time, and personnel – to achieve the intended results?

(Include examples of resources/lack of resources and draw specific conclusions from examples you cited)

d) What is the evidence regarding the sufficiency of opportunities for staff collaboration to support implementation of the strategy/ program/ initiative?

☐Agendas/minutes
☐Action plans
☐Email correspondence
☐Focus group and/or anonymous surveys

☐Budget sheets
☐Logs, school schedules
☐Inventories
☐Curriculum pacing guides
☐Collaboration models (such as PLCs, Collaborative Action Research, Lesson Study Teams)
☐Staff meeting results
☐Protocols for reviewing formative assessments
☐Other

What does the evidence show regarding the sufficiency of opportunities for staff collaboration to support implementation of the strategy/ program/initiative?

(Include examples of staff collaboration/lack of collaboration supported by data and draw conclusions from examples you cited)

e) What is the evidence regarding structures being in place to collect and review implementation data?

☐Agendas/minutes
☐Action plans
☐Email correspondence
☐Focus group and/or anonymous surveys
☐Budget sheets
☐Logs, school schedules
☐Inventories
☐Curriculum pacing guides
☐Collaboration models (such as PLCs, Collaborative Action Research, Lesson Study Teams)
☐Staff meeting results
☐Protocols for reviewing formative assessments
☐Other

What does the evidence show regarding structures being in place to collect and review implementation data?

(Describe structures in place to collect and review implementation data; derive conclusions from structures/lack of structures to collect and review implementation data)

Given the evidence you've assembled, choose one overall self-assessment of the opportunity for high quality implementation.

(Align rating to evidence)

☐Necessary support and resources (time, funding, and attention) were solidly in place.

☐Many necessary resources were aligned with program goals, but more are needed.

☐Basic resources and opportunities were available, but significant gaps need to be filled.

☐Opportunityandresourceswerejustbeginningtoalignsupportoftheprogram.

What action steps are needed to ensure opportunity for high quality implementation?

(Deduce action steps for OPPORTUNITY from evidence and rating)

4. Implementation with Fidelity: Was the strategy/program/initiative being implemented as intended?

IN AN IDEAL STRATEGY/PROGRAM/INITIATIVE, all personnel involved in the program implement the strategies with fidelity according to the research, carrying out responsibilities by their proposed timelines. They use clearly defined protocols to collect and review formative implementation data to identify unintended consequences. Program leaders consider adjustments guided by implementation data while maintaining the integrity of results.

a) What is the evidence regarding a process being in place to monitor fidelity of implementation of the non-negotiable or acceptable variations of the elements of the strategy/program/initiative, including timelines and responsibilities?

☐Principal's walkthroughs
☐Number of staff implementing with fidelity
☐Model lessons
☐Surveys
☐Coaching schedule
☐Agendas and minutes of common planning time/meetings
☐Record of funds used
☐Lists of acquired resources
☐Focus group interviews
☐Debriefing following model lessons
☐Collegial observations/visits
☐Training agendas & material
☐Program Time Line
☐Other

What does the evidence show regarding the fidelity of implementation of the non-negotiable or acceptable variations of the elements of the strategy/program/initiative, including timelines and responsibilities?

(Provide specific evidence of a process to monitor fidelity of staff implementation of the strategy/program/initiative; draw specific conclusions regarding fidelity of implementation from examples)

b) What is the evidence regarding positive or negative unintended consequences that may have occurred, if any?

☐Principal's walkthroughs
☐Number of staff implementing with fidelity
☐Model lessons

☐Surveys
☐Coaching schedule
☐Agendas and minutes of common planning time/meetings
☐Record of funds used
☐Lists of acquired resources
☐Focus group interviews
☐Debriefing following model lessons
☐Collegial observations/visits
☐Training agendas & material
☐Program Time Line
☐Other

What does the evidence show regarding positive or negative unintended consequences that may have occurred, if any?

(Provide specific examples of positive and/or negative unintended consequences, and draw conclusions from the examples provided)

c) What do implementation data and student achievement results suggest for implementing/modifying the strategy/program/initiative?

☐Principal's walkthroughs
☐Number of staff implementing with fidelity
☐Model lessons
☐Surveys
☐Coaching schedule
☐Agendas and minutes of common planning time/meetings
☐Record of funds used
☐Lists of acquired resources
☐Focus group interviews
☐Debriefing following model lessons
☐Collegial observations/visits
☐Training agendas & material
☐Program Time Line
☐Other

How might these affect the integrity of the results?

(Include modifications made/being considered and discuss possible impact of modifications on the integrity of implementation)

Given the evidence you've assembled, choose one overall self-assessment of the fidelity of high quality implementation.

(Align rating to evidence)

☐All research-based elements have been implemented with fidelity following the proposed timelines.

☐Critical elements have been implemented, but work on consistency and depth remains. ☐The overall design was in place, but variations in practice were evident and may be adversely affecting results. ☐Parts of the program were working, but others have yet to be implemented.
What action steps are needed to ensure faithful implementation of program plans?
(Deduce action steps for FIDELITY OF IMPLEMENTATION from evidence and rating)

5. Impact: What was the impact of the strategy/program/initiative's on students?

IN AN IDEAL STRATEGY/PROGRAM/INITIATIVE, the school's achievement results on state or district wide assessments meet proficiency standards. Achievement gaps between each of the relevant subgroups and their counterparts have been narrowed as proposed in the School Improvement Plan's measurable objectives. Interim assessment results indicate progress toward proficiency for all students to the satisfaction of all stakeholders

a) What is the evidence and what does it show regarding achievement of the measurable objective for all students when compared to baseline state and local data?

(Include data sources aligned to measurable objectives for all students and draw conclusions from data)

b) What is the evidence and what does it show regarding achievement of the measurable objective for subgroups and their counterparts when compared to baseline state and local data?

(Include data sources aligned to objectives for each subgroup and draw conclusions from the data for each subgroup)

c) What is the evidence and what does it show regarding stakeholder (staff/students/parents) satisfaction with the results?

(List stakeholders involved, describe methods used to measure each stakeholder's satisfaction and specific data results for each stakeholder group)

d) Were the objectives for this strategy/program/initiative met?

☐ Yes
☐ No

Impact Conclusion
Should the strategy/program/initiative be continued or institutionalized?
☐ Yes ☐ No
a) What is the evidence and what does it say regarding whether this was the right strategy/program/ initiative to meet your needs?
(Provide conclusion relating data to identified need)
b) What is the evidence and what does it say regarding whether the benefits of the strategy/program/ initiative are sufficient to justify the resources it requires?
(Provide conclusion relating data to cost effectiveness)
c) What adjustments, if any, might increase its impact while maintaining its integrity?
(Discuss potential adjustments with rationale)
d) What is needed to maintain momentum?
(Discuss specific actions, resources, changes that will maintain momentum)
e) How might these results inform the School/District Improvement Plan?
(Identify how results will impact measurable objectives, strategies, and/or activities in the School Improvement Plan (SIP)/District Improvement Plan (DIP).

APPENDIX B

Revisions to items on the Evaluation Capacity Assessment Instrument (ECAI) (Taylor-Ritzler et.al., 2013)

Changes were made to the instrument in order to correspond with the 3 second order factor hypothesis and reflect K-12 school structures (as opposed to community service organizations.

Awareness (a)[1]: I think that evaluation…

1. Will help me understand my program, strategy or initiative.
2. Will inform the decisions I make about my program, strategy or initiative.
3. Will justify funding for my program, strategy or initiative.
4. Will help to convince others that changes are needed in my program, strategy or initiative.
5. Will inform changes in our documentation systems.
6. Is absolutely necessary to improve my program, strategy or initiative.
7. Should involve program, strategy or initiative participants in the evaluation process.
8. Will influence policy relevant to my program, strategy or initiative.
9. Will help improve services to students from diverse ethnic/racial backgrounds and/or students with disabilities.
10. Is unnecessary because we already know what is best for our students. [This gets recoded in analysis.]
11. Of a program, strategy or initiative is too complex to do. [This gets recoded in analysis.]

Motivation (m)[1]: I am motivated to…

12. Learn about program, strategy or initiative evaluation.
13. Evaluate my program, strategy or initiative.
14. Support staff to evaluate programs, strategies or initiatives.
15. Encourage others to buy into evaluating our program, strategy or initiative.

Competence (c)[1]: I know how to…

16. Develop a program, strategy or initiative evaluation plan.
17. Clearly state measurable goals and objectives for my program, strategy or initiative.
18. Identify strategies to collect information from participants.
19. Define outcome indicators of my program, strategy or initiative.
20. Decide what questions to answer in an evaluation.
21. Decide from whom to collect the information.
22. Collect evaluation information.
23. Analyze evaluation information.
24. Develop recommendations based on evaluation results.
25. Examine the impact of my program, strategy or initiative on students from diverse ethnic/racial backgrounds and/or students with disabilities.

26. Write an evaluation report.
27. Conduct an evaluation of my program, strategy or initiative.
28. Conduct an evaluation of my program with support from others.
29. Present evaluation findings orally.

Leadership (*l*)[1]:

30. District leaders provide effective leadership.
31. Staff understands how everyone's duties fit together as part of the overall mission of the program, strategy or initiative.
32. District leaders communicate program, strategy or initiative goals and objectives clearly.
33. District leaders have a clear plan for accomplishing program, strategy or initiative goals.
34. District leaders have realistic expectations of what staff can accomplish given the resources they have available.

Learning Climate (clim)[1]: The school where I work fosters an environment in which…

35. Program, strategy or initiative evaluation information is shared in open forums.
36. Staff is supported to introduce new approaches in the course of their work.
37. It is easy for staff to meet regularly to discuss issues.
38. Staff is provided opportunities to assess how well they are doing, what they can do better, and what is working.
39. Staff can encourage others to make use of evaluation findings.
40. Staff respects each other's perspectives and opinions.
41. Staff errors lead to teachable moments rather than criticisms.
42. Staff participates in making long-term plans for the program, strategy or initiative.
43. Staff concerns are considered in most decisions regarding strategic planning and evaluation.

Resources for Evaluation (r)[1]: In my school…

44. Resources are allocated to provide accommodations for people from diverse ethnic backgrounds and for people with disabilities to collect evaluation information (e.g. interpreters, translated documents).
45. Staff has time to conduct evaluation activities (e.g. identifying or developing a survey, collecting information from participants).
46. Staff has access to technology to compile information into computerized records.
47. Staff has access to adequate technology to produce summary reports of information collected from participants (e.g. computerized database).
48. Resources are allocated for staff training (e.g. money, time, bringing in consultants).
49. Technical assistance is available to staff to address questions related to evaluation.
50. Grant funders provide resources (e.g. training, money, etc.) to conduct evaluation.
51. Grant funders provide leadership for conducting evaluation.
52. District leadership engages in ongoing dialogue with grant funders regarding evaluation.

Mainstreaming (mstr)[1]:

53. My school gathers information from diverse stakeholders to gauge how well the program, strategy or initiative is doing.
54. My school has adequate records of past program, strategy or initiative evaluation efforts and what happened as a result.
55. I have access to the information I need to make decisions regarding my work.
56. I am able to integrate program, strategy or initiative evaluation activities into my daily work practices.
57. The evaluation activities I engage in are consistent with the State's expectations.

Use of Evaluation Findings (u)[2]: My school currently uses evaluation results…

58. To report to the State.
59. To improve programs, strategies or initiatives.
60. To get additional funding.
61. To design ongoing monitoring processes.
62. To assess implementation of a program, strategy or initiative.
63. To assess quality of a program, strategy or initiative.
64. To improve community engagement.
65. To make informed decisions.
66. As a component of staff training.
67. To develop best practices.
68. To eliminate unneeded programs, strategies or initiatives.

Notes:

1. Response format was a 1-4 scale, where 1 = *strongly disagree*; 2 = *somewhat disagree*; 3 = *somewhat agree*; and 4 = *strongly agree*.
2. Response format was a 1-4 scale, where 1 = *not at all*; 2 = *to some extent*; 3 = *to a considerable extent*; and 4 = *to a very great extent*.

REFERENCES

Ackoff, R.L. (1999). *Ackoff's Best.* New York: John Wiley & Sons, pp. 170-172.

Axxin, W.G., Link, C.F. & Groves, R.M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography, 48,* 1127-1149.

Baizerman, M., Compton, D.W. & Stockdill, S.H. (2002). New directions for ECB. *New Directions for Evaluation, 93*, 109-119.

Boyd, A. Geerling, T., Gregory, W.J., Kagan, C., Midgley, G., Murray, P. & Walsh, M.P. (2007). Systemic evaluation: a participative, multi-method approach. *Journal of the Operational Research Society, 58*, 1306-1320. http://dx.doi.org/10.1057/palgrave.jors.2602281

Boyle, R., & Lemaire, D. (1999). *Building effective evaluation capacity.* New Brunswick, NJ: Transaction Publishers.

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research:* Rand-McNally.

CEPI (2017). EEM Reports. Retrieved from https://cepi.state.mi.us/eem/reports.aspx

Farley-Ripple, E. & Buttram, J. (2015). The development of capacity for data use: The role of teacher networks in an elementary school. *Teachers College Record, 117,* 1-34.

Fierro, L. A. (2012). *Clarifying the connections: Evaluation capacity and intended outcomes* (Doctoral Dissertation). (Ascension Order No. AAI3545210). Retrieved from http://search.proquest.com.proxy.lib.wayne.edu/docview/1520338457?accountid=14925

Fitzpatrick, J., Sanders, J. & Worthen, B. (2011). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.) Upper Saddle River, NJ: Prentice Hall.

General Accounting Office. (2002). *Performance Budgeting: Opportunities and challenges*. (GAO-02-1106T). Retrieved from http://www.gao.gov/assets/110/109585.pdf

General Accounting Office. (2011). *Performance Measurement and Evaluation: Definitions and relationships.* (GAO-11-646SP). Retrieved from http://www.gao.gov/products/GAO-11-646SP

Guerra-López, I. (2007). Planning a responsive evaluation: Establishing solid partnerships by clarifying expectations and purpose, part 2. *Performance Improvement, 46*(8), 32-36.

Goertz, M., & Duffy, M. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory into Practice, 42*(1), 4-11.

Haeffele. L., Hood, L. & Feldman, B. (2011). Evaluation capacity building in a school-university partnership grant program. *Planning and Changing 42*(1/2), 887-100.

Hall, D. (2004). Preparing for the data deluge: NCLB and data-driven decision making require educators not only to collect assessments but to be able to manage them. Getting ready for the challenge can be the hardest part. *Learning & Leading with Technology*, *32*(2), 32.

Hess, F.M. & Kelly, A.P. (2007). Learning to lead: what gets taught in principal-preparation programs. *Teachers College Record, 109*(1), 244-274

Huffman, D., Lawrence, F., Thomas, K. & Clarkson, L. (2006). Collaborative evaluation communities in urban schools: A model of evaluation capacity building for STEM education. *New Directions for Evaluation, 119,* 93-113.

Hanwright, J., & Makinson, S. (2008). Promoting evaluation culture: The development and implementation of an evaluation strategy in the Queensland Department of Education, Training and the Arts. *Evaluation Journal of Australasia, 8*(1), 20-25.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55. http://dx.doi.org/10.1080/10705519909540118

Huber, E. & Harvey, M. (2015). An analysis of internally funded learning and teaching project evaluation in higher education. *International Journal of Educational Management, 30*(5), 606-621. http://dx.doi.org/10.1108/IJEM-08-2014-0108

Janssens, F.J. & van Amelsvoort, G.H. (2008). School self-evaluations and school inspections in Europe: An exploratory study. *Studies in Educational Evaluation, 34*(1), 15-23. https://doi.org/10.1016/j.stueduc.2008.01.002

Kahan, B. (2008). Review of evaluation frameworks. *Kael Consulting, Saskatchewan Ministry of Education (March)*. Retrieved from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C23&q=kahan+review+of+evaluation+frameworks&btnG=

King, J.A. & Volkov, B. (2005). A framework for building evaluation capacity based on the experiences of three organizations. *CURA* [Center for Urban and Regional Affairs] *Reporter, 35*(3), 10-16.

Kline, R.B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Labaree, D. (1997). Public Goods, Private Goods: The American struggle over educational goals. *American Educational Research Journal, 34*(1), 39-81.

Lavinghouze, S.R. & Snyder, K. (2013). Developing your evaluation plans: A critical component of public health program infrastructure. *American Journal of Health Education, 44*(4), 237-243. http://dx.doi.org/10.1080/19325037.2013.798216

Lee, L.E. (1999). Building capacity for school improvement through evaluation: experiences of the Manitoba School Improvement Program Inc. *The Canadian Journal of Program Evaluation Special Issue*. 155-178.

Lord, F.M. (1953). On the Statistical Treatment of Football Numbers. *American Psychologist, 8*(12), 750-751.

Mandinach, E.B. & Gummer, E.S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher, 42*(1), 30-37.

Mandinach, E.B., Friedman, J.M. & Gummer, E.S. (2015). How can schools of education help to build educators' capacity to use data? A systemic view of the issue. *Teachers College Record 117*, 1-50.

MDE (2016). Program Evaluation Tool Retrieved from http://www.michigan.gov/mde/0,4615,7-140-6530_30334_51051-328384--,00.html

MDE (2018). District and School Accountability file. Retrieved from http://www.michigan.gov/mde/0,4615,7-140-81376_59490---,00.html

Micerri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Mills, L. (2006). Transforming data into knowledge. *Principal Leadership, 7*(2), 44-48.

MISchool Data (2017). Free and Reduced Lunch Counts. Retrieved from https://www.mischooldata.org/Other/DataFiles/StudentCounts/HistoricalFreeAndReducedLunchCounts.aspx

Mulford, B. & Silins, H. (2011). Revised models and conceptualisation of successful school principalship for improved student outcomes. *The International Journal of Educational Management, 25*(1), 61-82.

No Child Left Behind Act. (2001). Improving the academic achievement of the disadvantaged, section 101. Retrieved from https://www.congress.gov/bill/107th-congress/house-bill/1/text

Oliver, M., MacBean, J., Conole, G. & Harvey, J. (2002). Using a toolkit to support the evaluation of learning. *Journal of Computer Assisted Learning, 18*( ), 199-208.

Owen, J. M. (2007). *Program evaluation: Forms and approaches* (3rd ed.). New York: The Guilford Press.

Rose, S.A., Markman, B., & Sawilowsky, S. (2017). Limitations in the systematic analysis of structural equation model fit indices. *Journal of Modern Applied Statistical Methods*, *16*(1), 69-85. http://dx.doi.org/10.22237/jmasm/1493597040

Sawilowsky, S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research. 60*(1), 91-126.

Sawilowsky, S. & Blair, R. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360.

Schildkamp, K. & Teddlie, C. (2008). School performance feedback systems in the USA and in the Netherlands: a comparison. *Educational Research and Evaluation, 14*(3), 255-282. http://dx.doi.org/10.1080/13803610802048874

Schildkamp, K., Visscher, A. & Luyten, H. (2009). The effects of the use of a school self-evaluation instrument. *School Effectiveness and School Improvement 20*(1), 69-88. http://dx.doi.org/10.1080/09243450802605506

Shadish, W., Cook, T. & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*: Wadsworth.

Shen, J., Cooley, V.E., reeves, P., Burt, W.L., Ryan, L., Rainey, J.M. & Yuan, W. (2010). Using data for decision-making: perspectives from 16 principals in Michigan, USA. 56, 435-456. http://dx.doi.org/10.1007/s11159-010-9172-x

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51(2), 161-179.*

Stufflebeam, D.L. & Shinkfield, A.J. (2007). *Evaluation theory, models, & applications.* Sam Francisco, CA: John Wiley & Sons, Inc.

Taylor-Ritzler, T., Suarez-Balcazar, Y., Garcia-Iriarte, E., Henry, D.B., & Balcazar, F.E. (2013). Understanding and Measuring Evaluation Capacity: A Model and Instrument Validation Study. *American Journal of Evaluation, 34*(2), 190-206.

Vanhoof, J., Verhaeghe, G., Verhaeghe, J.P., Valcke, M. & Van Petegem, P. (2011). The influence of competences and support on school performance feedback use. *Educational Studies, 37*(2), 141-154. http://dx.doi.org/10.1080/03055698.2010.482771

Wayman, J.C., Jimerson, J.B. & Cho, V. (2012). Organizational considerations in establishing the data-informed district. *School Effectiveness and School Improvement, 23*(2), 159-178. http://dx.doi.org.proxy.lib.wayne.edu/10.1080/09243453.2011.652124

Yeagley, R. (2003). The demands of data under NCLB: procrastinating your collection and compilation could lead to a district's doom. *School Administrator*, *60*(11), 22.

Zumbo, B.D. & Zimmerman, D.W. (1993). Postscript: Some closing comments on “alternatives to classical statistics” (normal theory). *Canadian Psychology, 34*(4), 441-445.

ABSTRACT

RELIABILITY AND VALIDITY OF AN EVALUATION CAPACITY ASSESSMENT INSTRUMENT FOR PRINCIPALS OF PUBLIC SCHOOLS IN MICHIGAN AND THE CURRENT DISTRIBUTION OF EVALUATION CAPACITY BY POVERTY CONCENTRATION

by

CHRISTINE LEWIS

December 2019

Advisor: Dr. Shlomo S. Sawilowsky

Major: Education Evaluation and Research

Degree: Doctor of Philosophy

This study used Confirmatory Factor Analysis (CFA) to investigate the reliability and validity of an Evaluation Capacity Assessment Instrument (ECAI) for principals of kindergarten to twelfth grade public schools in Michigan. The ECAI was then used to investigate whether scores between principals of high poverty schools and low poverty schools were significantly different. The ECAI was administered to one hundred and twenty-one principals across Michigan. Based on the results of the CFA, evidence of fit to the *a priori* model was not established but internal consistency statistics provide some reliability evidence. A Principle Components Analysis (PCA) was conducted and factor loadings were examined. Some groups of ECAI items loaded onto the *a priori* factors of the model while others did not. The scores between principals of high poverty and low poverty schools were not significantly different however, as model fit was not established, questions regarding the model and questions regarding any significant difference between the two poverty groups remain additional topics for further research.

AUTOBIOGRAPHICAL STATEMENT

PROFESSIONAL EXPERIENCES

Warren Consolidated Schools - Warren, Michigan

SEPTEMBER 2001 TO PRESENT

Chief Human Resources Officer **2019 - PRESENT**

Primary Responsibilities:

- Labor Relations
- Talent Management
- Recruitment
- Employment, Benefits and Attendance Office Oversight

Administrator of State and Federal Programs **2011 - 2019**

Primary Responsibilities:

- Grants
- School Improvement

Interim Principal – Middle School **2013**

Mathematics and Music Curriculum Specialist, K-12 **2009 - 2011**

Mathematics Teacher - High School **2001 - 2009**

Wayne State University - Detroit, Michigan

SEPTEMBER 2016 TO 2018

Adjunct Faculty – School of Education **2016 - 2018**

EDUCATION AND CERTIFICATION

- ***Ph.D.***

Education Evaluation and Research: Quantitative Methods **ANTICIPATED 2019**
Wayne State University, Detroit, Michigan

- ***Administrative Certificate – State of Michigan (current)***

Endorsements in Elementary & Secondary Administration K-12 (ES) and Central Office (AC)
2011

- ***M.Ed.***

Educational Leadership: Curriculum, Instruction and Leadership **2008**
Saginaw Valley State University, University Center, Michigan

- ***Teaching Certificate – State of Michigan (current)***

Secondary, with Mathematics (EX) major and Sociology (CF) minor **2000**