1-1-2012

Miblsi program evaluuation of participatory elementary schools from 2003-20009

Marvin Gibbs
Wayne State University,

MiBLSi PROGRAM EVALUATION OF PARTICIPATORY ELEMENTARY SCHOOLS FROM 2003-2009

by

MARVIN GIBBS

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

2012

MAJOR: EVALUATION & RESEARCH

Approved by:

_____

Advisor Date

_____

_____

_____

DEDICATION

The source of my inspiration:

My Parents

My Daughter and Son-in-Law

My Granddaughter and Grandson

Family and Friends

ACKNOWLEDGMENTS

This dissertation could not have been written without Dr. Shlomo Sawilowsky who not only served as my supervisor but also encouraged and challenged me throughout my academic program. He and the other faculty members, Dr. Gail Fahoome, Dr. Monte E. Piliawsky, and Dr. Lee Warshay, guided me through the dissertation process, never accepting less than my best efforts. I thank them all.

This paper would not have been possible without the support and patience of Dr. Anna Harms and the other MiBLSi codirectors who encouraged me to narrow my focus on the analysis and the checklist for this dissertation. I am also grateful to Jennifer Barrington who kept me motivated, grounded, and laughing as I prepared for my defense of this project. In addition, last but not the least because they are equally as important as anyone, I acknowledge my friends and family who without their encouragement and devotion; I never could have written this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I – INTRODUCTION

Evaluation Approaches

According to Stufflebeam (1994), developer of the Context, Input, Process, Product (CIPP) Model of curriculum evaluation, evaluators have more efficacious evaluation approaches available than ever before:

> Following a period of relative inactivity in the 1950s, a succession of international and national forces stimulated the development of evaluation theory and practice. Main influences were the efforts to vastly strengthen the U.S. defense system spawned by the Soviet Union's 1957 launching of Sputnik I; the new U.S. laws in the 1960s to equitably serve persons with disabilities and minorities; the federal evaluation requirements of the Great Society programs initiated in 1965; the U.S. movement begun in the 1970s to hold educational and social organizations accountable for both prudent use of resources and achievement of objectives; the stress on excellence in the 1980s as a means of increasing U.S. international competitiveness; and the trend in the 1990s for various organizations, both inside and outside the U.S., to employ evaluation to assure quality, competitiveness, and equity in delivering services. (Stufflebeam, 1999)

Seeking "reforms, American society has repeatedly pressed [educational entities], healthcare organizations, and various social welfare enterprises to show whether services and improvement efforts are succeeding" (Stufflebeam, 2001b). The pursuit to reform, which led to the study of alternative evaluations, is "important for professionalizing program evaluation and for its scientific advancement and operation. Professionally, careful study of program evaluation approaches can help evaluators to legitimize methods that conform to sound principles of evaluation and discredit those that do not" (Stufflebeam, Madaus, & Kellaghan, 2000).

Standards for Educational Evaluation

The Joint Committee on Standards for Educational Evaluation (JCSEE) published "three sets of standards for educational evaluations: *Personnel Evaluation Standards* was first published in 1988, *Program Evaluation Standards* (second edition) was published in 1994, and *Student Evaluations Standards* was published in 2003. Each publication presents and expands on

a set of standards for use in an assortment of educational settings. In addition, the standards provide guidelines for designing, implementing, assessing, and improving the identified form of evaluation. JCSEE placed each of the standards in one of four fundamental categories to promote evaluations that are proper, useful, feasible, and accurate". They are as follows:

The Personnel Evaluation Standards

- The Propriety Standards require that evaluations be conducted legally, ethically, and with due regard for the welfare of evaluates and clients involved.
- The Utility Standards are intended to guide evaluations so that they will be informative, timely, and influential.
- The Feasibility Standards call for evaluation systems that are as easy to implement as possible, efficient in their use of time and resources, adequately funded, and viable from a number of other standpoints.
- The Accuracy Standards require that the obtained information be technically accurate and that conclusions be linked logically to the data. (JCSEE, 1988)

The Program Evaluation Standards

- The Utility Standards are intended to ensure that an evaluation will serve the information needs of intended users.
- The Feasibility Standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.
- The Propriety Standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.
- The Accuracy Standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated. (JCSEE, 1994)

The Student Evaluation Standards

- The Propriety Standards help ensure that student evaluations are conducted lawfully, ethically, and with regard to the rights of students and other persons affected by student evaluation.
- The Utility Standards promote the design and implementation of informative, timely, and useful student evaluations.
- The Feasibility Standards help ensure that student evaluations are practical; viable; cost-effective; and culturally, socially, and politically appropriate.
- The Accuracy Standards help ensure that student evaluations will provide sound, accurate, and credible information about student learning and performance. (JCSEE, 2003)

A MiBLSi Evaluation Study

The Individuals with Disabilities Act (IDEA) regulation 34 CFR 300.307 requires a state policy for determining Learning Disabilities (LD) that cannot require the discrepancy model, which refers to differences between IQ and performance or achievement. According to LaPointe and Heinzelman (2006), the regulations also include a Response to Intervention (RTI) approach to evaluation based on the student's ongoing "response to scientific, research-based intervention (34 CFR 300.309(a)[2][I]). RTI is a multitiered approach to help struggling learners, with students' closely monitored at each stage of intervention to determine the need for additional research-based instruction and/or intervention. State Michigan's Integrated Behavior and Learning Support Initiative (MiBLSi)" policy is beginning to reflect these provisions. Some districts have developed local MiBLSi/RTI policy and procedure that follows the IDEA and answers demands from the No Child Left Behind Act (NCLB) for higher levels of student literacy.

Although state policy is not completely implemented, the Office of Special Education and Early Intervention Services (OSE-EIS) supports local systemic development of MiBLSi/RTI through approximately 240 MIBLSI pilot projects. Support includes a significant amount of MiBLSi/RTI support for hosting of state and national level technical assistance, Internet-based data collection and connection to national research projects, state/regional networking, presentation of demonstration projects, and limited funding to support conference participation and release time, travel, etc. "Two focus areas are reading Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and behavior (SWIS)" (Heinzelman, LaPointe, & Vanderploeg, 2010).

Background Fact Pattern of MiBLSi

"Educational equity can be framed in terms of both equal opportunities and outcomes

including both the contexts in which students participate in educational experiences and the extent to which those experiences enable their academic growth" (Nieto, as cited in DeVanenzuela, Copeland, Huaqing, & Park, 2006, p. 425). The "disproportionate representation of minority students in special education has long been a concern in discussions of educational equity. These concerns relate to potential inequities in both educational opportunities and outcomes resulting from ineffective education. Disproportionate representation may also differentially diminish the opportunities of students identified with a disability to interact with teachers and others within the larger school context, especially when educated in segregated settings. This disproportionate representation has been a cause for suspicion of the use of PL-94-142" (Education of All Handicapped Children Act), the predecessor of the Individuals with Disabilities Education Act.

For example, the Florida Department of Education reported that the "percentage of Black students in classes for educable mentally retarded pupils has exceeded the generally accepted 2% expected level" (Peterz, 1999). According to Harry and Klingner (2007), African American students "across the United States are represented in the category of Educable Mental Retardation at twice the rate of their White peers. In the category of Emotional/Behavioral Disorders, they are represented at 1.5 times the rate of their White peers. And, in some states, Native American and Hispanic students are overrepresented in the Learning Disability category."

The foundation of this dilemma lies inherent in U.S. history. Emanating as an adjacent to the civil rights movement, special education emerged to dispel the inequities of those being denied a higher quality educational experience. However, despite the many educational reforms, disparities in referrals to special education presented a relationship between school integration and special education. "Looking at how the mandate for school integration intertwined with

special education, scholars analyzed public documents and newspaper articles dating from Brown v. Board of Education in 1954 to the inception of the Education for All Handicapped Children Act in 1975" (Connor & Ferri, 2005). Their findings highlight how "African American students entering public schools through forced integration were subject to low expectations and intense efforts to keep them separate from the White mainstream.

As the provision of services for students with disabilities became a legal mandate, clear patterns of overrepresentation of Mexican American and African American students in special education programs became apparent" (Connor & Ferri, 2005). "Plagued by ambiguous definitions and subjectivity in clinical judgments, these categories often had more to do with administrative, curricular, and instructional decisions than with students' inherent abilities" (Harry & Klingner, 2007). According to the Reauthorization of the Individuals with Disability Act 2004 Regulations Findings,

> Greater efforts are needed to prevent the intensification of problems connected with mislabeling and high dropout rates among minority children with disabilities. More minority children continue to be served in special education than would be expected from the percentage of minority students in the general school population. (Wright & Wright, 2007)

With the continuing support of MiBLSi Program Evaluation of Participatory Elementary Schools, educational institutions are learning to implement a school culture in which teachers are able to enhance academic success and behavior in a cohesive setting. It is important that student progress is monitored frequently to help make decisions about modifications in instruction or academic goals, thus allowing data to drive instruction as well as other educational decisions. "A school-wide support model provides the foundations for using prevention and intervention strategies for identified academic and/or behavioral problems" (Michigan Department of Education, n.d.). Although MiBLSi is not a research study/project but rather a state professional

development grant, schools officials value its ability to ”develop support systems and sustained implementation of data-driven, problem-solving” models that provide students with the strategies to become better readers, in addition to the social skills necessary for success (Michigan Department of Education, n.d.). Institutions must realize that a poorly designed academic and behavioral program and the implementation of it can lead to the stagnation of educators, the classroom setting, and their pupils.

MiBLSi conducted a 2-year study to “evaluate the implementation of programs assisted under this title and the impact of such programs on improving the academic achievement of children with disabilities. [In addition, it analyzed program effectiveness to enable a child’s ability to] reach challenging state academic content standards based on state academic assessments” (Harms, 2010).. The study was implemented in schools under typical conditions with existing staff and is continuously evolving. These findings were collected and shared throughout the study.

The conceptual framework included planned intervention, student outcomes, and actual implementation. There were 485 participating schools that were in collaboration with 45 independent school districts (ISDs). The unit of analysis was whole-school building: team-based self-assessment of implementation fidelity and aggregated student data. The measures of student performance were (a) reading--Dynamic Indicators of Basic Early literary Skills (DIBELS) and Michigan Educational Assessment Program (MEAP) and (b) behavior--Office Discipline Referral (Harms, 2010).

Discovering what works regarding improving the academic achievement of children with disabilities “does not solve the problem of program effectiveness. Once models and best practices are identified, practitioners are faced with the challenge of implementing programs

properly. A poorly implemented program can lead to failure as easily as a poorly designed one" (Mihalic, Irwin, Fagan, Ballard, & Elliott, 2004). Evaluator and client bias must be a concern during an evaluation process. Any evaluation study is going to be biased to some extent. The decisions that evaluators make about what to examine, what methods and instruments to use, and with whom to talk all influence the outcome of the evaluation. Evaluators'personal backgrounds, biases, professional training, and experience affect the way the study is conducted. Both evaluators and clients must be concerned about evaluation bias—evaluators because their personal standards and professional reputations are at stake, and clients because they do not want to invest (either politically or financially) in findings that are off the target (Fitzpatrick, Sanders, & Worthen, 2004).

Purpose of the Study

The purpose of this study is to show that Stufflebeam's (1999) Program Evaluations Metaevaluation Checklist (PEMC) can be used to determine the extent to which each of the JCSEE standards were included in the evaluation, "So, How Are We Doing? A Michigan Integrated Behavior Learning Support Initiative (MiBLSi) Evaluation Study." IDEA 2004 requires a national assessment to evaluate the implementation of programs funded under this title, as well as assess the impact of such programs on improving the academic achievement of children with disabilities to enable the children to reach challenging state academic content standards. Based on this perspective, the present paper reveals how the PEMC has been applied to the MiBLSi Program Evaluation of Participatory Elementary Schools from 2003-2009, a required program by the reauthorization of IDEA 2004. This evaluation was accomplished by describing MiBLSi, identifying participants in the MiBLSi Participatory Elementary Schools from 2003-2009, and describing their roles during and subsequent to the evaluation, using the

PEMC and metaevaluation model to assess the extent to which the programs evaluation of the MiBLSi Participatory Elementary Schools program met the evaluation standards established by Stufflebeam (1999).

Research Questions

The following four questions are addressed in this dissertation because they serve as the focus of this researcher's metaevaluation investigation.

1. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the utility standard developed by the JCSEE?
2. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the feasibility standard developed by the JCSEE?
3. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the propriety standard developed by the JCSEE?
4. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the accuracy standard developed by the JCSEE?

The psychometric properties of the Program Evaluations Metaevaluation Checklist will be assessed using the Cronbach's alpha and Kruskal-Wallis one-way analysis of variance by ranks (Siegel & Castellan, 1988).

Definition of Terms

The following terms are defined specifically for this study:

Learning Disability – As defined by the Individual with Disabilities Education Act (IDEA), a learning disability is "a disorder in one or more basic psychological processes involved in understanding or in using language, spoken or written, that may manifest itself in an imperfect ability to listen, speak, read, write, spell, or do mathematical calculations, including

conditions such as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia" 34 Code of Federal Regulations § 300.7(c) (10).

Special Education – The educational system for students with special needs that addresses the students' individual differences and needs.

Program Evaluation – Evaluation "means a study designed and conducted to assist some audience to assess an object's merit and worth. This definition should be widely acceptable since it agrees with common dictionary definitions of evaluation; it is also consistent with the definition that underlies published sets of professional standards for evaluations" (JCSEE, 1994).

MiBLSi – Michigan's Integrated Behavior and Learning Support Initiative (MiBLSi) school-wide multitiered system for Response to Intervention (RTI). This program offers three different examples of how RTI is improving outcomes for students in Michigan.

Response to Intervention (RTI) – "RTI is a multitiered approach to help struggling learners. Students' progress is closely monitored at each stage of intervention to determine the need for further research-based instruction and/or intervention in general education, in special education, or both" (The National Center for Learning Disabilities, Inc, 2012).

Assumptions

The researcher assumes that the evaluation, "So, How Are We Doing? A Michigan Integrated Behavior Learning Support Initiative (MiBLSi) Evaluation Study" was conducted following JCSEE standards and the information contained in the evaluation is accurate.

Limitations of the Study

The following limitations are acknowledged for this study. These limitations may reduce the generalizability of the study to populations outside of these parameters.

1. The study is limited to the evaluation of individual evaluation designs, studies and

reports.

2. The study is restricted to the metaevaluation of the "So, How Are We Doing? The MiBLSi Evaluation Study."
3. Schools in cohorts 4.3, 4.5, and 6 will be excluded, middle schools will be excluded, and schools with a whole set of missing data will be excluded.
4. The study is limited to revisiting the evaluator's assessments and understanding how successes and failures were explained.
5. The study is limited to the probability that a number of problems highlighted in the reviewed reports have already been acted upon, something which was outside the scope of this metaevaluation.

CHAPTER II – LITERATURE REVIEW

Part of a Larger Study

Reauthorization of Individuals with Disabilities Act (IDEA) became effective October 13, 2006. This act incorporated new requirements for identifying students with learning disabilities (LDs) and allowed districts to consider a child's "*response to scientific, research-based intervention*" as part of evaluation process. A subsection of the act, §300.309(a)(2)(i), was shortened to *response to intervention,* or RTI. House and Senate committee reports were concerned with severe discrepancy models and wanted to distinguish "more accurately between students who truly have LDs from those whose learning difficulties could be resolved with more specific, scientifically based, general education interventions." The IQ-achievement discrepancy model for LD included a description of "an educationally significant discrepancy between estimated intellectual potential and actual level of performance related to basic disorders in the learning processes" (Bateman, as cited in Kavale, n.d., p. 2). The President's Commission on Excellence in Special Education report also recommended RTI. Implementing a school-wide model for student success was conceptualized using this multitiered framework across the behavior and reading domains.

Metaevaluation Purpose

Because the quality of Michigan's Integrated Behavior and Learning Support Initiative (MiBLSi) can impact the education of youth, it is important that its evaluation be accurate and unbiased. Because metaevaluation is the evaluation of an evaluation, the need to conduct metaevaluations to ensure such evaluations are valid is important to the welfare of consumers (Cooksy, 1999). Scriven (1969) wrote that metaevaluation "is the methodological assessment of evaluation and is the concern with the evaluation of specific evaluative performance" (p. 36).

The purpose of this review, therefore, is to focus on metaevaluation as it is applied to the MiBLSi Program Evaluation of Participatory Elementary Schools from 2003-2009 that included cohorts. Cohorts are a group of Michigan Participating MiBLSi Schools who have shared a particular time together during a particular time span. Participating MiBLSi Schools include Cohort 1, Cohort 2, Cohort 3, Cohort 4.1, Cohort 4.2, Cohort 4.3, Cohort 4.4, and Cohort 5.

To understand these standards for this metaevaluation, the quality and robustness of the evaluation process should include the following:

1. "The American Evaluation Association has created a set of *Guiding Principles for Evaluators* (2004). The order of these principles does not imply priority among them; priority will vary by situation and evaluator role. The principles are as follows:

    - *Systematic Inquiry*: Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.
    - *Competence*: Evaluators provide competent performance to stakeholders.
    - *Integrity /Honesty*: Evaluators ensure the honesty and integrity of the entire evaluation process.
    - *Respect for People*: Evaluators respect the security, dignity and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.
    - *Responsibilities for General and Public Welfare*: Evaluators articulate and consider the diversity of interests and values that may be related to the general and public welfare". (pp. 5-6)

2. The JCSEE (2011) "has developed standards for program, personnel, and student evaluation. The Joint Committee standards are broken into four sections: utility, feasibility, propriety, and accuracy. They provide guidelines about basing value judgments on systematic inquiry, evaluator competence and integrity", respect, and regard for public welfare.

As the JCSEE is the benchmark, its origins and achievements must be reviewed. **"**In 1974, the committee jointly appointed by American Educational Research Association (AERA),

American Psychological Association (APA), and National Council on Measurement in Education (NCME) completed its revision of the 1966 edition of *Standards for Educational and Psychological Tests and Manuals*, published by American Psychological Association" (JCSEE, 1994). This committee felt that inclusion of the section on evaluation and test standards lay beyond its own area of authority or responsibility and recommended "creation of another committee to address this issue. The three organizations, therefore, appointed another committee that met for first time in the autumn of 1975, with the membership extending across 12 national organizations with an interest in the quality of evaluation in education" (JCSEE 1994).

The first edition of *The Program Evaluation Standards* was published by the JCSEE in 1981. The second edition from 1994 is the one that is referred to in this paper. The first edition of *The Personnel Evaluation Standards* was published by JCSEE in 1988. This edition was revised in 2008. The *Student Evaluation Standards* (2003) primarily addressed internal, everyday evaluations conducted by teachers in elementary and secondary schools. These standards were the result of the rigorous process of elaboration and testing which drew on the input and contributions of sources including the panel of writers, review panels, field test sites, public hearings, and the validation panel, and were subjected to periodical reviews that allowed them to constantly incorporate technical and scientific advances and to respond to new challenges and requirements emerging in field of evaluation

The Nature, Structure, and Importance of Evaluation Standards

Standards provide a framework of reference for defining good practice in evaluation (JCSEE, 2011). The standards of practice are based principles upon which professionals in the field have reached consensus. These principles, when observed, provide assurances of the quality of evaluation and suitable professional practice; they are not, however, an instruction manual or

list of specific technical standards or rules to be applied mechanically. The principles articulate a set of guidelines that may not be equally important or applicable in all situations, and which may even, in certain situations, conflict with one another. The evaluation process includes an evaluator qualification standard that is important for good quality evaluation. This particular standard is generally accepted as an essential condition in the decision plan. Consequently, although the standards represent the state-of-the-art in research in evaluation and contribute to improving quality, they do not of themselves guarantee this quality.

These standards can be divided into four categories. The three sets of standards, which largely share the same attributes across three primary domains of evaluation practice--personnel, student, program--are predicated on four major categories for high-quality evaluation: (a) propriety, (b) utility, (c) feasibility, and (d) accuracy. The four categories of evaluation standards are expression of the unified and consistent perspective on educational evaluation, and they maintain a right to the joint analysis of three publications of JCSEE. Although directed at different audiences, evaluators should consider using the four categories of evaluation standards together. “There is no shortage of examples of interpetration of evaluation of learning, evaluation of performance of education professionals, and evaluation of programs” (JCSEE, 2003).

“Although they were developed in the [United States] and are based on ideas, laws, [respective education systems], and circumstances, these standards articulate the practical philosophy of evaluation which has gained universal acceptance in the western world–with promotion and safeguarding of quality of educational services as the ultimate objective. The first two standards presented and examined by the JCSEE for evaluations were those relating to propriety and utility of evaluation” (2003). The aim of the propriety standard is to ensure that the evaluation is conducted ethically and legally, with respect for the well-being of all those who are

involved in and affected by the evaluation. Service orientation is a key concept used to provide satisfaction of students' educational needs, and by extension, community and society. "Conflicts of interest" must therefore be avoided or managed in such a fashion that the evaluator is independent and impartial, neither benefiting nor suffering from any result which the evaluation may produce (JCSEE, 2003).

The aim of the utility standards is to ensure that the evaluation is applied in a clear and timely fashion (providing the response to needs for information of users), and as the constructive guideline which informs recommendation, planning (including implementation), supervision, and evaluation of followup actions designed to consolidate or develop strengths, while eliminating, correcting, or improving weaknesses–"impact of evaluation" (JCSEE, 2003).

The feasibility standards are designed to ensure that political and material conditions exist for implementation of the evaluation as intended. This requires diplomacy and procedures that do not interfere with educational activity, are practical/practicable and can mobilize necessary resources (JCSEE, 2003).

The accuracy standards address production of reliable and representative information that permits valid interpretations, justified conclusions, and appropriate followup actions. In this context, "metaevaluation" is of prime importance. Each group of standards addresses an essential aspect of evaluation. Each of these aspects is strongly interdependent, meaning all aspects must be taken into consideration in each particular evaluation (JCSEE, 2003).

Checklists

A checklist is a series of items or tasks that need to be accomplished. Checklists have been used for medicine, education, business, aviation, and other purposes to help guide a project to success (formative evaluations) or determine the merit or worth of a project (summative

evaluations; Stufflebeam, 2001a). Checklists can be used for a wide variety of evaluations: program, personnel, and product, as well as providing criteria and guidance for metaevaluations and systems of evaluations.

A checklist includes factors, properties, aspects, components, criteria, tasks, or dimensions that are needed to complete a task. The order and extent to which each of these components is included are considered separately (Scriven, 2007). While checklists differ to type and purpose, they all have a common function–being a mnemonic device. As professional evaluations require a systematic approach to assess the value, worth, merit, etc., the availability of a checklist of the required components is invaluable for program evaluations. Scriven (2007) listed the reasons that checklists are used for evaluations:

1. Checklists are mnemonic devices that minimize the probability that an important element in an evaluation will be forgotten. There is a direct reduction of errors of omission and indirect reduction of errors of commission.
2. Lay stakeholders are better able to understand and verify checklists than complex theories or statistical analyses.
3. Checklists reduce the halo effect (i.e., the overvaluing of a highly-valued component of the evaluation). Checklists accomplish this task by requiring the evaluator to consider each component separately, and allocating the worth of the component appropriately.
4. Checklists force evaluators to make specific judgments about each component and draw conclusions based on their judgments.
5. Checklists consolidate large amounts of information about a program that is going to be evaluated in an economical format. Checklists are a form of knowledge about a

domain that is specific to the evaluation and is designed to perform certain tasks (e.g., personnel, overall evaluation).

6. Checklists can provide evaluations with improved reliability and validity, add credibility to an evaluation, as well as provide useful knowledge about the program being evaluated.

Scriven (2007) asserted that checklists should meet the following standards:

1. The checkpoints should refer to criteria and not mere indicators.
2. The list should be complete (no significant omissions).
3. The items should be contiguous (i.e., nonoverlapping--essential if the list is used for scoring).
4. The criteria should be commensurable.
5. The criteria should be clear (comprehensible, applicable).
6. The list should be concise (to assist its mnemonic function; i.e., it should contain no superfluous criteria).
7. The criteria should be confirmable (e.g., measurable or reliably inferrable).

According to Scriven (2007), the use of evaluation checklists is important to assess and characterize the general merit, worth, or importance of the program being evaluated. One difficulty in evaluating specific components of a program is assigning weights. Scriven argued that equal weighting should be used unless there is overwhelming evidence that a criterion has greater or less merit than another criterion. Establishing weights must be done with caution and must be done consistently across the entire criterion. He asserted that consistent ratings are a better way and have less inherent bias than providing weights to certain criteria in the evaluation.

Program Evaluation Models Metaevaluation Checklist

This checklist is for performing metaevaluations of program evaluation models. It is organized according to the Joint Committee Program Evaluation Standards. For each of the 30 standards, the checklist includes 10 checkpoints drawn from the substance of the standard. It is suggested that each standard be scored on each checkpoint. Then, judgments about the adequacy of the subject evaluated (evaluation model) can be made as follows: 0-*2 Poor, 3-4 Fair, 5-6 Good, 7-8 Very Good, 9-10 Excellent.* It is recommended that an evaluation model be failed if it scores *Poor* on standards P1 Service Orientation, A5 Valid Information, A10 Justified Conclusions, or A11 Impartial Reporting. Figure 1 presents the metaevaluation standard model. Stufflebeam (1999) advised users of this checklist to consult the full text of JCSEE (1994).

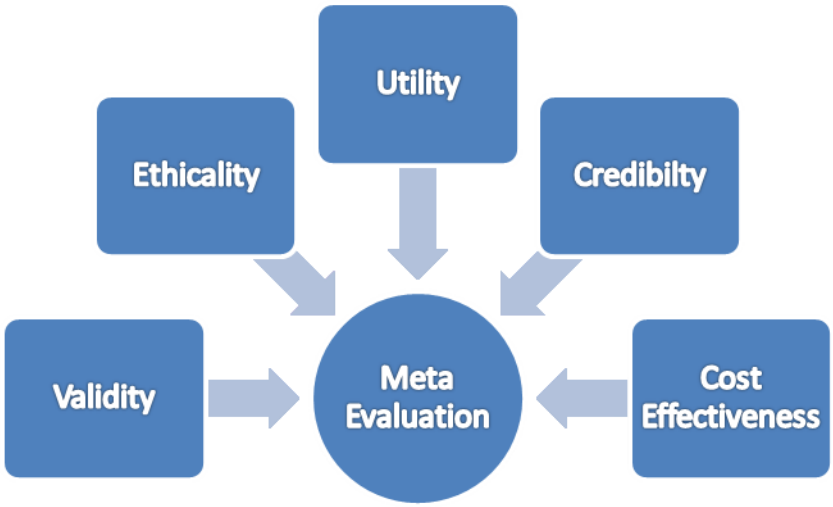

Figure 1. Metaevaluation standard model (Sinjindawong, Lawthong, & Kajanawasee, n.d.).

The indicators are as follows:

Validity means the evaluation should be a managed document, context analyzed, defined evaluation goal, and designed evaluation, so that evaluation can be verified accurately and quality of collection, analysis, interpretation and conclusion, can be divided into 13 indicators, as follows: Va1.1 Context Identification Va1.2 Prominent Identification Va1.3 Described Purpose Va1.4 Evaluation Design Va1.5 Analysis of Document Sources Va1.6 Reliable Information Sources Va1.7 Verifiable of Information Va1.8 Quality of Information Va1.9 Systematic Data Analysis Va1.10 Justified Interpretations and Conclusions Va1.11 Disclose Positive and Negative Evaluation Report Va1.12 Fair Evaluation Results Va1.13 Verifiable Evaluation results
Utility means the evaluation that will be useful for stakeholders and the others.The evaluation can be judged, reported clearly, disseminated in time, and guided for improving plan, with 10 indicators, as follows: Ut2.1 Stakeholder Identification Ut2.2 Period and Timeline Identification Ut2.3 Collecting Data Technique Ut2.4 Actual Evaluation Judgment Ut2.5 Useful Evaluation Results Ut2.6 Format of Evaluation Report Ut2.7 Clarified Evaluation Report Ut2.8 Comprehensible Evaluation Report Ut2.9 Report in Time Ut2.10 Dissemination of Evaluation Report
Ethicality means the evaluation should be a suitable set of assessment procedures for realistic situations and can be considered for many groups of humans. Evaluation can be a mean of continuous improvement by considering protection of human rights and utilization of public standards of conduct that evaluate completely and fairly for participants, in addition, disclosure of evaluation results, with 9 indicators, as follows: Et3.1 Assessment Communication Et3.2 Acceptation of Evaluation Results Et3.3 Continuous Improvement for Evaluation Quality Et3.4 Formal Agreements Et3.5 Disclosure and Limitation of Evaluation Et3.6 Protection of Human Rights Et3.7 Divergent Human Interaction Et3.8 Complete and Fair Assessment

Et3.9 Assessment according to the Standards
Credibility means the evaluation should be by competent assessors and no conflict of interest that can be underminded and contracdict reliable findings and information, with 4 indicators, as follows: Cr4.1 Evaluator Competence Cr4.2 Communication skills of Evaluators Cr4.3 Evaluation Management Cr4.4 Conflict of Interest
Cost-Effectiveness means the evaluation should be considered its worth. Credibilityneedsresources for assessment and cost accountability, which have 2 indicators, as follows: Ce5.1 Resources Management of Evaluation Ce5.2 Budget Accountability

Figure 2. Program evaluation indicators (Sinjindawong, Lawthong, & Kajanawasee, n.d.).

Metaevaluation by Scriven, Stufflebeam, Wingate, and Other Evaluators

Wingate (2009) stated that the range of purposes for metaevaluation may be put into four distinct categories, but a single metaevaluation may serve multiple purposes. She noted that this configuration was her typology and other evaluators may use a greater or fewer categories. The four categories are formative evaluation (Fitzpatrick et al., 2004; Scriven, 1975; Stufflebeam, 2001b), summative evaluation (Patton, 2008; Stufflebeam, 2001b; Stufflebeam & Shinkfield, 2007), synthesis (Cooksy & Caracelli, 2005; Scott-Little, Hamann, & Jurs, 2002; Stufflebeam & Shinkfield, 2007), and research (Fitzpatrick et al., 2004; Stufflebeam & Shinkfield, 2007).

Scriven (1991) described metaevaluation in his assessment of the thesaurus: "meta-evaluation is to assess indirectly, estimates by experts and represents a scientific and ethical obligation, in the best interests of the other part" (p. 228). He added that metaevaluation should be performed by the verifier and on an external object. Stufflebeam (2007) emphasized the difference between the active metaevaluation that was designed to be assessed by experts and metaevaluation that used audience judges to evaluate a program.

Stufflebeam (1999) created a metaevaluation checklist based on *The Program Evaluation Standards* (JCSEE, 1994), which are as follows:

The Program Evaluation Standards (Joint Committee, 1994)	
Utility Standards: The utility standards are intended to ensure that an evaluation will serve the informational needs of intended users.	
Ul Stakeholder	Persons involved in or affected by the evaluation should be identified, so that their needs can be addressed.
U2 Evaluator Credibilty	The persons conducting the evaluation should be both trustworthy and credibility competent to perform the evaluation, so that the evaluation findings achieve maximum credibility and acceptance.
U3 Information Scope and Selection	Scope Information collected should be broadly selected to address pertinent questions and selection about the program and be responsive to the needs and interests of clients and other specified stakeholders.
U4 Values Identification	The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear.
U5 Report Clarity	Evaluation reports should clearly describe the program being evaluated, including its context, and the purposes, procedures, and findings of the evaluation, so that essential information is provided and easily understood.
U6 Report: Timeliness and Dissemination	Significant interim findings and evaluation reports should be disseminated to intended users, so that they can be used in a timely fashion.
U7 Evaluation Impact	Evaluations should be planned, conducted, and reported in ways that encourage follow-through by stakeholders, so that the likelihood that the evaluation will be used is increased.
Feasibility Standards: The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.	
Fl Practical Procedures	The evaluation procedures should be practical to keep disruption to a minimum while procedures needing information is obtained.
F2 Political Viabliltiy	The evaluation should be planned and conducted with anticipation of different viability positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

The Program Evaluation Standards (Joint Committee, 1994)	
F3 Cost Effectiveness	The evaluation should be efficient and produce information of sufficient value, so that effective resource expenditure can be justified.
Propriety Standards: The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.	
PI Service Orientation	Evaluations should be designed to assist organizations to address and effectively serve the needs of the full range of targeted participants.
P2 Formal Areement	Obligations of the formal parties to an evaluation (what is to be done, how, by whom, and when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it.
P3 Rights of Human Subjects	Evaluations should be designed and conducted to respect and protect the rights and welfare of human subjects.
P4 Human Interactions	Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation, so that participants are not threatened or harmed.
P5 Complete and Fair Assesment	The evaluation should be complete and fair in its examination and recordings of strengths and weaknesses of the program being evaluated can be built upon and problem areas addressed.
P6 Disclosure of Findings	The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation and any others with expressed legal rights to receive the results.
P7 Conflict of Interest	It should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results.
P8 Fiscal Responsibility	The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible, so that expenditures are accounted for and appropriate.
Accuracy Standards: The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about die features that determine worth or merit of the program being evaluated.	
Al Program Documentation	The program being evaluated should be described and documented clearly and accurately, so that the program is clearly identified.

The Program Evaluation Standards (Joint Committee, 1994)	
A2 Context Analysis	The context in which the program exists should be examined in enough detail, so that its likely influences on the program can be identified.
A3 Described Purposes and Procedure	The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assessed.
A4 Defensible Information Sources	The information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed.
A5 Valid Information	The information-gathering procedures should be chosen or developed and then implemented so that they will assure that the interpretation arrived at is valid for the intended use.
A6 Reliable Information	The information-gathering procedures should be chosen or developed and then implemented so that they will assure that the information obtained is sufficiently reliable for the intended use.
A7 Systematic Information	The information collected, processed, and reported in an evaluation should be systematically reviewed, and any errors found should be corrected.
A8 Analysis of Quantitative Information	An evaluation should be appropriately and systematically analyzed so that evaluation questions are effectively answered.
A9 Analysis of Qualitative Information	An evaluation should be appropriate and analyzed so that evaluation questions are effectively answered.
A10 Justified Conclusions	The conclusions reached in an evaluation should be explicitly justified, so that stakeholders can assess them.
All Impartial Reporting	Reporting procedures should guard against distortion caused by personal feelings and biases of any party to the evaluation, so that evaluation reports fairly reflect the evaluation findings.
A12 Metaevaluation	The evaluation itself should be formatively and summatively evaluated against these and other pertinent standards, so that its conduct is appropriately guided and, on completion, stakeholders can closely examine its strengths and weaknesses.

Figure 3. *Program Evaluation Standards* (JCSEE, 1994).

The *Program Evaluation Standards* have certification from the American National Standards through the American National Standards Institute (ANSI), which requires that the standards it certifies are developed in accordance with "essential requirements for due process" (American National Standards Institute, 2009, p. 4). Guidelines for metaevaluation using evaluation standards were found during the evaluation of the literature. Patton (1997) suggested questions should focus on the metaevaluation: "Was there an assessment well done? Have the evaluator applied professional assessment standards and principles?" (p. 193). Similarly, Scriven (1991) argued that metaevaluation can be either formative or final and aided by the use of checklists or standards such as program evaluation standards (JCSEE, 1994). The JCSEE (1994) stipulated that "the self-assessment and summative evaluation design should be on these and other relevant standards, so that its conduct was appropriate advice and, after completion, participants can examine in detail the strengths and weaknesses" (p. 185). Stufflebeam and Shinkfield (2007) supported the increased use of metaevaluation, noting that both formative assessment, final assessment, and metaevaluation provided descriptive and subjective assessment information for the assessment guide and presented strengths and weaknesses. They also described the structure of the metaevaluation process, on which standards for the evaluation of the program were based.

Stufflebeam and Shinkfield (2007) discussed the increased use of metaevaluation in detail: "Proactive metaevaluation is necessary to focus the expert, design, budget, and contract and carry out sound evaluations. Retrospective metaevaluation is needed to judge how the audience concluded the assessments." (p. 82). The two types of metaevaluations, formative metaevaluation and final metaevaluation, were highlighted in accordance with the frequent association of the metaevaluation standards and evaluation of program standards. In their

chapter, the two types were used to evaluate the assessments and explain how standards were used to improve assessment practices (Fitzpatrick et al., 2004). Prior to publishing the standards, however, noted evaluators Nilsson and Hogben (1983) commented on the need for the ratings for both evaluation of the specific research programs as well as for the entire evaluation.

Henry and Mark (2003) also "preferred the broader influence and offered a framework for representing how evaluation affects various program changes and ultimately leads to social betterment. Their distinction of levels of influence as being between intra- and interpersonal change processes brings up a consideration absent from Kirkhart's three dimensions of source, intention, and time". Kirkhart (as cited in Cummings, 2002) indicated that the source of influence is change at the starting point of a process and sources can either be a part of the evaluation process or a result of the evaluation. The second dimension, intention of the influence, is defined as "the extent to which evaluation influence is purposefully directed, consciously recognized and planfully anticipated" (as cited in Cummings, 2002, p. 4). The time of influence in Kirkhart's model is the timing of the influence, categorized into three levels: (a) immediate (during the study), (b) end of cycle, and (c) long-term. These three dimensions provide an integrated theory of influence that can occur at the level of the individual or at the level of more than one interacting individual. Henry and Mark (2003) "argued that any evaluation has anticipated outcomes and that mapping influence through the individual, interpersonal, and collective levels can trace change all the way from the evaluation to the policy level". Henry and Mark's taxonomy, "drawing from several bodies of literature in social science disciplines, categorizes evaluation influence into three levels, each of which has several change processes representing what evaluation influence could look like in any given context. Their 47 levels of influence offer a menu from which the evaluator or the researcher may select in order to cater a

theory of influence to a particular situation". Figure 4 depicts how their levels of influence break down into levels and menu items:

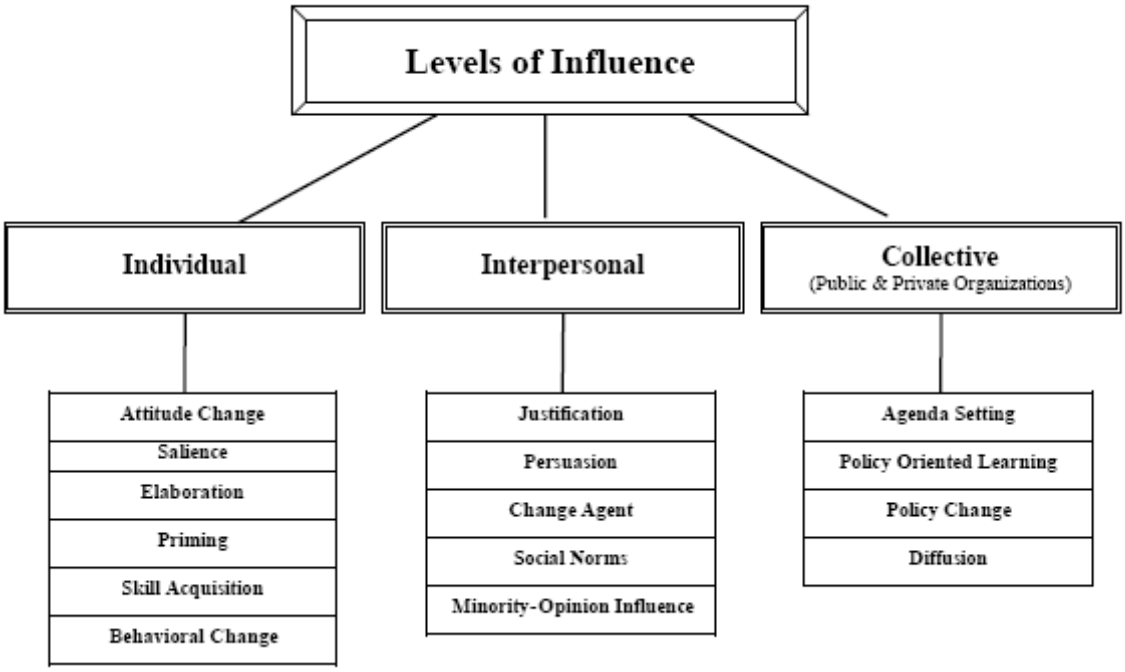

Figure 4. Mechanisms through which evaluation produces influences (Mark & Henry, 2004).

Although Mark and Henry (2004, p. 36) did not address metaevaluation specifically but rather discussed an integrated program of basic research and applied research, they believed that metaevaluations provided an excellent opportunity to collect data for research support. Despite the interest in the metaevaluation and opportunities for research support raised by Mark and Henry (2004), several examples of metaevaluations appear in the analysis of the literature. One of the earliest examples of metaevaluation was the set of articles that were critical of corporate training programs evaluations (Burt & Celotto, 1992; Ives, 1992; Jinkerson, Cummings, Neisendorf, & Schwandt, 1992; Niemiec, Sikorski, Clark, & Walberg, 1992). The intent of this series of articles was to examine existing evaluation processes and determine the strengths and

weaknesses inherent in these processes and provide suggestions to improve the process. The Advanced Technology Education (ATE) provided funding for four external metaevaluation programs (Hanssen, Lawrenz, & Dunet, 2008). The metaevaluators (Gullickson, Wingate, Lawrenz, & Coryn, as cited in Hanssen et al., 2008) provided suggestions on methods to improve formative evaluation processes. The metaevaluation process was validated and concerns regarding the evaluation were addressed during this process. As part of the program evaluation, the quality of the process used was assessed rather than providing suggestions to improve the outcomes of the evaluation (Hanssen et al., 2008).

Several studies in literature argue in support of the use of metaevaluation, yet they are more meta-analytic in nature as compared to metaevaluation (Ashworth, Cebulla, Greenberg, & Walker, 2004; Woodside & Sakai, 2001). Metaevaluation assesses the dignity and worth of the assessment while the meta-analysis is a quantitative synthesis of research related to the general question (Stufflebeam & Shinkfield, 2007). In response to the lack of published metaevaluations presented in the *American Journal of Evaluation*, a new section in 1999, assessment of assessments that presented a metaevaluation and the efforts of the tax base (Cooksy, 1999; Grasso, 1999) was presented. The section's aim was to improve assessment practices and show the usefulness of metaevaluation (Cooksy, 1999). However, this section was short-lived.

Scott-Little, Hamann, and Jurs (2002) "described the use of metaevaluation of after school programs with the Program Evaluation Standards. They showed that this type of study is an important mechanism for the results, and documentation strengthens procedural knowledge. They described a metaevaluation of Teach for America, a teacher evaluation system that used 10 other metaevaluations as guidelines for conducting a metaevaluation". Thus, the evaluations

could improve the program and help develop knowledge in the field and improve the implementation of metaevaluation, assessment, and evaluation practices.

Calls for greater use of metaevaluation also are found in the literature. For example, Fitzpatrick et al. (2004) confirmed the absence of metaevaluation in the literature and recommended its use to improve evaluation practices. Stufflebeam and Shinkfield (2007) emphasized the importance of metaevaluation and provided concrete proposals on how metaevaluations can produce the valuable results. At the same time, selective use of metaevaluation is recommended. For example, "metaevaluation gave the parties an independent assessment of evaluation but stated it would not be cost effective for all grades" (Patton, 1997). If disagreements or political unrest results from the measurement outcomes, an independent metaevaluation could provide evidence that the evaluation is important. In the same vein, Patton recommended carrying out the decision on a metaevaluation because the benefits sometimes outweigh the costs. Although the case of concurrent metaevaluation described here has not been politicized in the classical sense, there are reasons for the methodology of metaevaluation of support services as a model for a series of evaluations. The developers wanted to establish a method that met the highest standards. In addition, the cost of metaevaluation was small when compared with the costs of testing (Patton, 1997).

Criticism presented here provides a retrospective study of the process of concurrent metaevaluation. Not surprisingly, this type of metaevaluation has both strengths and weaknesses (Cousins & Shulha, 2006). The methods used to conduct a metaevaluation are briefly described, and then the strengths and weaknesses of this approach are discussed.

Decision-oriented evaluations provide a knowledge and value base for making and defending conclusions. Encouraging the use of evaluation to plan and implement needed

programs helps justify decisions about plans and actions. Necessary collaboration between evaluator and decisionmaker provides opportunity to bias results. Policy studies broader issues as well as identifies and assesses potential costs and benefits of competing policies. It provides general direction for broadly focused actions and is often corrupted or subverted by politically-motivated actions of participants.

Consumer-oriented generalized needs and values affect and judge the relative merits of alternative goods and services. They provide independent appraisals to protect practitioners and consumers from shoddy products and services. High public credibility might not help practitioners do a better job; however, evaluation methods require credible and competent evaluators. Accreditation/certification standards and guidelines determine if institutions, programs, and personnel should be approved to perform specified functions. They help the public to make informed decisions about quality of organizations and qualifications of personnel. Standards and guidelines typically emphasize intrinsic criteria to the exclusion of outcome measures.

Concurrent metaevaluation. Metaevaluations reported in the literature, although rare, often have focused on retrospective assessment of completed evaluations. Conducting a metaevaluation concurrently with the evaluation modifies this approach. This method provides the opportunity for the metaevaluators to advise evaluators and provides the basis for a summative judgment about the quality of the evaluation. The authors conducted a concurrent metaevaluation of a new evaluation technique being developed by a federal governmental agency; the new evaluation technique was expected to be highly visible and widely applied. The differences between concurrent metaevaluation and other metaevaluations were continuous involvement, attendance at data collection events, and external verification of the evaluation

data. The author's experience conducting the concurrent metaevaluation is described and challenges are discussed in this critique. The author concluded that concurrent metaevaluation holds promise for improving the practice of evaluation and of metaevaluation (Hersey et al., 2010). Fetterman and Wandersman (2007) reported,

> An important element of this metaevaluation was to verify the data analysis. Often, metaevaluators collect contact with project participants their views on valuation, but it is unusual to have a separate, parallel bar quite sure that the evaluation results are accurate and reproducible process. Although some estimates are used to replicate data provided promising technique of metaassessment as tools to evaluate this aspect, it is not confirmed and these quantitative results were not as useful for metaevaluation reporting. The idea of verifying the information and decisions during the evaluation process is an important potential role for concurrent metaevaluation made in the future, but this aspect should be carefully designed to provide useful data if necessary. The idea promises of concurrent metaevaluation is to improve ratings because the traditional metaevaluation is done after completion of the initial evaluation is conducted. It is useful to have an idea of how assessments can be improved, and to provide information for future assessment practice. (p.180)

Metaevaluation Reliability

According toWingate (2009),

> Professional evaluation rests on the premise that evaluation is a systematic endeavor. The Standards represent a major effort toward making evaluation practice more systematic. There are at least two important underlying assumptions embodied within the Standards: (1) Adherence to the Standards will produce higher quality evaluations (i.e., evaluations that are more useful, feasible, ethical, and accurate) and (2) Similar judgments about the quality of a given evaluation will be reached by different individuals when using the Standards as criteria of merit. Both assumptions are worthy of empirical investigation, but it is the latter one that is investigated in this dissertation. Reliability is a necessary but not sufficient condition for validity. (p. 7)

Measurement that rests largely with human judgment increases the potential for error greatly. In the context of metaevaluation, the Program Evaluation Standards serve as a common set of criteria against which to measure the quality of program evaluations. However, the standards do not constitute a precise measuring instrument but serve more as a heuristic device to facilitate

analysis and judgment. Metaevaluators' interpretation and application of the standards may be mediated by numerous factors that are unrelated to the actual quality of the evaluation being assessed. Such factors may include an evaluator's previous experiences with similar programs, paradigmatic predilections, conscious or unconscious biases, and technical expertise. The endeavor of professionalizing evaluation has been focused, in part, on increasing rigor to militate against inherent threats to validity in program evaluation practice while also enhancing the usefulness of evaluations. Achieving reliability is a strong defense against reaching erroneous, invalid conclusions (Stemler, 2007).

A MiBLSi Evaluation Study

IDEA regulation 34 CFR 300.307 requires a state policy for determining LDs that cannot require the discrepancy model. The regulations also include a RTI approach to evaluation based on the student's ongoing response to scientific, research-based intervention (34 CFR 300.309(a)[2][I]). Michigan's MiBLSi policy is beginning to reflect these provisions. Some districts have developed local MiBLSi/RTI policy and procedure that follow the IDEA and answer demands from the NCLB for higher levels of student literacy (LaPointe & Heinzelman, 2006).

MiBLSi is the Michigan Department of Education initiative that works with schools to develop the multitiered system of support for both reading and behavior. The MiBLSi program does this by providing professional development and technical assistance to building leadership teams with coaching support. The mission of MiBLSi is to develop support systems and sustained implementation of the data-driven, problem-solving model in schools to help students become better readers with social skills necessary for success. Although state policy is not completely implemented, the Office of Special Education and Early Intervention Services (OSE-

EIS) supports local systemic development of MiBLSi/RTI through approximately 240 MIBLSI pilot projects. A substantial amount of MiBLSi/RTI support for schools include hosting state and national level technical assistance, providing internet-based data collection, connecting to national research projects; networking on a state/regional basis, presenting demonstration projects, and providing limited funding to support conference participation, release time, travel, etc. The two areas on which the MiBBLSi concentrates are Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and the School-Wide information System (SWIS; Heinzelman et al., 2010).

Integrating Response to Intervention and Cognitive Assessment Methods

IDEA was reauthorized by the U.S. Congress in 2004, yet ongoing regulatory efforts are required to monitor its operation and implementation. Of particular concern to school psychologists and others involved in the educational process are guidelines for identifying children with specific learning disabilities (SLD). Two seemingly opposite camps have been arguing for either an RTI approach for SLD identification or a methodology that includes comprehensive evaluations for SLD identification and intervention purposes. The authors of IDEA proposed a resolution to these important issues by emphasizing a multitiered approach to serving children with learning problems–one that begins with RTI but then provides for comprehensive evaluation of cognitive processes once RTI methods are determined to successful in ameliorating the child's learning difficulties. If a child fails to respond to intervention and demonstrates a deficit in the basic psychological processes following comprehensive evaluation, both the definitional criteria for SLD and the method for determining SLD eligibility can be addressed. This methodology integrates the best aspects of both the RTI and comprehensive evaluation perspectives to forge a balanced practice model that ensures diagnostic accuracy and

optimizes educational outcomes for children with SLD (Hale, Kaufman, Naglien, & Kavale, 2006).

Why an Integrated Approach to Behavior and Reading?

Emerging research provides evidence to suggest that there are benefits to an integrated school-wide approach to supporting all students. Models of integrated behavior and reading supports produce larger gains in literacy skills than the reading-only model (Stewart, Benner, Martella, & Marchand-Martella, 2007). Improving students' social behavior can result in more minutes spent on academic instruction (Putnam, Handler & O'Leary-Zonarich, 2003; Putnam, Handler, Rey, & O'Leary-Zonarich, 2002). High quality instruction engages students and leads to reduction of problem behavior (Preciado, Horner, Baker, 2009; Sanford, 2006). Students who experience difficulty with reading may have found ways to escape or avoid reading activities (McIntosh, Horner, Chard, Dickey, & Braun, 2008). Additionally, similarities in supports for behavior and reading are implemented at the school level. Both are similar in their use of (a) a continuum of support; (b) action planning guided by a team; (c) the problem solving process (e.g., identification of need based on data); (d) the use of data for program development, progress monitoring, and evaluation; and (e) reliance on evidence-based practices.

Schoolwide, effective reading support can involve a three-tiered approach to prevention and intervention for reading problems in schools. The approach involves team-based training in strategies to prevent reading problems and support children with intense reading problems, as well as assimilate valuable academic and instructional systems. Important features of this approach include strong comprehensive, research-based initial instruction that addresses the needs of most students; a valid assessment system that includes screening and progress monitoring**;** and high quality, intensive interventions for struggling readers.

Positive behaviorial interventions and supports (PBIS) is

> a proactive, team-based process for creating and sustaining safe and effective schools. PBIS is a systems framework that improves the capacity of schools to educate all children using research-based schoolwide and classroom interventions. An emphasis is placed on preventing the occurrence of problem behavior as well as the use of data-based problem solving for addressing existing behavior concerns…. In order to effectively implement a problem-solving model, information must be collected and used continuously to evaluate and improve the systems of supports. (Michigan Department of Education, n.d.)

Continuum of Support

According to Sugai and Horner (2002), the fundamental goal for any educational practice (and support system) is the development of students who are competent in academics and social skills. The interaction that takes place between teacher and students within the classroom should be the main focus on implementation structures at all educational levels, including school, district, and state. The most important question is, "Does the program make a difference for students over time and across settings?" MiBLSi is in the ongoing process of creating a sustainable and scalable statewide system of support. The following figure describes this system at each level of implementation (Sugai & Horner, 2002).

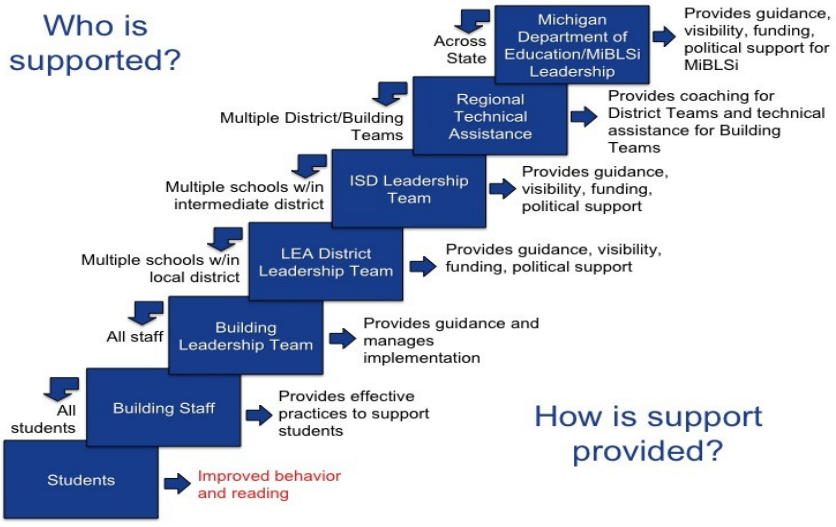

Figure 5. MiBLSi systems of support (Alpena-Montmorency-Alcona Educational Service District, 2011).

Figure 5 presents the importance placed on practices and supporting structures. As the structures move further away from direct student instruction, less emphasis is placed on the idiosyncratic aspects of the educational practice and more emphasis is placed on the infrastructure to support the implementation of the practice. The implementation drivers are integrated into the supporting infrastructure to ensure fidelity of implementation that is sustainable (Sugai & Horner, 2002).

Figure 6 illustrates the continuum of support that is being designed in the MiBLSi Statewide Structure of Support, laying out the levels and the support provided.

MiBLSi Statewide Structure of Support		
Level	How is support provided?	Who is supported?
Michigan Department of Education/MiBLSi Leadership	Across state	Provides guidance, visibility, funding, political support for MiBLSi
Regional Technical Assistance	Multipe District/Building Teams	Provides coaching for District Teams and technical assistance for Building Teams
District/Regional Leadership Team	Multiple schools within local or intermediate district	Provides guidance, visibility, funding, political support
Building Team Leadership	All staff	Provides guidance and manages implementation
Building Staff	All students	Provides effective practices to support students
Students		Improved behavior and reading

Figure 6. MiBLSi statewide structure of support (Alpena-Montmorency-Alcona Educational Service District, 2011).

Through the application process, school teams participate with MiBLSi for a period of 3 years (Sugai & Horner, 2002). During this time, school teams receive professional development through a training sequence that focuses on school-wide implementation of behavior and reading supports. Principals and coaches participate in meetings on specific topics regarding

implementation efforts. Technical assistance is provided by regional facilitators, with schools teams and coaches attending state conferences for implementation.

After the 3-year participation period has ended, continued support is provided through the following:

- Technical assistance provided by regional facilitators.
- The development of regional trainers who can assist in professional development for new as well as existing staff.
- Coaches and principal meetings that are scheduled throughout the school year for problem solving and as an implementation support network.
- Development of new materials (tools, manuals, training PowerPoints/handouts) and revision of earlier materials made available through the MiBLSi website.
- Future development of webinars on implementation topics.
- Focused training topics available for staff (registration fee).
- The development of Intermediate School District (ISD) support structure for implementation problem solving and support.
- Participation in State Coaches Conference (registration fee).
- Participation in State Implementer's Conference (registration fee).

MiBLSi Model

The MiBLSi is an integrated model of behavior and reading support. The practices are provided by staff to improve student outcomes. The systems are structures created to provide staff with support in implementing successful practices. Information is used for successful decision making, identifying appropriate (evidence-based) practices that meet student need,

evaluation of implementation efforts (Sugai & Horner, 2002), and student outcomes as a result of the practices.

Schoolwide Evaluation Tool (SET)

Description of measures. The Schoolwide Evaluation Tool (SET) is designed to assess and evaluate the critical features of schoolwide effective behavior support across each academic school year**.** The SET results are used to "evaluate the current status of schoolwide PBIS and to assist school teams to strengthen schoolwide behavior supports" (Michigan Department of Education, n.d.). Also, the SET is designed to assess and evaluate the important features of schoolwide effective behavior support across each academic school year**.** The SET results are used to accomplish the following:

1. Assess features that are in place,
2. Determine annual goals for schoolwide effective behavior support,
3. Evaluate ongoing efforts toward schoolwide behavior support,
4. Design and revise procedures as needed, and
5. Compare efforts toward schoolwide effective behavior support from year to year.

Information necessary for this assessment tool is collected through multiple sources including reviews of permanent products, classroom observations, and staff (minimum of 10) and student (minimum of 15) interviews or surveys. Multiple steps are used for collecting the necessary information. The first step is to identify teacher or staff member at the school as the contact person. This person will collect each of the available products and to identify a time for the SET data collector to preview the products and set up observations and interview/survey opportunities. Once the process for collecting the necessary data are established, reviewing the data and scoring the SET averages takes 2 to 3 hours. Results of the SET can provide schools

with a measure of the proportion of features that are (a) not targeted or started, (b) in the planning phase, and (c) in the implementation/ maintenance phases of development toward a systems approach to schoolwide effective behavior support. The SET is designed to provide trend lines measuring improvement and sustainability over time (Sugai, Lewis-Palmer, Todd, & Horner, 2001).

In their publication titled *The Reading Evaluator: A How to Manual for Success,* Hasbrouck and Denton (2005) recounted that the reading evaluator is somebody who works competently with participatory elementary school professionals to advance their abilities in teaching reading to students. A reading evaluator guides school staff to conceive their reading dream and assist by turning dreams into the reality through implementation method (Shanklin, 2006). In essence, reading advisers can assist educators to persist in fulfilling participatory elementary school enhancement by using research-based data-driven practices to boost literacy for all students.

To further analyze what the evaluator's function is, one should realize anticipated conclusions or goals of a productive reader and evaluator. One main objective of the evaluator is to work with schools to advance scholastic aptitude. This was accomplished by "heavy" advising. Joellen Killion (2008), Deputy Executive Director, NSDC, stated that two kinds of coaching exist: "coaching light" and "coaching heavy." She advocated for coaching heavy and asserting that coaches who coach heavy typically are extending their skills, subject area knowledge, leadership skills, interpersonal relation skills, and instructional strategies. In a similar manner, Killion argued that coaches challenge themselves and present teachers with appropriate challenges to encourage them to develop an enhanced sense of professionalism and

improved effectiveness. Killion continued that using these challenges, a greater sense of collaborative responsibility for every student's positive outcomes is created.

Another important evaluator outcome is continuous change. This change can be made evident by a shift in school culture and staff attitudes (Reiss, 2007). Lasting change or sustainability occurs when the momentum and enthusiasm are persistant throughtout the more difficult times of implementation, even after funding has diminished or dissipated. The outcome of problem solving is among evaluator goals suggested by Hasbrouch and Denton (2005). The immediate situations are disentangled, and future ones can be prevented. To assist an evaluator in creating results with the desired outcomes, an evaluator needs to engage specific skills.

If the purpose of reading evaluators is to help teachers to educate children to become better readers, then a number of skills exist in which they need to engage to do this successfully:

1. A reading evaluator should contribute to the profession by sharing his/her knowledge of research-based instructional practices. This sharing can be achieved by teaching educators during grade level groups, conducting workshops, and modeling/demonstrating within the classroom (Hasbrouk & Denton, 2005; Riddle-Buly, Coskie, Robinson, & Egawa, 2006; Shanklin, 2006).
2. A reading evaluator should be able to recommend reading assessments, train others how to use them, and monitor their use for fidelity (Shanklin, 2006). The reading evaluator must ensure the data collected from the assessments are reviewed in a timely manner and plans are created from that data for student achievement. Even more importantly, the plans have to be carried out.

3. A reading evaluator should guide schools with "organizing and managing their reading programs" (Hasbrouck & Denton, 2005, p. 2). By doing so, the school is more apt to understand and make use of the suggested reading programs.
4. A reading evaluator should be available to reinforce and give encouragement to teachers (Shanklin, 2006). By giving positive feedback, teachers are more likely to continue to implement a new strategy and gain confidence about his/her skills.
5. A reading evaluator coach must be a good problem solver (Hasbrouk & Denton, 2005). This skill can be carried out by analyzing facts and numbers and producing proposals for future scholar progress.
6. A reading evaluator coach should help set aims and goals with teachers (Riddle-Buly et al., 2004). That way, educators have the concentrated aim and can stay on course to glimpse the task through to end.
7. A reading evaluator should should spend much of his/her time and effort up front with groups and/or educators. Eventually, the reading evalulation gradually decreases engagement to encourage educators to extend implementation of new abilities on their own (Shanklin, 2006).

The first and second important behavior evaluator skills include experience with school team implementation and problem solving (Sugai & Horner, 2006; Sugai, Horner, & Todd, 2003). Having school team experience, the evaluator understands group dynamics and can help teams to move past problems and engage in possibilities. The third skill a behavior evaluator must engage in is making sure the building team meets regularly. Teams accomplish more when meeting on a consistent basis, which promotes a sense of ownership and commitment. Evaluators should attend all Positive Behavior Support (PBS) building team meetings and help to create the

agenda for those meetings (Sugai, Todd, & Newcomer, 2008). The fourth skill needed is the ability to set data-based goals and adhere to them in a timely manner. This is known as being the "positive nag." Positive nags remind team members of assignment due dates beforehand which promotes success vs. failure (Simonsen, Sugai, & Negron, 2008). A fifth effective behavior evaluator skill is competency with data collection tools (Sugai, Todd, & Newcomer, 2008). The evaluator needs to be able to recommend and use tools, interpret data, and analyze the statistics collected from the tools. A sixth skill includes guiding and assisting schools with implementation but not accomplishing the team's duties for them (Killion, 2008). By taking over team duties, an evaluator does not allow the staff to learn and become independent. Sustainability would be minimal, as there would be no investment and buy-in. Seventh, an effective behavior evaluator keeps a log and frequently updates the team's performance (Sugai et al., 2003).

Purposes of the Staff Evaluation

Staff evaluations are a necessary endeavor, albeit not always the easiest subject to approach. The Family Business Experts website (Family Business Institute, 2012) stated that performance appraisals can lead to a relationship strain between an employee and an employer as well as among coworkers. In addition, history has shown teacher evaluations have not been productive activities to improve job performance or boost confidence levels in employees (Peterson, 2000). Despite evaluation difficulties, it is still the best way schools have to document job efforts for duties such as evaluators. According to Hasbrouk and Denton (2005), it is important to monitor evaluators because "evaluators will also have to be supervised and evaluated. If a principal has never worked with a reading evaluator before, how will he or she be able to make these important decisions?" (p. 23).

Margulus and Melin (2005) stated the three main purposes for staff evaluations are as follows:

1. Evaluations are used to give feedback on how effective a person is at his/her job.
2. Evaluations provide a way of communicating at a personal level when talking about job objectives.
3. Evaluations allow administrators to evaluate a person and decide if he/she is right for job assignments and promotions.

Other purposes for conducting a staff evaluation include protecting children and shaping the professional practice (Peterson, 2000). Staff appraisal measures can provide fidelity and aid in promoting the sustainability of new skills (Fixsen, Blase, Horner, & Sugai, 2008). They can be used to assess required basic performance skills and documenting poor, good, or great job performance. Giving opportunities for staff growth and improvement also are good reasons for evaluating staff (Stronge & Tucker, 2003). In addition, evaluations help administrators look for future leaders.

Currently, the nation is in an educational era of "program accountability" where NCLB has mandated schools to assess and use data to enhance the education for all students. Underperforming schools who do not improve are subject to losing federal funding (Arends, 2006). Although accountability for funding purposes is essential, it is even more important to ensure quality for student success. Stronge and Tucker (2005) purported improved teacher performance is equal to school improvement. Moreover, staff performance appraisals "ensure that students are well served and that a school continues to function efficiently" (Fields, Reck, & Egley, 2006, p. 12).

Components of Staff Evaluations

Basically, there are two types of staff evaluations: formative and summative. The difference between the two is the former is used to give feedback for the employee to improve skills and the latter is used for performance accountability (Knapper & Cranton, 2001). Ideally, an evaluation should use a combination of formative and summative appraisals. Danielson and McGreal (2000) identified three elements of a teacher evaluation to consider. To begin, evaluators must understand levels of performance and know the difference between exemplary practice and what is "good enough." They need to have an instrument that can differentiate between beginning teachers and veteran teachers. One suggestion is to create a tool that includes levels for unsatisfactory, basic, proficient, and distinguished performances. Next, the evaluator must know how the assessment will be conducted. For example, will there be an observation? How will evidence be collected? Will it be through a required portfolio? Finally, an evaluation should be conducted so that "No matter who conducts the evaluation, the results must be the same" (Danielson & McGreal, 2000, p. 22).

Once the type of assessment has been decided upon and the performance levels have been established, the body of the evaluation has to be constructed. In their book, *Writing Meaningful Evaluations for NonInstructional Staff--Right Now!,* Barker and Searchwell (2004) suggested dividing staff performance appraisal areas into five components: specific tasks, level of expertise, preparation and organization, related responsibilities, and interpersonal domain. Within each of the above five areas, sets of subskills should be developed and included. Each subskill has to be observable and measurable to make an evaluation objective.

Implementing a Multitiered Model

There are several key features of implementing multitiered models. These features include establishing commitment, establishing a team, conducting an audit, establishing information systems, developing an action plan, implementing the plan, and using the data to revise the action plan. Implementation of the multitiered model provides for three layers of support: 100% of students receive Universal Supports. This involves core instruction that is both preventative and proactive. About 15% of students receive Secondary Supports. This is supplemental support that reduces risk. Roughly 5% of students receive Tertiary Supports. This instruction is functionally based and highly specific (MiBLSi, 2010).

MiBLSi Evaluation Tools and Timelines

Implementation fidelity was measured using the Planning and Evaluation Tool for Effective Schoolwide Reading Programs-Revised (PET-R; Kame'enui & Simmons, 2003), Positive Behavioral Interventions and Supports Self Assessment Survey (PBIS-SAS) and the Positive Behavioral Interventions and Supports Team Implementation Checklist (PBIS-TIC, Sugai et al., 2003). Student outcomes were measured using school-level aggregate data from the Dynamic Indicators of Basic Early Literacy Skills--6th Edition (DIBELS; Good & Kaminski, 2002) and average major discipline referrals per 100 students per day, as measured by the Schoolwide Information System (SWIS) (May et al., 2000).

Evaluation

Schoolwide,

> Effective reading support involves a three-tiered approach to prevention and intervention for reading problems in schools. The approach involves team-based training in strategies to prevent reading problems, support children with the most intense reading problems, and integrate effective academic and instructional systems. Critical features of this approach include strong research-based initial instruction that is comprehensive and addresses the needs of most students, a valid assessment system that involves screening

> and progress monitoring, and high quality, intensive interventions for struggling readers. (MiBLSi, n.d.)

Positive Behaviorial Interventions and Supports (PBIS) is

> a proactive, team-based process for creating and sustaining safe and effective schools. PBIS is a systems framework to increase the capacity of schools to educate all children utilizing research-based schoolwide and classroom interventions. An emphasis is placed on preventing the occurrence of problem behavior as well as data-based problem solving for addressing existing behavior concerns. In order to effectively implement a problem-solving model, information must be collected and used to continuously evaluate and improve the systems of supports. (Michigan Department of Education, n.d.)

Harms (2010) examined outcomes of a statewide, integrated RTI project and the relation between implementation fidelity and student outcomes in the context of a statewide integrated three-tier model. A three-tier model of integrated behavior and learning supports linking systems-wide implementation to student outcomes. This study explored elementary schools' implementation of an integrated three-tier model of reading and behavior supports as they participated with a statewide RTI project. The purpose of the study was to examine the process of implementing an integrated three-tier model and to explore the relation between implementation fidelity and student outcomes. This study evaluated the 2003-2009 outcomes of elementary schools participating with MiBLSi (Cohorts 1-5), including 21 schools in 2004, 31 schools in 2005, 50 schools in 2006, 165 schools in 2007, 95 schools in 2008, and 123 schools in 2009 in collaboration with 45 ISDs. Connections will be made to the status of this type of research nationally. This particular study began about 2 years ago. Research questions were the following: To what extent do schools implement three-tier reading and behavior systems with fidelity across time, and what is the relation between implementation fidelity and student outcomes?

A combination of descriptive analyses and generalized estimating equations were used to evaluate implementation fidelity over time and the relation between implementation fidelity and

student outcomes. Major results included (a) average implementation fidelity scores improved over time, although individual schools started with different scores and made various amounts of growth over time; (b) approximately half of the elementary schools included in the study attained criterion levels of implementation during their participation with the RTI project; (c) schools made the most amount of implementation growth between years 1 and 2; (d) overall implementation improvements and most year-to-year improvements were statistically significant; (e) the reading implementation checklist was a better predictor of student reading outcomes than the behavior implementation checklists as predictors of student behavior outcomes; and (f) the combination of reading and behavior implementation checklists added to the prediction of student behavior outcomes beyond the behavior measures alone (Harms, 2010). Table 1 presents the schools participating in the MiBLSi for the 5 years beginning with 2003.

Table 1

Elementary Schools Participating with MiBLSi Cohorts 1-5 2003-2009 (Harms, 2010)

Cohort	Total Schools Participating in the Project	Elementary Schools in this Study	Percent of Schools Included
1 – January 2004	22	13	59
2 – January 2005	31	25	81
3 – January 2006	50	44	88
4 – January 2007	165	85	52
5 – August 2008	96	71	75
Total	363	238	66

$r = .138$, $p = .01$
Note: See Appendix A for performance indicators.

Reviewing the above data, the purpose of this study is to inquire to what extent schools implement three-tier reading and behavior systems with fidelity across time.

Measures

Measures of implementation fidelity for reading and behavior are Planning and Evaluation Tool (PET) for Effective Schoolwide Reading Programs, Effective Behavior Support Team Implementation Checklist (EBS-TIC), and Effective Behavior Support Self-Assessment Survey (EBS-SAS). Units of analysis are whole-school building aggregated student data and Team based self- assessment of implementation fidelity. Several terms have been used to label the process of providing behavior suports to students. These include Positive Behavioral Interventions and Supports (PBIS) and Effective Behavior Support (EBS).

Conclusions

Wingate (2007) proposed the use of the blend of formative and summative metaevaluations. Also, she suggested the use of the ranking rubric to make metaevaluations less subjective. According to Brinkerhof, Brethower, Hluchyj, and Nowakowski (1983), metaevaluation was added as a standard by the joint committee in 1994. No longer is metaevaluation merely a nicety. It is now an expectation. Nearly everyone does informal metaevaluation, but formal evaluation is something else entirely. Not only should they (metaevaluators) be competent enough to do the original evaluation, but they also have to be able to tell if it was a good or bad one and be able to convince others that they know the difference.

Spouse (2001) mentioned one of central constituents of advising is to perform and educate simultaneously. This advising required the MiBLSi evaluators being involved as an integral part of the stakeholder group by taking part in trainings. Reiss (2007) proposed that the evaluator act as the "possibility thinker" because a productive evaluator can assist to proceed from "I can't" or "I won't" bivouac into an "I can" camp. According to Harms (2010), for reading we see a positive relation between the PET and percent of students at benchmark.

However, additional work is needed to determine why a strong relation was not found among the TIC, SAS, and discipline referral data.

CHAPTER III – METHODOLOGY

The purpose of this chapter is to provide a methodological procedure for a metaevaluation of the evaluation, "So, How Are We Doing? A Michigan Integrated Behavior Learning Support Initiative (MiBLSi) Evaluation Study." The metaevaluation will apply the four attributes of an evaluation--utility, feasibility, propriety, and accuracy--to determine the strength of the evaluation. The metaevaluation required evaluators to score 30 subsets measuring a standard established by JCSEE on metaevaluation to address the following research questions:

1. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the utility standard developed by the JCSEE?
2. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the feasibility standard developed by the JCSEE?
3. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the propriety standard developed by the JCSEE?
4. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the accuracy standard developed by the JCSEE?

Description of the MiBLSi Evaluation Study

MiBLSi is a Mandated Activities Project (MAP), funded under the Individuals with Disabilities Education Act (IDEA) through the Michigan Department of Education, Office of Special Education. A program evaluation was completed on the MiBLSi Participatory Elementary Schools program for presentation at the MiBLSi State Conference in March 2010. Stufflebeam's PEMC was used to determine the extent to which the original evaluation met the standards established for program evaluations.

The evaluation "So How Are We Doing? A MiBLSi Evaluation Study" was developed by Anna Harms, Project Specialist, and three codirectors (Steve Goodman, Margie McGlinchey, and Kathryn Schalimo) to describe changes in student behaviors and reading outcomes over a 6-year period from 2003 through 2009. The results are presented in graphs and charts to provide program results to educators. However, the evaluation was presented as a PowerPoint presentation that is publicly available on the Internet. A formalized evaluation of the program was not available.

Program Evaluations Metaevaluation Checklist

The PEMC (Stufflebeam, 1999) provides a checklist for determining if program evaluations meet the standards developed by the JCSEE. The PEMC is publicly available and can be adapted to meet the needs of the evaluation. For the purposes of the present study, the PEMC will be used as written. The checklist includes measures four categories of evaluations: (a) utility, (b) feasibility, (c) propriety, and (d) accuracy. Each category is further divided into specific subsets. Table 2 presents the categories and associated subsets.

Table 2

PEMC Categories and Subsets

PEMC Categories	Subsets	Checkpoints
Utility: The general consensus that program evaluations respond to the needs of the stakeholders	Stakeholder Identification	6
	Evaluator Credibility	6
	Information Scope and Selection	6
	Values Identification	6
	Report Clarity	6
	Report Timeliness and Dissemination	6
	Evaluation Impact	6
Feasibility: Evaluations are cost effective and possible in politically-charged settings	Practical Procedures	6
	Political Viability	6
	Cost Effectiveness	6
Propriety: Evaluations consider JCSEE standards regarding ethical issues, constituional concerns, human rights, and freedom of information	Service Orientation	6
	Formal Agreements, Reach Advanced Written Agreements	6
	Rights of Human Subjects	6
	Human Interactions	6
	Complete and Fair Assessment	6
	Disclosure of Findings	6
	Conflict of Interest	6
	Fiscal Responsibility	6
Accuracy: Evaluations meet the standards for technical merit of the information included in the evaluations	Program Documentation	6
	Context Analysis	6
	Described Purposes and Procedures	6
	Defensible Information Sources	6
	Valid Information	6
	Reliable Information	6
	Systematic Information	6
	Analysis of Quantitative Information	6
	Analysis of Qualitative Information	6
	Justified Conclusions	6
	Impartial Reporting	6
	Metaevaluation	6

Note: Burrows, n.d.

The evaluator reads the evaluation report and places a check mark in each box where the item is included in the evaluation. The Program Metaevaluation Checklist is an instrument in which the required elements (utilities, feasibility, propriety, accuracy) of a performance are listed and a score is assigned based on whether the element is present or not. This is a useful device for assessing simple performances or achievement in which the individual elements being assessed typically involve dichotomous types of judgments. For example, “Engage leadership figures to

identify other stakeholders," a *yes* response would earn 1 point and a *no* response would earn 0 points. Notice that this checklist element does not address the concept of quality of the work and does not easily inform the rater what to do with partial performances.

Next, the number of items is counted for each subset. The number of ratings is then totaled (minimum = 0, maximum = 6). This number is then weighted, with the number of subsets with *Excellent* ratings (6) multiplied by 4, *Very Good* ratings (5) multiplied by 3, *Good* (3) multiplied by 2, and *Fair* (2-3) multiplied by 1. The weighted scores are then summed to obtain a total score. According to Stufflebeam (2001a), the overall scores for each category can range from *Poor* to *Excellent*. These scores differ for each category and depend on the number of subsets within the categories. Table 3 presents the breakdown of scores for each category.

Table 3

Category Scores

Category	Subsets		Scoring
Utility	7	Excellent	26 to 28 (93 to 100%)
		Very Good	19 to 25 (68 to 92%)
		Good	14 to 18 (50 to 67%)
		Fair	7 to 13 (25 to 49%)
		Poor	0 to 6 to 24%)
Feasibility	3	Excellent	11 to 12 (93 to 100%)
		Very Good	8 to 10 (68 to 92%)
		Good	6 to 7 (50 to 67%)
		Fair	3 to 5 (25 to 49%)
		Poor	0 to 2 (0 to 24%)
Propriety	8	Excellent	30 to 32 (93 to 100%)
		Very Good	22 to 29 (68 to 92%)
		Good	16 to 21 (50 to 67%)
		Fair	8 to 15 (25 to 49%)
		Poor	0 to 7 (0 to 24%)
Accuracy	12	Excellent	45 to 48 (93 to 100%)
		Very Good	33 to 44 (68 to 92%)
		Good	24 to 32 (50 to 67%)
		Fair	12 to 23 (25 to 49%)
		Poor	0 to 11 (0 to 24%)

Procedures

The inclusion of the items on the evaluation "So, How Are We Doing? A Michigan Integrated Behavior Learning Support Initiative (MiBLSi) Evaluation Study" was rated by the researcher using the Stufflebeam's PEMC. The checklist was used to rate each of the 30 standards to determine the extent to which each standard was included in the evaluation and the strength of each of the four attributes of the evaluation--utility, feasibility, propriety, and accuracy. The metaevaluation did not use other raters to verify the researcher's findings.

Data Analyses

Scores obtained from each of the 30 standards (utility = 7, feasibility = 3, propriety = 8, accuracy = 12) were entered into an SPSS database. Means and standard deviations were calculated for each domain. Using Stufflebeam's PEMC, the internal consistency reliability was measured by the Cronbach's alpha coefficient of the 30 standards. The Kruskal-Wallis one-way analysis of variance by ranks (Siegel & Castellan, 1988) was conducted to assess differences among the domains.

CHAPTER IV - RESULTS

This chapter reports the results of the scoring of the PEMC. The data were entered into a SPSS dataset and analyzed using the statistical procedures of the Kruskal-Wallis one-way analysis of variance by ranks that assessed the differences among the standards (utility = 7, feasibility = 3, propriety = 8, accuracy = 12) and the Cronbach's alpha coefficients of the four major areas of (a) utility, (b) feasibility, (c) propriety, and (d) accuracy that assessed the reliabilities of the standards. Also reported is the scoring of each checkpoint for the 30 standards of the four domains.

Table 4 presents the 30 standards of the four domains. Each standard consisted of six checkpoints and were coded a 1 or a 0 to determine the extent to which each standard was included in the evaluation. The six checkpoints were summed (values ranged from 0-6 with 0-1 being *Poor;* 2-3, *Fair;* 4, *Good;* 5, *Very Good*; 6, *Excellent*) and were weighted. *Excellent* ratings (6) were given a value of 4, *Very Good* (5), 3; *Good* (4), 2; *Fair* (2-3), 1; and *Poor* (0-1), 0. For the domain of utility of the seven standards, 1 was rated *Excellent*, 3 were rated *Very Good*, and 3 were rated *Fair*. For the domain of feasibility of the three standards, 1 was rated *Good* and 2 were rated *Fair*. For the domain of propriety of the eight standards, 4 were rated *Excellent*, 3 were rated *Fair,* and 1 was rated *Poor*. For the domain of accuracy of the 12 standards, 3 were rated *Excellent*, 1was rated *Very Good*, 3 were rated *Good*, 2 were rated *Fair*, and 3 were rated *Poor.*

Table 4

Program Evaluations Metaevaluation Checklist

Program Evaluations Metaevaluation Checklist N=30	
Utility	
Subsection	**Score**
U1 Stakeholder Identification	1
U2 Evaluator Credibility	3
U3 Information Scope and Selection	4
U4 Values Identification	3
U5 Report Clarity	3
U6 Report Timeliness and Dissemination	1
U7 Evaluation Impact	1
Feasibility	
Subsection	**Score**
F1 Practical Procedures	1
F2 Political Viability	2
F3 Cost Effectiveness	1
Propriety	
Subsection	**Score**
P1 Service Orientation	4
P2 Formal Agreements	1
P3 Rights of Human Subjects	4
P4 Human Interactions	1
P5 Complete and Fair Assessment	4
P6 Disclosure of Findings	4
P7 Conflict of Interest	1
P8 Fiscal Responsibility	0
Accuracy	
Subsection	**Score**
A1 Program Documentation	4
A2 Context Analysis	2
A3 Described Purposes and Procedures	0
A4 Defensible Information Sources	4
A5 Valid Information	2
A6 Reliable Information	1
A7 Systematic Information	1
A8 Analysis of Quantitative Information	3
A9 Analysis of Qualitative Information	4
A10 Justified Conclusions	2
A11 Impartial Reporting	0
A12 Metaevaluation	0

Descriptive statistics of the domains are presented in Table 5. The domain of utility is composed of seven standards, the range of values are 1-4, with a mean of 2.3 and a standard deviation of 1.25. Feasibility has three standards, ranging in value from 1-2 and a mean and standard deviation of 1.3 and 0.58, respectively. The domain of propriety is composed of eight standards, the range of values are 0-4, with a mean of 2.4, and a standard deviation of 1.77. Accuracy has 12 standards, ranging in value from 0-4 and a mean of 2.0 and a standard deviation 1.56 respectively.

Table 5

Program Evaluations Metaevaluation – Descriptive Statistics of Domains

Program Evaluations Metaevaluation Descriptive Statistics of Domains						
Domain	**N**	**Total Score**	**Min/Max**	**Mean**	**Std Dev**	**Median**
Utility	7	16	1-4	2.3	1.25	3.0
Feasibility	3	4	1-2	1.3	0.58	1.0
Propriety	8	19	0-4	2.4	1.77	2.5
Accuracy	12	23	0-4	1.9	1.56	2.0

Table 6 presents the total scores, strength, and the quality of the four domains. The seven standards of utility were summed, divided by 28, and multiplied by 100 to determine the strength of the evaluation's provisions for Utility. The domain of utility was assessed a total score of 16 with a strength of 57.1%, thereby indicating a quality of *Good.* The three standards of feasibility were summed (4), divided by 12, and then multiplied by 100 to determine the strength of the evaluation's provisions for feasibility. This domain was assessed a strength of 33.3%, thereby indicating a quality of *Fair.* The eight standards of propriety were summed (19), divided by 32, and then multiplied by 100 to determine the strength of the evaluation's provisions for propriety. This domain was assessed a strength of 59.4%, thereby indicating a quality of *Good.* The 12

standards of accuracy were summed (23), divided by 48, and then multiplied by 100 to determine the strength of the evaluation's provisions for accuracy. This domain was assessed a strength of 47.9%, thereby indicating a quality of *Fair.*

Table 6

Program Evaluations Metaevaluation – Domain Scores

Program Evaluations Metaevaluation Domain Scores			
Domain	**Total Score**	**Strength**	**Quality**
Utility	16	57.1%	Good
Feasibility	4	33.3%	Fair
Propriety	19	59.4%	Good
Accuracy	23	47.9%	Fair

Cronbach's Alpha Reliability of the Standards of the Domains

A multiple-item instrument with Likert-type scaling was developed to assess the reliability of the domains. Each checkpoint of the domain's standard was scored from 0 to 6 (0-1 being *Poor*; 2-3, *Fair;* 4, *Good*; 5, *Very Good;* 6, *Excellent*). For example, U1 Stakeholder Identification was scored a 6. The Cronbach's alpha reliability coefficient ranges between 0 and 1. The closer Cronbach's alpha coefficient is to 1.0, the greater the internal consistency of the items in the scale (Gliem & Gliem, 2003). The standards of the four domains due to the small number of replications (n=6) produced negative Cronbach's alpha reliability coefficients and hence are not reported.

Reliability of the PEMC was assessed using the 30 standards. It resulted in an alpha of .203 with a split-halves correlation of .272. To improve the alpha, item deletion via reanalysis of Cronbach's alpha was used only after a factor analysis was attempted. Traditionally, factor analysis is used to try and reduce the number of variables in a scale while preserving all the

subscales by maintaining at least two items per subscale (Brown, 2006); however, due to the lack of variance on at least one item and the extremely small sample size, the factor analysis approach could not be conducted. Therefore, the item deletion by Cronbach's alpha approach was taken. Table 7 presents the item total statistics which inform what the Cronbach's alpha would be if the item were deleted. Upon examination of these findings, it was determined to delete standards U4, F1, A7, and A11 for they indicated that Cronbach's Alpha would increase over .300.

Table 7

Item-Total Statistics of the 30 Standards of the PEMC

Standard	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
U1	100.1667	186.567	.031	.201
U2	98.8333	184.567	.114	.179
U3	97.1667	195.767	-.009	.205
U4	98.0000	228.400	-.551	.345
U5	97.8333	170.967	.318	.116
U6	100.0000	193.200	-.073	.239
U7	100.6667	167.067	.280	.111
F1	101.1667	216.167	-.404	.302
F2	98.5000	123.900	.905	-.200
F3	100.0000	265.600	-.847	.453
P1	96.6667	196.667	-.082	.208
P2	99.5000	128.700	.756	-.144
P3	96.5000	195.900	.000	.204
P4	99.5000	128.700	.756	-.144
P5	96.6667	202.267	-.563	.231
P6	97.1667	200.967	-.248	.228
P7	100.0000	198.400	-.130	.253
P8	102.5000	195.900	.000	.204

A1	96.5000	195.900	.000	.204
A2	99.0000	129.200	.951	-.166
A3	102.5000	195.900	.000	.204
A4	97.6667	157.867	.535	.041
A5	99.0000	208.800	-.256	.294
A6	100.5000	164.700	.272	.106
A7	99.8333	231.367	-.486	.372
A8	98.8333	185.767	.024	.204
A9	96.8333	192.167	.242	.189
A10	99.0000	199.200	-.144	.263
A11	101.0000	216.000	-.358	.311
A12	101.0000	186.000	.161	.174

The statistical findings with 30 standards and 26 standards are presented in Table 9. It can be observed that with the deletion of the four standards (in the column titled 26 Standards), the Cronbach's alpha reliability coefficient increased to .600. A Cronbach's alpha of .600 indicates a scale of questionable internal consistency (Gliem & Gliem, 2003).

To assess further the reliability of the 26 standards (and the 30 standards), a split-halves method was conducted (Carmines & Zeller, 1979). The split half correlations cited in Table 8 are based upon splitting the sample of the 26 into two parts of 13 standards each. The correlation between these two parts was .657. However, this correlation is the reliability for each half of the scale rather than the total scale. To correct for this, the Spearman-Brown formula, r = (2r)/1+ r, where r=the correlation between parts, was applied. The estimated reliability of the 26-standard PEMC was .793, indicating an acceptable to good internal consistency (Gliem & Gliem, 2003).

Table 8

Reliability Statistics of the 30/26 Standards of the PEMC

Statistical Tests	30 Standards	26 Standards
Cronbach's Alpha	.203	.600
Correlation between Parts [a]	.272	.657
Spearman-Brown Split-Half Coefficient [a]	.427	.793

[a] Correlation between two Parts, 15 or 13 standards in each part.

Kruskal-Wallis One-Way Analysis of Variance Nonparametric Test

The Kruskal-Wallis one-way analysis of variance by ranks (Siegel & Castellan, 1988) was conducted to assess differences among the domains. Four standards (U4 Values Identification, F1 Practical Procedures, A7 Systematic Information, and A11 Impartial Reporting) were deleted from the analysis based upon the results of the Cronbach's Alpha test. Therefore, the domains of utility had 6 standards, of feasibility had 2, while propriety had its original 8, and accuracy was reduced to 10. It was hypothesized that there would be no differences among these domains. This test is used for nonparametric data and for deciding whether independent samples are from different populations. Sum of scores were calculated and divided by the number of standards to provide a Wilcoxon score for each domain. See the upper part of Table 9. For utility, the Wilcoxon score was 13.17; for feasibility, 10.50; for propriety, 14.31; and for accuracy, 13.65. These mean scores were compared and the Kruskal-Wallis statistic was calculated by the X^2 distribution with $df = k - 1$ (Siegel & Castellan, 1988). Statistical analysis indicated there to be no differences among the domains, X^2 (3) = 0.441, $p>.05$. These findings are in the lower part of Table 10.

Table 9

Program Evaluations Metaevaluation – Wilcoxon Scores (Rank Sums) for Domain's Score Classified by Domain

Program Evaluations Metaevaluation Wilcoxon Scores (Rank Sums) for Domain's Score Classified by Domain					
Domain	N	Sum of Scores	Expected under HO	Std Dev under HO	Mean Score
Utility	6	79.00	81.00	15.90	13.17
Feasibility	2	21.00	27.00	10.06	10.50
Propriety	8	114.50	108.00	17.42	14.31
Accuracy	10	136.50	135.00	18.36	13.65
Kruskal-Wallis Test of Domain Scores					
Chi-Square			DF		Pr
0.441			3		0.932

Figure 7 is a box plot of the distribution of Wilcoxon scores by domains. Average Wilcoxon scores are indicated by the diamonds.

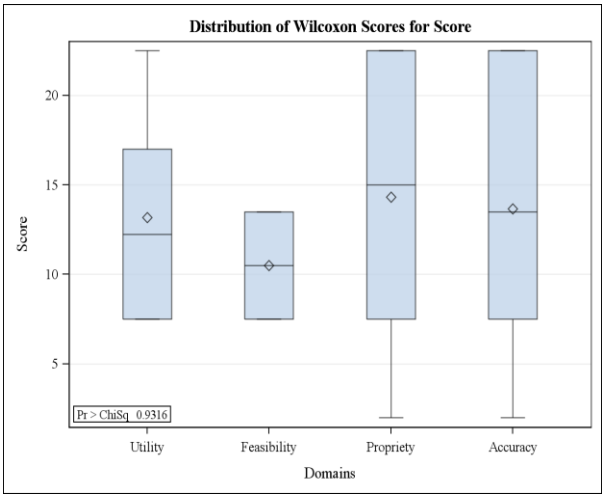

Figure 7. Box plot of domains by Wilcoxon scores.

Explanations for Scores of the 29 Standards of the Four Domains

There were four domains with a total of 29 standards, and each standard was scored on six checkpoints drawn from the substance of the standard. The six checkpoints were scored as either as 1, *present* or 0, *not present* for each of the 156 checkpoints.

Utility scoring results and explanations. Table 10 presents the explanations for scoring each checkpoint of the utility standard. The table includes the questions for the PEMC standard and a score of 0 for *not present* and 1 for *present* for each checkpoint.

For the checkpoints of U1 Stakeholder Identification, three of the six were scored as 1, *present* and three as 0, *not present.* The reasons that checkpoints 1, 3, and 6 were scored as present centers around the MiBLSi evaluation identifying the participating schools and the implied objectives (Harms, 2010, p.2). Checkpoints 2, 4, and 5 were scored as *not present* because the word "stakeholders" was not in the report and there was no evidence that evaluator consulted with stakeholders.

For the checkpoints of U2 Evaluator Credibility, five of the six were scored as 1, *present* and one as 0, *not present*. Checkpoints 1, 2, 3, 5, and 6 were scored as *present* because some evidence was postulated or inferred, and the evaluator was the MiBLSi state project specialist (Harms, 2010) and part of a 25-member technical team (slide 54). Checkpoint 4 was scored as *not present* because there was no evidence of these issues.

For the checkpoints of U3 Information Scope and Selection, all of the six were scored as 1, *present* and none as 0, *not present.* Checkpoints 1, 2, 3 4, 5, and 6 were scored as *present* due to the positive evidence presented, postulated, or inferred. Table 10 presents the evidence for checkpoints 2 and 3: MiBLSi has been collecting data from the beginning of the evaluation, which began 2 years ago (Harms, 2010, p. 5).

For the checkpoints of U5 Report Clarity, five of the six were scored as 1, *present* and one as 0, *not present.* Checkpoints 1, 2, 3 4, and 5 were scored as *present* due to the positive evidence presented, postulated, or inferred. Checkpoint 6 was scored as *not present* because there was no evidence of this issue.

For the checkpoints of U6 Timeliness and Dissemination, three of the six were scored as 1, *present* and three as 0, *not present.* Checkpoints 1, 4, and 6 were scored as *present* because of positive evidence postulated or presented, such as made special efforts to identify, reach, and inform all those intending to use the website and publishing on the web. Checkpoints 2, 3, and 5 were scored as *not present* because there was no evidence that evaluator addressed the issues of timeliness and dissemination.

For the checkpoints of U7 Evaluation Impact, 4 two of the six were scored as 1, *present* and four as 0, *not present.* Checkpoints 2 and 4 were scored as *present* because of positive evidence postulated or presented such as written reports with ongoing oral communication, such as a PowerPoint presentation. Checkpoints 1, 3, 4, and 5 were scored as *not present* because there was no evidence of the issue.

Table 10

Utility Scoring Results and Explanation

Utility	0=Not Present 1=Present	Explanation
U1 Stakeholder Identification		
1. Clearly identify the evaluation client.	1	Slides 1, 2, 6, and 13. Slide 6 defines MiBLSi as a state professional grant and page 13 lists the participating number of participating schools by cohort years.

2. Engage leadership figures to identify other stakeholders.	0	The word *stakeholders* is not present in the report.
3. Consult stakeholders to identify their information needs.	1	Slide 2 and implied in session objective outline.
4. Ask stakeholders to identify other stakeholders.	0	No evidence that evaluator consulted with stakeholders.
5. Arrange to involve stakeholders throughout the evaluation, consistent with the formal evaluation agreement.	0	No evidence that evaluator consulted with stakeholders.
6. Keep the evaluation open to serve newly identified stakeholders.	1	Although present on page 2 by implication of session objective, it is not directly addressed.
U2 Evaluator Credibility		
1. Engage competent evaluators.	1	Evaluator was one of 3 co-directors as shown on slide 54, and her credentials as an evaluator are assumed.
2. Engage evaluators whom the stakeholders trust.	1	The evaluator was the MiBLSi state project specialist as recorded on the cover page and part of a 25 technical team, slide 54.
3. Engage evaluators who can address stakeholders' concerns.	1	Postulated due to familiarity of work.
4. Engage evaluators who are appropriately responsive to issues of gender, socioeconomic status, race, and language and cultural differences.	0	Cannot make this assumption because there was no evidence of these issues.
5. Help stakeholders understand and assess the evaluation plan and process.	1	Postulated based on work area.
6. Attend appropriately to stakeholders' criticism and suggestions.	1	Evidence is inferred by discussion of implementation fidelity on page 4.
U3 Information Scope and Selection		
1. Assign priority to the most important questions.	1	Expressed on slide 4 as what it boils down to: Did we do what we said we would do, how and when we said we would do it?
2. Allow flexibility for adding questions during the evaluation.	1	Background: MiBLSi has been collecting data from the beginning and study began 2 years ago, slide 5.
3. Obtain sufficient information to address the stakeholders' most important evaluation questions.	1	Background: MiBLSi has been collecting data from the beginning and study began 2 years ago, slide 5.
4. Obtain sufficient information to assess the program's merit.	1	Postulated based on slide 5.
5. Obtain sufficient information to assess the program's worth.	1	Postulated based on slide 5.

6. Allocate the evaluation effort in accordance with the priorities assigned to the needed information.	1	Stakeholders include State of Michigan as evidenced by a state professional grant based on slide 6.
U5 Report Clarity		
1. Issue one or more reports as appropriate.	1	Multiple reports were provided including Pet school's attainment of criterion scores slide 20, PET mean scores over time slide 21, EBC=TIC school's attainment of criterion scores slide 22, and EBS-SAS school attainment of criterion scores slide 24.
2. Address the special needs of the audiences.	1	Audience identified as schools that implement 3 tier reading and behavior that address the need of student who receive 3 tier reading and behavior support slide 19.
3. Focus reports on contracted questions and convey the essential information in each report.	1	Page 13- Focused reports on contracted questions.
4. Write and/or present the findings simply and directly.	1	Slides 34-46 Wrote and present the findings simply in chart form.
5. Employ effective media for informing the different audiences.	1	The whole PowerPoint presentation employed effective media for informing the different audiences, using both presentations loaded with graphs.
6. Use examples to help audiences relate the findings to practical solutions.	0	No indication. Failure to use examples to help audiences relate the findings to practical solutions.
U6 Report Timeliness and Dissemination		
1. Make special efforts to identify, reach, and inform all intended users.	1	Website http://miblsi.cenmi.org Made special efforts to identify, reach, and inform all intended users using website and publishing on the web.
2. Make timely interim reports to intended users.	0	No evidence that evaluator attempted to notify timely interim reports to intended users.
3. Have timely exchanges with the pertinent audiences.	0	No evidence that evaluator had timely exchanges with the pertinent audiences.
4. Deliver the final report when it is needed.	1	Postulated that report was timely.
5. Issue press releases to the public media.	0	Cannot postulate that evaluator issued press releases to the public media.
6. Make findings publicly available via such media as the Internet.	1	Evaluator published findings via website http://miblsi.cenmi.org.
U7 Evaluation Impact		
1. Keep audiences informed throughout the Evaluation.	0	No evidence that evaluator kept audiences informed throughout the evaluation.

2. Forecast and serve potential uses of findings.	1	Pages 49, 52 stated potential uses of findings.
3. Provide interim reports.	0	No evidence that evaluator provided interim reports.
4. Supplement written reports with ongoing oral communication.	1	Supplemented written reports with ongoing oral communication such as PowerPoint presentation.
5. Conduct feedback sessions to go over and apply findings.	0	No evidence that evaluator had feedback sessions.
6. Make arrangements to provide following assistance in interpreting and applying the findings.	0	No indication evaluator made arrangements to provide following assistance in interpreting and applying the findings.

Note: Pages and slides refer to Harms (2010).

Feasibility scoring results and explanations. Table 11 presents the explanations for scoring each checkpoint of the feasibility standard. The table includes the questions for the PEMC standard and a score of 0 for *not present* and 1 for *present* for each checkpoint.

For the checkpoints of F2 Political Viability, four of the six were scored as *present* and two as *not present*. Checkpoints 1, 2, 3, and 4 were scored as *present* because evidence was postulated and hypothesized based on Harms (2010, pp. 52-53), indicating divergent views regarding the need to provide more support to schools.

For the checkpoints of F3 Cost Effectiveness, three of the six were scored as 1, *present* and three as 0, *not present*. Checkpoints 4, 5, and 6 were scored as *present* because evidence was postulated or hypothesized based on Harms (2010, pp. 52-53), which presented limited amounts of data that was actually submitted and available for analysis. Checkpoints 1, 2, and 3 were scored as *not present* was because there was no evidence of these issues.

Table 11

Feasibility Scoring Results and Explanation

Feasibility	0=Not Present 1=Present	Explanation
F2 Political Viability		
1. Anticipate positions of interest groups.	1	Postulated based on slide 50 evaluator anticipated different positions of different interest groups.
2. Anticipate actions designed to impede or destroy the evaluation.	1	Postulated based on slide 52 was vigilant on actions designed to impede or destroy the evaluation.
3. Foster cooperation.	1	Postulated based on slide 53, we need to provide more support to our schools in order to get the process data submitted.
4. Report divergent views.	1	Hypothesized based on slide 52, divergent views reported.
5. Make constructive use of diverse political forces to achieve the evaluation's purposes.	0	No evidence found if evaluator made constructive use of diverse political forces to achieve the evaluation's purposes.
6. Terminate a corrupted evaluation.	0	No evidence found if evaluation was efficient use of data.
F3 Cost Effectiveness		
1. Be efficient.	0	No evidence found if report was most efficient use of data
2. Make use of in-kind services.	0	No evidence found if evaluator made use of in-kind services.
3. Inform decisions.	0	No evidence found if evaluator used informed decisions.
4. Foster program improvement.	1	Hypothesized based on page 52 Limited amounts of data actually submitted and available for analysis.
5. Provide accountability information.	1	Hypothesized based on page 53 Provided accountability information.
6. Generate new insights.	1	Postulated based on page 53 Generated new insights into Cohort 2.

NOTE: Pages and slides refer to Harms (2010).

Propriety scoring results and explanations. Table 12 presents the explanations for scoring each checkpoint of the propriety standard. The table includes the questions for the PEMC standard and a score of 0 for *not present* and 1 for *present* for each checkpoint.

For the checkpoints of P1 Service Orientation, six of the six were scored as 1, *present* and none as 0, *not present*. Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *present* because some evidence was postulated based on Harms (2010, p. 49-53) in the evaluation.

For the checkpoints of P2 Formal Agreements, Reach Advance Written Agreement, three of the six were scored as 1, *present* and three as 0, *not present.* Checkpoints 1, 2, and 4 were scored as *present* because some evidence was postulated or presented. For example, Table 12 checkpoint 4 presents the following evidence: Release of reports data available on Harms (2010, pp. 34-49). Checkpoints 3, 5, and 6 were scored as *not present* because there was no evidence of this issue.

For the checkpoints of P3 Rights of Human Subjects, six of the six were scored as 1, *present* and none as 0, *not present.* Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *present* because of level of details postulated in evaluation. Checkpoint 4 was scored as *not present* was because there was no evidence of these issues.

For the checkpoints of P4 Human Interactions, three of the six were scored as 1, *present* and three as 0, *not present.* Checkpoints 1, 2, and 4 were scored as *present* because of the quality of the evaluation. Checkpoints 3, 5, and 6, were scored as *not present* because there was no evidence of these issues.

For the checkpoints of P5 Complete and Fair Assessment, six of the six were scored as, 1 *present* and none as 0, *not present.* Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *present* because of the evidence presented in Harms (2010, pp. 50-53). Checkpoint was scored as *not present* because of the strong evidence of these issues.

For the checkpoints of P6 Disclosure of findings, six of the six were scored as 1, *present* and none as 0, *not present*. Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *present* because of the

relevant points presented on Harms (2010, pp. 34-53). No checkpoint was scored as *not present* because the strong evidence of these issues.

For the checkpoints of P7 Conflict of Interest, three of the six were scored as, 1 *present* and three as 0, *not present.* Checkpoints 1, 3, and 6 were scored as *present* because the PowerPoint presentation (Harms, 2010) was evidence of evaluation records for independent review for checkpoint 3 in Table 12. Checkpoints 2, 4, and 5 were scored as *not present* because there was no evidence of these issues.

For the checkpoints of P8 Fiscal Responsibility, none of the six were scored as 1, *present* and six as 0, *not present.* Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *not present* because there was no evidence of these issues.

Table 12

Propriety Scoring Results and Explanation

Propriety	0=Not Present 1=Present	Explanation
P1 Service Orientation		
1. Assess program outcomes against targeted and nontargeted customers' assessed needs.	1	Postulated based on pages 34-47, assessed program outcomes against targeted customers.
2. Help assure that the full range of rightful program beneficiaries are served.	1	Postulated based on page 49, rightful program beneficiaries were served.
3. Promote excellent service.	1	Postulated based on page 50, promoted excellent service.
4. Identify program strengths to build on.	1	Postulated based on page 49, identify program strengths to build on.
5. Identify program weaknesses to correct.	1	Postulated based on pages 50, 51, identify program weaknesses to correct.
6. Expose persistently harmful practices.	1	Postulated based on page 53, expose persistently harmful practices.
P2 Formal Agreements		
1. Evaluation purpose and questions.	1	Evaluated purpose and questions present on slides 2, 8.
2. Audiences.	1	Postulated based on pages14.
3. Editing.	0	Unknown about editing of report and data.

4. Release of reports.	1	Release of reports data available on pages 34-49.
5. Evaluation procedures and schedule.	0	No evidence found if evaluation procedures and schedule was not discussed.
6. Evaluation resources.	0	No evidence found if evaluator discussed resources.
P3 Rights of Human Subjects		
1. Follow due process and uphold civil rights.	1	Followed due process and uphold civil rights. Postulated based on Department of Education regulations.
2. Understand participants' values.	1	Understood participants' values postulated based on Department of Education regulations.
3. Respect diversity.	1	Respected diversity of students postulated based on Department of Education regulations.
4. Follow protocol.	1	Followed protocol; postulated based on Department of Education regulations.
5. Honor confidentiality/anonymity agreements.	1	Honored confidentiality of students; postulated based on Department of Education regulations.
6. Minimize harmful consequences of the evaluation.	1	Minimized harmful consequences of the evaluation on subjects; postulated based on Department of Education regulations.
P4 Human Interactions		
1. Consistently related to all stakeholders in a professional manner.	1	Consistently related to all stakeholders in a professional manner; postulated based on Department of Education regulations.
2. Honor participants' privacy rights.	1	Honored privacy rights. Postulated based on Department of Education regulations.
3. Honor time commitments.	0	No evidence if evaluator honored time commitments.
4. Be sensitive to participants' diversity of values and cultural differences.	1	Sensitive to participants' diversity of values and cultural differences; postulated based on Department of Education regulations.
5. Be evenly respectful in addressing different stakeholders.	0	No evidence if respectful in addressing different stakeholders.
6. Do not ignore or help cover up any participant's incompetence, unethical behavior, fraud, waste, or abuse.	0	No evidence if evaluator attempted to cover up incompetence, unethical behavior, fraud, waste, or abuse.
P5 Complete and Fair Assessment		
1. Assess and report the program's strengths and weaknesses.	1	Pages 51, 52 assessed and reported the program's strengths and weaknesses.
2. Report on intended and unintended outcomes.	1	Page 53 reported on intended and unintended outcomes.
3. Show how the program's strengths could be used to overcome its weaknesses.	1	Page 53 discussed how the program's strengths could be used to overcome its weaknesses.

4. Appropriately address criticisms of the draft report.	1	Page 50 addressed criticisms of the draft report.
5. Acknowledge the final report's limitations.	1	Slide 52 acknowledged the final report's limitations.
6. Estimate and report the effects of the evaluation's limitations on the overall judgment of the program.	1	Slide 53 reported the effects of the evaluation's limitations on the overall judgment of the program.
P6 Disclosure of Findings		
1. Clearly define the right-to-know audience.	1	Page 4 defined the right-to-know audience.
2. Report relevant points of view of both supporters and critics of the program.	1	Pages 51, 52 reported relevant points of view of both supporters and critics of the program.
3. Report balanced, informed conclusions and recommendations.	1	Pages 50-54 informed conclusions and recommendations.
4. Report all findings in writing, except where circumstances clearly dictate otherwise.	1	Data available on pages 34-49 reported all findings.
5. In reporting, adhere strictly to a code of directness, openness, and completeness.	1	Page 52 acknowledged the final report's limitations.
6. Assure the reports reach their audiences.	1	Assured the reports reach their audiences via website http://miblsi.cenmi.org.
P7 Conflict of Interest		
1. Identify potential conflicts of interest early in the evaluation.	1	Pages 4, 10, 11 identified potential conflicts of interest
2. As appropriate and feasible, engage multiple evaluators.	0	No evidence if evaluator engaged other evaluators.
3. Maintain evaluation records for independent review.	1	PowerPoint presentation is evidence of evaluation records for independent review.
4. Contract with the funding authority rather than the funded program.	0	No evidence if evaluator contracted with the funding authority rather than the funded program.
5. Have the lead internal evaluator report directly to the chief executive officer.	0	No evidence if evaluator had the lead internal evaluator report directly to the chief executive officer.
6. Engage uniquely qualified persons to participate in the evaluation, even if they have a potential conflict of interest, but take steps to counteract the conflict.	1	Postulated based on credentials of evaluator.
P8 Fiscal Responsibility		
1. Specify and budget for expense items in advance.	0	No evidence if evaluator provided budget information.
2. Keep the budget sufficiently flexible to permit appropriate reallocations to strengthen the evaluation.	0	No evidence if evaluator provided budget information.
3. Maintain accurate records of sources of funding and expenditures and resulting evaluation services and products.	0	No evidence if evaluator provided budget information.
4. Maintain adequate personnel records concerning job allocations and time spent on the evaluation project.	0	No evidence if evaluator provided budget information.

5. Be frugal in expending evaluation resources.	0	No evidence if evaluator provided budget information.
6. Include an expenditure summary as part of the public evaluation report.	0	No evidence if evaluator provided budget information.

NOTE: Pages and slides refer to Harms (2010).

Accuracy scoring results and explanations. Table 13 presents the explanations for scoring each checkpoint of the accuracy standard. The table includes the questions for the PEMC standard and a score of 0 for *not present* and 1 for *present* for each checkpoint.

For the checkpoints of A1 Program Documentation, six of the six were scored as 1, *present,* and none were scored as 0, *not present*. Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *present* because some evidence was postulated or inferred through qualitative research over time and documented program progress.

For the checkpoints of A2 Context Analysis, four of the six were scored as 1, *present* and two as 0, *not present*. Checkpoints 1, 2, 3, and 4 were scored as *present* because some evidence was postulated or inferred by maintaining a log of unusual circumstances and contextual features and influences. Checkpoints 5 and 6 were scored as *not present* because there was no evidence of competitors and people's perceptions of program.

For the checkpoints of A3 Described Purposes and Procedures, none of the six were scored as 1, *present* and all six were scored as 0, *not present*. Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *not present* because no evidence was postulated or inferred on these issues.

For the checkpoints of A4 Defensible Information Sources, five of the six were scored as 1, *present* and one as 0, *not present*. Checkpoints 1, 2, 3, 4, and 6 were scored as *present* because some evidence was postulated or inferred about data collection sources and methods. Checkpoint 5 indicates that no evidence that data collection instruments were included in evaluation.

For the checkpoints of A5 Valid Information, three of the six were scored as 1, *present* and three as 0, *not present.* Checkpoints 1, 4, and 6 were scored as *present* because some evidence was postulated or inferred that the evaluator established key questions, inferences, and meaningful categories. Checkpoints 2, 3, and 5 were scored as *not present* because there was no evidence of information on procedures.

For the checkpoints of A6 Reliable Information, two the six scored as 1, *present* and four as 0, *not present*. Checkpoints 2 and 4 were scored as *present* because some evidence was postulated or inferred on issues on measuring devices and consistency of scoring. Checkpoints 1, 3, 5, 6 were scored as *not present* because there was no evidence of instrument devices or training.

For the checkpoints of A8 Analysis of Quantitative Information, five of the six were scored as 1, *present* and one as 0, *not present.* Checkpoints 1, 2, 3, 5, and 6 were scored as *present* because some evidence was postulated or inferred on the analysis of the quantitative data. Checkpoint 4 was scored as *not present* was because there was no evidence of examination variability as central tendencies.

For the checkpoints of A9 Analysis of Qualitative Information, six of the six were scored as 1, *present* and none as 0, *not present.* Checkpoints 1, 2, 3, 4, 5, and 6 were scored as *present* because some evidence was postulated or inferred on the analysis of the qualitative data collected.

For the checkpoints of A10 Justified Conclusions, four of the six were scored as 1, *present* and two as 0, *not present.* Checkpoints 1, 2, 3, and 5 were scored as *present* because some evidence was postulated or inferred on limited conclusions of information. Checkpoint 4 was scored as 0 because the evaluator did not discuss program side effects.

For the checkpoints of A12 Metaevaluation, one of the six was scored as 1, *present* and five as 0, *not present.* Checkpoint 1 was scored as *present* because evidence was postulated or inferred regarding proper budget because report is completed. Checkpoints 2, 3, 4, 5, and 6 were scored as *not present* because there was no evidence that evaluator designated standards and controls or inferred formative or summative information.

Table 13

Accuracy Scoring Results and Explanation

Accuracy	0=Not Present 1=Present	Explanation
A1 Program Documentation		
1. Collect descriptions of the intended program from various written sources and from the client and other key stakeholders.	1	There was evidence of program documentation on slide 9 that stated what we know about implementation.
2. Maintain records from various sources of how the program operated.	1	There was evidence of qualitative research (slide 9).
3. Analyze discrepancies between the various descriptions of how the program was intended to function.	1	Growth over time is statistically significant slides 34-49.
4. Analyze discrepancies between how the program was intended to operate and how it actually operated.	1	What we can celebrate (slides 48, 49)?
5. Record the extent to which the program's goals changed over time.	1	Inclusion of Cohort 2 results in cohort effects being more significant predictors than change over time page 40.
6. Produce a technical report that documents the programs operations and results.	1	Technical reports were produced from slides 36-47. Fiscal responsibility to include expenditure summary as part of the public evaluation report.
A2 Context Analysis		
1. Describe the context's technical, social, political, organizational, and economic features.	1	Multiple technical graphs were provided that showed reading and behavior measurements of cohorts slides 34-47.
2. Maintain a log of unusual circumstances.	1	Measures of implementation fidelity for reading and behavior although not a log slide18.
3. Report those contextual influences that appeared to significantly influence the program and that might be of interest to potential adopters.	1	MEAP percent of students at or above the state average.

4. Estimate the effects of context on program outcomes.	1	What do we need to work on and what can we celebrate? Slides 43 and 44.
5. Identify and describe any critical competitors to this program that functioned at the same time and in the program's environment.	0	No evidence of competitors; question posed but never answered.
6. Describe how people in the program's general area perceived the program's existence, importance, and quality.	0	No evidence of how people in the program general area perceived the program's existence.
A3 Described Purposes and Procedures		
1. Monitor and describe how the evaluation's purposes stay the same or change over time.	0	No evidence that evaluator monitored or described how the evaluation's purposes stay the same or change over time.
2. Update evaluation procedures to accommodate changes in the evaluation's purposes.	0	No evidence that evaluator updated evaluation procedures to accommodate changes in the evaluation's purposes.
3. Record the actual evaluation procedures, as implemented.	0	No evidence that evaluator recorded the actual evaluation procedures.
4. When interpreting findings, take in to account the extent to which the intended procedures were effectively executed.	0	No evidence that evaluator took in to account the extent to which the intended procedures were effectively executed.
5. Describe the evaluation's purposes and procedures in the summary and full-length evaluation reports.	0	No evidence that evaluator described the evaluation's procedures in the summary and full-length evaluation reports. Purpose described on slide 8.
6. Engage independent evaluators to monitor and evaluate the evaluation's purposes and procedures.	0	No evidence if evaluator engaged independent evaluators to monitor and evaluate the evaluation's purposes and procedures.
A4 Defensible Information Sources		
1. Once validated, use pertinent, previously collected information.	1	Evaluator used previously collected information Growth over time is statistically significant slides 34-48.
2. Employ a variety of data collection sources and methods.	1	Employed a variety of data collection sources and methods slides 34-48.
3. Document and report information sources.	1	Document and report information sources slides 34-48.
4. Document, justify, and report the means used to obtain information from each source.	1	Measuring implementation fidelity at your school slides 13, 14.
5. Include data collection instruments in a technical appendix to the evaluation report.	1	Did not include data collection devises slides 12.
6. Document and report any biasing features in the obtained information.	1	Reported any biasing features in the obtained information slide 44. What's go on in cohort 2.

A5 Valid Information		
1. Focus the evaluation on key questions.	1	Focused the evaluation on two key questions on slide 9: To what extent do schools implement 3 tier reading and behavior systems with fidelity across time? What is the relation between implementation fidelity and student outcomes?
2. Assess and report what type of information each employed procedure acquires.	0	No evidence that evaluator addressed procedures so question regarding them were never addressed.
3. Document how information from each procedure was scored, analyzed, and interpreted.	0	No evidence that evaluator addressed procedures so question regarding them were never addressed.
4. Report and justify inferences singly and in combination.	1	Justified inferences singly and in combination slides 34-47 PET and DiBELs scoring.
5. Assess and report the comprehensiveness of the information provided by the procedures as a set in relation to the information needed to answer the set of evaluation questions.	1	No evidence that evaluator addressed procedures so question regarding them were never addressed.
6. Establish meaningful categories of information by identifying regular and recurrent themes in information collected using qualitative assessment procedures.	0	Establish meaningful categories of information using graphs slides 34-49.
A6 Reliable Information		
1. Identify and justify the type(s) and extent of reliability claimed.	0	No evidence if evaluator justified the type and extent of reliability claimed.
2. Choose measuring devices that in the past have shown acceptable levels of reliability for their intended uses.	1	Evaluator used previously collected information; growth over time is statistically significant slides 20-47.
3. In reporting reliability of an instrument, assess and report the factors that influenced the reliability, including the characteristics of the examinees, the data collection conditions, and the evaluator's biases.	0	Unknown did not report the factors that influenced the reliability.
4. Check and report the consistency of scoring, categorization, and coding.	1	Reported the consistency of scoring, categorization, and coding slides 20-47.
5. Train and calibrate scorers and analysts to produce consistent results.	0	No evidence if evaluator trained and calibrated with scorers.
6. Pilot test new instruments in order to identify and control sources of error.	0	Evaluator never discussed pilot testing new instruments in order to identify and control sources of error.
A8 Analysis of Quantitative Information		
1. Begin by conducting preliminary exploratory analyses to assure the data's correctness and to gain a greater understanding of the data.	1	Inclusion/exclusion criteria demonstrated preliminary. Slide 13-15, preliminary analyses of quantitative information to gain an understanding of the data slide 14.
2. Report limitations of each analytic procedure, including failure to meet assumptions.	1	Limitation of analysis of quantitative information listed (slide 51).

3. Employ multiple analytic procedures to check on consistency and replicability of findings.	1	Multiple analytic procedures for analysis of quantitative information in graphs on slides 20-47.
4. Examine variability as well as central tendencies.	0	No evidence of analysis of variability and central tendencies of quantitative information.
5. Identify and examine outliers, and verify their correctness.	1	Outliners were postulated existed in each graph but not shown for the analysis of quantitative information pages.
6. Identify and analyze statistical interactions.	1	Multiple graphs depict correlation for DIBEL and student referrals used for analysis of quantitative data slides 46 and 47.
A9 Analysis of Qualitative Information		
1. Define the boundaries of information to be used.	1	Schools participating in the study defined as qualitative information slides 13 and 14.
2. Derive a set of categories that is sufficient to document, illuminate, and respond to the evaluation questions.	1	What does it mean to do RTI and MiBLSi? Part of documentation that respond to the evaluation slide 17.
3. Classify the obtained information into the validated analysis categories.	1	Schools attainment of criterion scores for PET, EBS_TIC and EBS_SAS were obtained to classify the analysis of information slides 20- 47.
4. Verify the accuracy of findings by obtaining confirmatory evidence from multiple sources, including stakeholders.	1	Schools attainment of criterion scores for PET, EBS_TIC and EBS_SAS were obtained to classify the analysis of information slides 20 -47.
5. Derive conclusions and recommendations, and demonstrate their meaningfulness.	1	Limitation of analysis of quantitative information listed slide 51.
6. Report limitations of the referenced information, analyses, and inferences.	1	Limitation of analysis of quantitative information listed slide 51.
A10 Justified Conclusions		
1. Limit conclusions to the applicable time periods, contexts, purposes, questions, and activities.	1	What we need to work on: Investigate the impact of meeting criterion on the behavior and reading implementation measures on student outcomes.
2. Report alternative plausible conclusions and explain why other rival conclusions were rejected.	1	Partially addressed on slide 44; failure to explain what happened in Cohort 2. Did not postulate a theory.
3. Cite the information that supports each conclusion.	1	Cited the information that supports each conclusion. Slides 49-53.
4. Identify and report the program's side effects.	0	Failure to identify and report the program's side effects.
5. Warn against making common misinterpretations.	1	Warned against making common misinterpretations slides 49-53.
6. Obtain and address the results of a prerelease review of the draft evaluation report.	0	Did not obtain and address the results of a prerelease review of the draft evaluation report.

A12 Metaevaluation		
1. Budget appropriately and sufficiently for conducting an internal metaevaluation and, as feasible, an external metaevaluation.	1	Postulated that evaluation was budgeted appropriately, as report is completed.
2. Designate or define the standards the standards the evaluators used to guide and assess their evaluation.	0	No evidence evaluation defined the standards the standards the evaluators used to guide and assess their evaluation. Postulated based on Department of Education regulations.
3. Record the full range of information needed to judge the evaluation against the employed standards.	0	Never postulated a full range of information needed to judge the evaluation against the employed standards.
4. As feasible and appropriate, contract for an independent metaevaluation.	0	Never postulated the contract for an independent metaevaluation.
5. Evaluate all important aspects of the evaluation, including the instrumentation, data collection, data handling, coding, analysis, synthesis, and reporting.	0	No evidence if evaluator evaluated all information provided by Department of Education.
6. Obtain and report both formative and summative metaevaluations to the right-to-know audiences.	0	Not able to postulated evaluator reported on both formative and summative metaevaluations.

NOTE: Pages and slides refer to Harms (2010).

CHAPTER V -- DISCUSSION

The findings of this study support the use of the PEMC to assess the extent the evaluation of the MiBLSi Participatory Elementary Schools from 2003-2009 required by the reauthorization of IDEA 2004 and meet the requirements for the program evaluations standards established by Stufflebeam (1999). First, the level of internal consistency, based on 26 of the 30 standards, was .79. This is the first evidence presented to date on the viability of Stufflebeam's checklist.

The deletion of four of the 30 standards, based on their psychometric properties, occurred based only on the sample examined in this study. Therefore, caution must be invoked prior to permanent deletion of those standards. However, should further replications indicate these four standards are heteroscedastic with regard to the overall checklist, then their permanent deletion should be considered.

Second, as regards the applied findings of the Stufflebeam checklist, no statistical significant differences were found among the four domains of utility, feasibility, propriety, and accuracy in the MIBLSI Participatory Elementary School meta-evaluation. The PEMC was used to answer the following questions:

1. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the utility standard developed by the JCSEE?
2. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the feasibility standard developed by the JCSEE?
3. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the propriety standard developed by the JCSEE?
4. To what extent does the evaluation of the MiBLSi Participatory Elementary Schools meet the accuracy standard developed by the JCSEE?

The domains of utility and propriety were assessed strengths of 57.1% and 59.4%, respectively. Therefore, it was determined that the evaluation's provisions for utility and propriety were *Good* (see Table 6). However, the domains of feasibility and accuracy were assessed strengths of 33.3% and 47.9%, respectively, indicating only a *Fair* quality in the evaluation's provisions for these two domains. The assessed strengths of the domains are fairly widespread.

Note that Wingate (2009) reported spreads in the intraclass correlation, which assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. She found the standards with the highest ICC values were from the accuracy domain while the standards from propriety and feasibility had the lowest ICC values. The standards of the utility domain presented a mixture–some low, others high. Wingate (2010) stated that there are some significant challenges to using the PEMC when the metaevaluation uses only evaluation reports. Although agreement was generally low across all the standards, the uncalibrated raters had the least agreement on standards in the feasibility and propriety domains, which are largely concerned with issues related to the manner in which an evaluation is carried out. With only reports in hand to judge the evaluation, raters had to infer quite a bit in order to make judgments about evaluation process (Wingate, 2010). The results reported here are only from the evaluation report.

To further complicate the metaevaluation, Stufflebeam's (1999) recommended that an evaluation be failed it if scored *Poor* on standards P1 Service Orientation, A5 Valid Information, A10 Justified Conclusions, or A11 Impartial Reporting. The standard of P1 Service Orientation scored *Excellent* because all six of the checkpoints were scored as *present* (see Table 12 section P1). The A5 Valid Information standard was scored *Fair* because three of the six checkpoints

were scored as 1, *present* and three as 0, *not present* (see Table 13 section A5). The A10 Justified Conclusions standard was scored *Good* because four of the six checkpoints were scored as 1, *present* and two as 0 (see Table 13 section A10). The A11 Impartial Reporting standard was scored *Poor* because only one of the six checkpoints was scored as 1, *present* and five as 0, *not present* (see Table 4). Therefore, the evaluation failed because the standard A11 Impartial Reporting was scored *Poor* as recommended by Stufflebeam, (1999). This standard may have been rated as such because the content concerning this standard was not included in the report. It does not necessarily hold that the standard was not met.

Conclusion

It can be concluded that the evaluation failed because the standard A11 Impartial Reporting was scored *Poor* as recommended by Stufflebeam, (1999). According to the scoring rubric, a single *Poor* result must result in determining the evalution has failed. This is probably too harsh, because the general standards are, in fact, not precise enough to measure a specific program or project. These need to be supported and concretized by specific, tailored standards, such as those used in the FOFS evaluation. Nevertheless, these general standards could be seen as useful tools for evaluators when preparing evaluations. The consideration of such standards could help to ameliorate evaluation studies and safeguard utilization of the results by means of a more user friendly (or in the words of an evaluator, "stakeholder oriented") format (Becker et al., 2004).

However, a caveat is called for here, because Stufflebeam (1999) asserted that the provider of the checklist had not modified or adapted the checklist to fit the specific needs of the user, and the user should execute his or her own discretion and judgment when using the checklist. The current study has processed to modify the PEMC through statistical analysis.

When the items contributing poorly to the scale's internal reliability were removed from analysis, the scale demonstrated a better internal consistency. The items of the scale seemed to be measuring more the same construct. Removing the four standards based on their poor psychometric characteristics improved the X^2 from p = .78 to 0.441 (Table 9), which at least represents a change in the desired direction.

It can be recommended that a specific and deliberate set of evaluation standards, or tailored standards, should be adapted and calibrated in accordance to the examined topic. However, it is helpful for evaluators and can furthermore greatly facilitate a worthwhile evaluation study if a set of established and accepted standards are consulted when preparing the evaluation. Such improvements would increase the likelihood that evaluation results will be utilized, encourage greater acceptance of the outcomes, and thus justify evaluation itself (Becker et al., 2004).

APPENDIX

PROGRAM EVALUATIONS METAEVALUATION CHECKLIST

PROGRAM EVALUATIONS METAEVALUATION CHECKLIST
(Based on The Program Evaluation Standards)
Daniel L. Stufflebeam, 1999

This checklist is for performing final, summative metaevaluations. It is organized according to the Joint Committee Program Evaluation Standards. For each of the 30 standards, the checklist includes 6 checkpoints drawn from the substance of the standard. It is suggested that each standard be scored on each checkpoint. Then judgments about the adequacy of the subject evaluation in meeting the standard can be made as follows: 0-1 Poor, 2-3 Fair, 4 Good, 5 Very Good, 6 Excellent. It is recommended that an evaluation be failed it if scores Poor on standards P1 Service Orientation, a5 Valid Information, a10 Justified Conclusions, or a11 Impartial Reporting. Users of this checklist are advised to consult the full test of The Joint Committee (1994) Program Evaluation Standards. Thousand Oaks, CA: Sage Publications.

TO MEET THE REQUIREMENTS FOR UTILITY, PROGRAM EVALUATIONS SHOULD:

U1 Stakeholder Identification

- ❽ Clearly identify the evaluation client
- ❽ Engage leadership figures to identify other stakeholders
- ❽ Consult stakeholders to identify their information needs
- ❽ Ask stakeholders to identify other stakeholders
- ❽ Arrange to involve stakeholders throughout the evaluation, consistent with the formal evaluation agreement
- ❽ Keep the evaluation open to serve newly identified stakeholders.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

U2 Evaluator Credibility

- ❽ Engage competent evaluators
- ❽ Engage evaluators whom the stakeholders trust
- ❽ Engage evaluators who can address stakeholders' concerns
- ❽ Engage evaluators who are appropriately responsive to issues of gender, socioeconomic status, race, and language and cultural differences
- ❽ Help stakeholders understand and assess the evaluation plan and process
- ❽ Attend appropriately to stakeholders criticisms and suggestions

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

U3 Information Scope and Selection

- ❽ Assign priority to the most important questions
- ❽ Allow flexibility for adding questions during the evaluation
- ❽ Obtain sufficient information to address the stakeholders' most important evaluation questions.
- ❽ Obtain sufficient information to assess the program's merit
- ❽ Obtain sufficient information to assess the program's worth
- ❽ Allocate the evaluation effort in accordance with the priorities assigned to the needed information.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

U4 Values Identification

- Consider all relevant sources of values for interpreting evaluation findings, including societal needs. Customer needs, pertinent laws, institutional mission, and program goals.
- Determine the appropriate party(s) to make the valuation interpretations.
- Provide a clear, defensible basis for value judgments.
- Distinguish appropriately among dimensions, weights, and cut scores on the involved values.
- Take into account the stakeholders' values.
- As appropriate, present alternative interpretations based on conflicting, but credible value bases.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

U5 Report Clarity

- Issue one or more reports as appropriate, such as an executive summary, main report, technical report, and oral presentation.
- As appropriate, address the special needs of the audiences, such as persons with limited English proficiency.
- Focus reports on contracted questions and convey the essential information in each report.
- Write and/or present the findings simply and directly.
- Employ effective media for informing the different audiences.
- Use examples to help audiences relate the findings to practical solutions.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

U6 Report Timeliness and Dissemination

- In cooperation with the client, make special efforts to identify, reach, and inform all intended users.
- Make timely interim reports to intended users.
- Have timely exchanges with the pertinent audiences (e.g., the program's policy board, the program's staff, and the program's customers).
- Deliver the final report when it is needed.
- As appropriate, issue press releases to the public media.
- If allowed by the evaluation contract and as appropriate, make findings publicly available via such media as the Internet.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

U7 Evaluation Impact

- As appropriate and feasible, keep audiences informed throughout the evaluation.
- Forecast and serve potential uses of findings.
- Provide interim reports.
- Supplement written reports with ongoing oral communication.
- To the extent appropriate, conduct feedback sessions to go over and apply findings.
- Make arrangements to provide following assistance in interpreting and applying the findings.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

Scoring the Evaluation of UTILITY

Add the following:

Number of excellent ratings (0-7) ____ x 4 =_____

Number of very good (0-7) ____ x 3 = _____

Number of Good (0-7) ____ x 2 = _____

Number of Fair (0-7) ____ x 1 = _____

Total Score _____

Strength of the evaluation's provisions for UTILITY

- 26 (93%) to 28: Excellent
- 19 (68%) to 25: Very Good
- 14 (50%) to 18: Good
- 7 (25%) to 13: Fair
- 0 (0%) to 6: Poor

_____ (Total Score) ÷ 28 = _____ x 100 = _____

TO MEET THE REQUIREMENTS FOR FEASIBILITY, PROGRAM EVALUATIONS <u>SHOULD</u>:

F1 Practical Procedures

- Minimize disruption and data burden.
- Appoint competent staff and train them as needed.
- Choose procedures in light of known resource and staff qualifications constraints.
- Make a realistic schedule.
- As feasible and appropriate, engage locals to help conduct the evaluation.
- As appropriate, make evaluation procedures a part of routine events.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

F2 Political Viability

- Anticipate different positions of different interest groups.
- Be vigilant and appropriately counteractive concerning pressures and actions designed to impede or destroy the evaluation.
- Foster cooperation.
- Report divergent views.
- As possible, make constructive use of diverse political forces to achieve the evaluation's purposes.
- Terminate an corrupted evaluation.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

F3 Cost Effectiveness

- Be efficient.
- Make use of in-kind services.
- Inform decisions.
- Foster program improvement.
- Provide accountability information.
- Generate new insights.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

Scoring the Evaluation of Feasibility
Add the following:
Number of excellent ratings (0-4) ____ x 4 =_____
Number of very good (0-4) ____ x 3 = _____
Number of Good (0-4) ____ x 2 = _____
Number of Fair (0-4) ____ x 1 = _____
Total Score _____

Strength of the evaluation's provisions for FEASIBILITY
- 11 (93%) to 12: Excellent
- 8 (68%) to 10: Very Good
- 6 (50%) to 7: Good
- 3 (25%) to 5: Fair
- 0 (0%) to 2: Poor

_____ (Total Score) ÷ 12 = _____ x 100 = _____

TO MEET THE REQUIREMENTS FOR PROPRIETY, PROGRAM EVALUATION *SHOULD*

P1 Service Orientation

- Assess program outcomes against targeted and nontargeted customers' assessed needs.
- Help assure that the full range of rightful program beneficiaries are served.
- Promote excellent service.
- Identify program strengths to build on.
- Identify program weaknesses to correct.
- Expose persistently harmful practices.

6 Excellent	5 Very Good	4 Good	2-3 Fair	0-1 Poor

P2 Formal Agreements, reach advance written agreements on:

- Evaluation purpose and questions
- Audiences.
- Editing.
- Release of reports.

❽ Evaluation procedures and schedule. ❽ Evaluation resources.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
P3 Rights of Human Subjects:				
❽ Follow due process and uphold civil rights. ❽ Understand participants' values. ❽ Respect diversity. ❽ Follow protocol. ❽ Honor confidentiality/anonymity agreements ❽ Minimize harmful consequences of the evaluation.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
P4 Human Interactions:				
❽ Consistently related to all stakeholders in a professional manner. ❽ Honor participants' privacy rights. ❽ Honor time commitments. ❽ Be sensitive to participants' diversity of values and cultural differences. ❽ Be evenly respectful in addressing different stakeholders. ❽ Do not ignore or help cover up any participant's incompetence, unethical behavior, fraud, waste, or abuse.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
P5 Complete and Fair Assessment:				
❽ Assess and report the program's strengths and weaknesses. ❽ Report on intended and unintended outcomes. ❽ As appropriate, show how the program's strengths could be used to overcome its weaknesses. ❽ Appropriately address criticisms of the draft report. ❽ Acknowledge the final report's limitations. ❽ Estimate and report the effects of the evaluation's limitations on the overall judgment of the program.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
P6 Disclosure of Findings:				
❽ Clearly define the right-to-know audience. ❽ Report relevant points of view of both supporters and critics of the program. ❽ Report balanced, informed conclusions and recommendations. ❽ Report all findings in writing, except where circumstances clearly dictate otherwise. ❽ In reporting, adhere strictly to a code of directness, openness, and completeness. ❽ Assure the reports reach their audiences.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

P7 Conflict of Interest:
❽ Identify potential conflicts of interest early in the evaluation. ❽ As appropriate and feasible, engage multiple evaluators. ❽ Maintain evaluation records for independent review. ❽ If feasible, contract with the funding authority rather than the funded program. ❽ If feasible, have the lead internal evaluator report directly to the chief executive officer. ❽ Engage uniquely qualified persons to participate in the evaluation, even if they have a potential conflict of

interest, but take steps to counteract the conflict.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

P8 Fiscal Responsibility:

❽ Specify and budget for expense items in advance.
❽ Keep the budget sufficiently flexible to permit appropriate reallocations to strengthen the evaluation.
❽ Maintain accurate records of sources of funding and expenditures and resulting evaluation services and products.
❽ Maintain adequate personnel records concerning job allocations and time spent on the evaluation project.
❽ Be frugal in expending evaluation resources.
❽ As appropriate, include an expenditure summary as part of the public evaluation report.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

Scoring the Evaluation of PROPRIETY

Add the following:

Number of excellent ratings (0-8) ____ x 4 = _____
Number of very good (0-8) ____ x 3 = _____
Number of Good (0-8) ____ x 2 = _____
Number of Fair (0-8) ____ x 1 = _____
Total Score _____

Strength of the evaluation's provisions for PROPRIETY

❽ 32 (93%) to 22: Excellent
❽ 22 (68%) to 29: Very Good
❽ 16 (50%) to 21: Good
❽ 8 (25%) to 15: Fair
❽ 0 (0%) to 7: Poor

_____ (Total Score) ◑ 32 = _____ x 100 = _____

TO MEET THE REQUIREMENTS FOR ACCURACY, PROGRAM EVALAUTIONS <u>SHOULD</u>:

A1 Program Documentation

❽ Collect descriptions of the intended program from various written sources and from the client and other key stakeholders.
❽ Maintain records from various sources of how the program operated.
❽ Analyze discrepancies between the various descriptions of how the program was intended to function.
❽ Analyze discrepancies between how the program was intended to operate and how it actually operated.
❽ Record the extent to which the program's goals changed over time.
❽ Produce a technical report that documents the programs operations and results.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

A2 Context Analysis

❽ Describe the context's technical, social, political, organizational, and economic features.
❽ Maintain a log of unusual circumstances.
❽ Report those contextual influences that appeared to significantly influence the program and that might be of interest to potential adopters.
❽ Estimate the effects of context on program outcomes.
❽ Identify and describe any critical competitors to this program that functioned at the same time and in the program's environment.
❽ Describe how people in the program's general area perceived the program's existence, importance, and quality.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

A3 Described Purposes and Procedures

❽ Monitor and describe how the evaluation's purposes stay the same or change over time.
❽ As appropriate, update evaluation procedures to accommodate changes in the evaluation's purposes.
❽ Record the actual evaluation procedures, as implemented.
❽ When interpreting findings, take in to account the extent to which the intended procedures were effectively executed.
❽ Describe the evaluation's purposes and procedures in the summary and full-length evaluation reports.

❽ As feasible, engage independent evaluators to monitor and evaluate the evaluation's purposes and procedures.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
A4 Defensible Information Sources				
❽ Once validated, use pertinent, previously collected information. ❽ As appropriate, employ a variety of data collection sources and methods. ❽ Document and report information sources. ❽ Document, justify, and report the means used to obtain information from each source. ❽ Include data collection instruments in a technical appendix to the evaluation report. ❽ Document and report any biasing features in the obtained information.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
A5 Valid Information				
❽ Focus the evaluation on key questions. ❽ Assess and report what type of information each employed procedure acquires. ❽ Document how information from each procedure was scored, analyzed, and interpreted. ❽ Report and justify inferences singly and in combination. ❽ Assess and report the comprehensiveness of the information provided by the procedures as a set in relation to the information needed to answer the set of evaluation questions. ❽ Establish meaningful categories of information by identifying regular and recurrent themes in information collected using qualitative assessment procedures.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
A6 Reliable Information				
❽ Identify and justify the type(s) and extent of reliability claimed. ❽ As feasible, choose measuring devices that in the past have shown acceptable levels of reliability for their intended uses. ❽ In reporting reliability of an instrument, assess and report the factors that influenced the reliability, including the characteristics of the examinees, the data collection conditions, and the evaluator's biases. ❽ Check and report the consistency of scoring, categorization, and coding. ❽ Train and calibrate scorers and analysts to produce consistent results. ❽ Pilot test new instruments in order to identify and control sources of error.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
A7 Systematic Information				
❽ Establish protocols and mechanisms for quality control of the evaluation information. ❽ Verify data entry. ❽ Proofread and verify data tables generated from computer output or other means. ❽ Systematize and control storage of the evaluation information. ❽ Strictly control access to the evaluation information according to established protocols. ❽ Have data providers verify the data they submitted.				
❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor
❽ Whenever possible, begin by conducting preliminary exploratory analyses to assure the data's correctness and to gain a greater understanding of the data. ❽ Report limitations of each analytic procedure, including failure to meet assumptions. ❽ Employ multiple analytic procedures to check on consistency and replicability of findings. ❽ Examine variability as well as central tendencies. ❽ Identify and examine outliers, and verify their correctness. ❽ Identify and analyze statistical interactions.				

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

A9 Analysis of Qualitative Information

- ❽ Define the boundaries of information to be used.
- ❽ Derive a set of categories that is sufficient to document, illuminate, and respond to the evaluation questions.
- ❽ Classify the obtained information into the validated analysis categories.
- ❽ Verify the accuracy of findings by obtaining confirmatory evidence from multiple sources, including stakeholders.
- ❽ Derive conclusions and recommendations, and demonstrate their meaningfulness.
- ❽ Report limitations of the referenced information, analyses, and inferences.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

A10 Justified Conclusions

- ❽ Limit conclusions to the applicable time periods, contexts, purposes, questions, and activities.
- ❽ Report alternative plausible conclusions and explain why other rival conclusions were rejected.
- ❽ Cite the information that supports each conclusion.
- ❽ Identify and report the program's side effects.
- ❽ Warn against making common misinterpretations.
- ❽ Whenever feasible and appropriate, obtain and address the results of a prerelease review of the draft evaluation report.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

A11 Impartial Reporting

- ❽ Engage the client to determine steps to ensure fair, impartial reports.
- ❽ Safeguard reports from deliberate or inadvertent distortions.
- ❽ As appropriate and feasible, report perspectives of all stakeholder groups and, especially, opposing views on the meaning of the findings.
- ❽ As appropriate and feasible, add a new, impartial evaluator late in the evaluation to help offset any bias the original evaluators may have developed due to their prior judgments and recommendations.
- ❽ Describe steps taken to control bias.
- ❽ Participate in public presentations of the findings to help guard against and correct distortions by other interested parties.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

A12 Metaevaluation

- ❽ Budget appropriately and sufficiently for conducting an internal metaevaluation and, as feasible, an external metaevaluation.
- ❽ Designate or define the standards the standards the evaluators used to guide and assess their evaluation.
- ❽ Record the full range of information needed to judge the evaluation against the employed standards.
- ❽ As feasible and appropriate, contract for an independent metaevaluation.
- ❽ Evaluate all important aspects of the evaluation, including the instrumentation, data collection, data handling, coding, analysis, synthesis, and reporting.
- ❽ Obtain and report both formative and summative metaevaluations to the right-to-know audiences.

❽ 6 Excellent	❽ 5 Very Good	❽ 4 Good	❽ 2-3 Fair	❽ 0-1 Poor

Scoring the Evaluation of ACCURACY		Strength of the evaluation's provisions for ACCURACY	
Add the following:		❽ 45 (93%) to 48:	Excellent
Number of excellent ratings (0-12)	____ x 4 = _____	❽ 33 (68%) to 44:	Very Good
Number of very good (0-12)	____ x 3 = _____	❽ 24 (50%) to 32:	Good
Number of Good (0-12)	____ x 2 = _____	❽ 12 (25%) to 23:	Fair
Number of Fair (0-12)	____ x 1 = _____	❽ 0 (0%) to 11:	Poor

Total Score _____	_____ (Total Score) ➗ 48 = _____ x 100 = _____

This checklist is being provided as a free service to the user. The provider of the checklist has not modified or adapted the checklist to fit the specific needs of the user and the user is executing his or her own discretion and judgment in using the checklist. The provider of the checklist makes no representations or warranties that this checklist is fit for the particular purpose contemplated by user and specifically disclaims any such warranties or representations.

REFERENCES

Alpena-Montmorency-Alcona Educational Service District. (2011). *Michigan's Integrated Behavior and Learning Support Initiative (MiBLSi)*. Retrieved from http://www.amaesd.k12.mi.us./MiBLsi.asp

American Evaluation Association (2004). Guiding principles for evaluators. Retrieved from http://www.eval.org/publications/guidingprinciples.asp

American National Standards Institute. (2009). Education organization is accredited by ANSI. *ANSI Reporter, 4*(23), 1, 2, 4.

Arends, R. I. (2006*)*. Performance assessment in perspective: History, opportunities and challenges. In S. Castle & B. Shaklee (Eds.), *Assessing Teacher Performance* (pp. 3-22). Lanham, MD: Rowman & Littlefield Education.

Ashworth, A.*,* Cebulla, A.*,* Greenberg, D.*, &* Walker, A. (2004). Meta evaluation: Discovering what works best in welfare provision. *Evaluation, 10,* 193-216.

Baizerman, M.*,* Compton, D.W.*, &* Stockdill, S. (2002). New directions for ECB. *In* D. W. Compton*,* M. Baizerman*, &* S. Stockdill (Eds.), *The Art, Craft and Science of Evaluation Capacity Building: New Directions for Evaluation,* (93,pp. 109-116). San Francisco, CA*:* Jossey-Bass.

Barker, C. L., & Searchwell, C. J. (2004). *Writing meaningful evaluations for non-instructional staff-right now!* Thousand Oaks, CA: Corwin Press.

Becker, C., Ekert, S., Sommer, J., & Zorn, A. (2004). Meta-evaluation of action plans–The case of the German Federal Organic Farming Scheme.

Brinkerhoff, R.O., Brethower, D. M., Hluchyj, T., & Nowakowski, J. R. (1983). *Program evaluation: A practitioner's guide for trainers and educators*. Boston, MA: Kluwer-Nijhoff.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Burrows, S. (n.d.). *Daniel Stufflebeam's contribution to programme evaluation.* Retrieved from http://www.aseesa-ed.co.za/bullh.htm

Burt, R., & Celotto, N. (1992). The network structure of management roles in a large matrix firm. *Evaluation and Program Planning, 15*, 303-326.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage.

Connor, D. J., & Ferri, B. A. (2005). Integration and inclusion: A troubling nexus: Race, disability, and special education. *The Journal of African American History, 90*(1-2), 107-127.

Cooksy, L. J. (1999). The meta-evalaund: The evaluation of project TEAMS. *American Journal of Evaluation, 20*(1), 123-136

Cooksy, L. J., & Caracelli, V. J. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evaluation, 26*(1), 31-42.

Cousins, J.B., & Shulha, L.M. (2006). A comparative analysis of evaluation and its cognate fields of inquiry: Current issues and trends. In I. F. Shaw, J. C. Greene, & M. M. Mark *(Eds.), The SAGE Handbook of Evaluation* (pp. 233-254). Thousand Oaks, CA: Sage.

Cummings, R. (2002). *Rethinking evaluation use*. Paper presented at the 2002 Australasian Evaluation Society International Conference, Wokongong, Australia.

Danielson, C., & McGreal, T. (2000). *Teacher metaevaluation to enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

DeValenzuela, J. B., Copeland, S. R., Huaqing, C., & Park, M. (2006). Examining educational equity: Revisiting the disproportionate representation of minority students in special education. *Exceptional Children, 72*(4), 425-441.

Education of all Handicapped Children Act. Public Law 94-142. (1975).

Family Business Institute. (2012). *Employee performance appraisals: Part of the process.* Family Business Experts. Retrieved from http://www.family-business-experts.com/employee-performance-appraisals.html

Fetterman, D.*, &* Wandersman, A. *(*2007*).* Empowerment evaluation: Yesterday, today and tomorrow. *American Journal of Evaluation, 28,* 179-198.

Fields, L., Reck, B., & Egley, R. (2006). *Managing marginal participatory elementary school employees: Applying standards-based performance measures*. Lanham, MD: Rowman and Littlefied Publishers.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston, MA: Pearson.

Fixsen, D. L., Blase, K. A., Horner, R. H., & Sugai, G. (2008). *Developing the capacity for scaling up the effective use of evidence-based programs in state departments of education.* Chapel Hill, NC: State Implementation of Scaling-up Evidence-based Practices (SISEP) Center.

Gliem, J. A., & Gliem, R. R. (2003). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales*. Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8-10, 2003.

Good, R. H., & Kaminski, R. A. (2002). *DIBELS oral reading fluency passages for first through third grades* (Technical Report No. 10). Eugene: University of Oregon. Retrieved from http://www.salkeiz.k12.or.us/system/files/ORF_Readability.pdf

Grasso, P. *(*1999*).* Metaevaluation of an evaluation of reader focused writing for the Veterans Benefits Administration*. American Journal of Evaluation, 20,* 355-370*.*

Hale, J. B., Kaufman, A., Naglien, J.A., & Kavala, K. A. (2006). Implementation of IDEA: Integrating response to intervention and cognitive assessment methods. *Psychology in the Schools, 43*(7), 753-770. doi: 10.1002/pits.20186

Hanssen, C. E., Lawrenz, F., Dunet, D. O. (2008). Concurrent meta-evaluation: A critique. *American Journal of Evaluation, 29,* 572-582. doi: 10.177/1098214008320462

Harms, A. (2010). *So, how are we doing?* A MiBLSi Evaluation Study presented at MiBLSi State Conference. Retrieved from http://miblsi.cenmi.org/LinkClick.aspx?fileticket=L-KOEpI0hbY%3D&tabid=1240

Harry, B., & Klingner, J. (2007). Discarding the deficit model. *ASCD.* Retrieved April 16, 2012 from http://www.ascd.org/publications/educational-leadership/feb07/vol64/num05/Discarding-the-Deficit-Model.aspx?&lang=en_us&output=json&session-id=adf7369124f8ce47da9952967f9ea799

Hasbrouck, J., & Denton, C. (2005). *The reading evaluator: A how-to manual for success*. Longmont, CO: Sopris West.

Heinzelman, D., LaPointe, S., & Vanderploeg, L. (2010). *A template for the SLD evaluation and eligibility determination process*. Retrieved from http://www.maase.org/Files/MAASE_PD_Handout-12-08_LaPointe-et_al.pdf

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*(3), 293-314. doi: 10.1016/S1098-2140(03)00056-0

Hersey, J., Williams-Piehota, P., Sparling, P., Alexander, J., Hill, M., Isenberg, K. B, . . . Dunet, D. *(2010).* Promising practices in promotion of healthy weight at small and medium-sized U.S. work sites. *Preventing Chronic Disease, 5*(4). Retrieved from http://www.cdc.gov/pcd/issues/2008/

Ives, W. (1992). Evaluating new multimedia technologies for self-paced instruction. *Evaluation and Program Planning, 15,* 287-296.

Jinkerson, D., Cummings, O., Neisendorf, B., & Schwandt, T. *(*1992*).* A case study of methodological issues in cross-cultural evaluation. *Evaluation and Program Planning, 15,* 273-285.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed). Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards*. Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (2008). *The personnel evaluation standards* (2nd ed). Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (2012). *The program evaluation standards*. Retrieved from http://www.jcsee.org/program-evaluation-standards

Kame'enui, E. J., & Simmons, D. C. (2003). *Planning and Evaluation Tool for effective school-wide reading programs – Revised (PET-R).* Eugene, OR: Institute for the Development of Educational Achievement.

Kavale, K. (n.d.). *Discrepancy models in the identification of learning disability.* Unpublished manuscript. Ames: University of Iowa. Retrieved from http://www.nrcld.org/resources/ldsummit/kavale.pdf

Killion, J. (2008). Are you coaching heavy or light? *Teachers Teaching Teachers, 3*(8), 1-4.

Knapper, C., & Cranton, P. (Eds.). (2001). Fresh approaches to the evaluation of teaching. *New Directions for Teaching and Learning, 88*. San Francisco, CA: Jossey-Bass.

Lapointe, S., & Heinzelman, D. (2006). *Response to intervention: Enhancing the learning of all children*. Arlington, VA: MAASE.

Margulus, L. S., & Melin, J.A. (2005). *Performance appraisals made easy.* Thousand Oaks, CA: Corwin Press.

Mark, M.*, &* Henry, G. (2004*).* The mechanisms and outcomes of evaluation influence. *Evaluation, 10*(1*),* 35-57.

May, S., Ard, W., Todd, A.W., Horner, R.H., Glasgow, A., Sugai, G., & Sprague, J.R. (2000). *School-wide information system: Educational and community supports*. Eugene: University of Oregon.

McIntosh, K., Horner, R. H., Chard, D. J., Dickey, C. R., & Braun, D. H. (2008). Reading skills and function of problem behavior in typical school settings. *Journal of Special Education, 42*(3), 131-147. doi: 10.1177/0022466907313253

Michigan Department of Education. (n.d.). MiBLSi Model. *Michigan's Integrated Behavior and Learning Support Initiative.* Retrieved April 16, 2012 from http://miblsi.cenmi.org/MiBLSiModel.aspx

Michigan's Integrated Behavior and Learning Support Initiative. (n.d.). *MiBLSi model.* Retrieved from http://miblsi.cenmi.org/MiBLSiModel.aspx

Mihalic, S., Irwin, K, Fagan, A., Ballard, D., & Elliott, D. (2004). Successful program implementation: Lessons from blueprints. *Juvenile Justice Bulletin*. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention.

Niemiec, R., Sikorski, M., Clark, G., & Walberg, H. (1992). Effects of management education: A quantitative synthesis. *Evaluation and Program Planning, 15,* 297-302.

Nilsson, N., & Hogben, D. (1983). Metaevaluation. In E. R. House (Ed.), Philosophy of Evaluation: *New Directions for Program Evaluation* (pp. 83-97). San Francisco, CA: Jossey-Bass.

Patton, M.Q. (1997). *Utilization focused evaluation: The new century text* (3rd Ed.). London, England: Sage.

Patton, M.Q. (2008). *Utilization-focused evaluation* (4th ed). Los Angeles, CA: Sage.

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.).Thousand Oaks, CA: Corwin Press.

Peterz, K. (1999, May). The overrepresentation of Black students in special education. *Motion Magazine*. Retrieved from http://www.inmotionmagazine.com/peterz1.html

Putnam, R. F., Handler, M., & O'Leary-Zonarich, C. (2003). *Improving academic achievement using school-wide behavioral support interventions*. San Francisco, CA: Annual Conference of the Association of Behavior Analysis.

Putnam, R. F., Handler, M., Rey, J.C., & O'Leary-Zonarich, C. (2002). *Classwide behavior support interventions: Using functional assessment practices to design effective interventions in general classroom settings*. Toronto, Canada: Annual Conference of the Association of Behavior Analysis.

Preciado, J. A., Horner, R. H., & Baker, S. K. (2009). Using a function-based approach to decrease problem behaviors and increase academic engagement for Latino English language learners. *The Journal of Special Education, 42*, 227-240.

Reiss, K. (2007). *Leadership coaching for educators: Bringing out best in participatory elementary school administrators.* Thousand Oaks, CA: Corwin Press.

Riddle-Buly, M., Coskie, T., Robinson, L., & Egawa, K. (2006). Literacy coaching: Coming out of the corner. *Voices from the Middle, 13*(4), 24-28.

Sanford, E. (2006). *The effects of function-based literacy instruction on problem behavior and reading growth.* Unpublished Dissertation. Eugene: University of Oregon.

Scriven, M. (1969). An introduction to metaevaluation. *Educational Products Report, 2,* 36-38.

Scriven, M. (1975). Evaluation bias and its control. *Occasional Paper Series*, *4.* Kalamazoo: The Evaluation Center, Western Michigan University. Retrieved from http://www.wmich.edu/evalctr/pubs/ops/ops04.pdf

Scriven, M. (1991) *Evaluation thesaurus* (4th ed.) Thousand Oaks, CA: Sage.

Scriven, M. (2007). *The logic and methodology of checklists.* Retrieved from http://www.wmich.edu/evalctr/archive_checklists/papers/logic&methodology_dec07.pdf

Scott-Little, C., Hamann, M. S., & Jurs, S. G. (2002). Evaluations of after-school programs: A metaevaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation, 23*(4), 387-419. doi: 10.1177/109821400202300403

Shanklin, N. L. (2006). What are characteristics of effective literacy coaching*? Literacy Coaching Clearinghouse*. Retrieved from http://www.usdb.org/curr/literacy/Administrator%20Materials/Characteristics%20of%20Literacy%20Coaching.pdf

Siegel, S., & Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd Ed.). New York, NY: McGraw-Hill.

Simonsen, B., Sugai, G., & Negron, M. (2008). Schoolwide positive behavior supports: Primary systems and practices. *Teaching Exceptional Children, 40*(6), 32-40.

Sinjindawong, S., Lawthong, N., & Kanjanawasee, S. (n.d.). The development and application of the meta-evaluation standards for Thai higher education institutions. *Research in Higher Education Journal.* Retrieved from http://www.aabri.com/manuscripts/10604.pdf

Spouse, J. (2001). Bridging theory and practice in supervisory relationship: A sociocultural perspective. *Journal of Advanced Nursing, 33*(4), 512-522.

Stemler, S. E. (2007). Interrater reliability. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 2, pp. 484-485). Thousand Oaks, CA: Sage.

Stewart, R., Benner, G., Martella R., & Marchand-Martella, N. (2007). Three-tier models of reading and behavior: A research review. *Journal of Positive Behavior Interventions 9*, 239. Retrieved from http://pbi.sagepub.com/content/9/4/239.full.pdf

Streiner D. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80,* 99-103.

Stronge, J. H., & Tucker, P. D. (2003). *Handbook on teacher evaluation: Assessing and improving performance.* Larchmont, NY: Eye on Education.

Stufflebeam, D. L. (1999). *Foundational models for 21st century program evaluation*. Retrieved from https://www.globalhivmeinfo.org/CapacityBuilding/Occasional%20Papers/16%20Foundational%20Models%20for%2021st%20Century%20Program%20Evaluation.pdf

Stufflebeam, D. L. (2001a). Evaluation checklists: Practical tools for guiding and judging evaluations. *American Journal of Evaluation, 22*, 71-79. doi: 10.1177/109821400102200107.

Stufflebeam, D. L. (2001b). The metaevaluation imperative. *American Journal of Evaluation, 22*(2), 183-209.

Stufflebeam, D., Madaus, D., & Kellaghan, T. (Eds.). (2000). *Evaluation models: Viewpoints in educational and human services evaluation* (2nd ed.). Norwell, MA: Kluwer Academic.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass.

Sugai, G., & Horner, R. H. (2002). The evolution of discipline practices: School-wide positive behavior supports. *Child and Family Behavior Therapy, 24,* 23-50.

Sugai, G., Horner, R., & Todd, A. (2003). *Effective behavior support self-assessment survey*. Eugene: University of Oregon.

Sugai, G., Lewis-Palmer, T., Todd, A., & Horner, R. (2001). *School-wide evaluation tool*. Eugene: University of Oregon.

Sugai, G., & Todd, A., & Newcomer, L. (2008, March). *MiBLSi coaching roles, and responsibilities, structures, skills & activities to support effective coaching.* PowerPoint presented at MiBLSi Coaches' Conference, Lansing, MI.

The National Center for Learning Disabilities, Inc. (2012) What is RTI? Retrieved April 19, 2012 from http://www.rtinetwork.org/learn/what

Wingate, L. (2002). *The evaluation checklist project: The inside scoop on content, process, policies, impact, and challenges.* Paper presented at the American Evaluation Association

Conference, Washington, D.C. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/papers/insidescoop.pdf

Wingate, L. (2009). *The program evaluation standards applied for metaevaluation purposes: Investigating interrater reliability and implications for use*. Kalamazoo: Western Michigan University.

Wingate, L. (2010, July 25). Lori Wingate on the use of the Program Evaluation Standards for Metaevaluation. *AEA365: A Tip-A-Day By and For Evaluators*. Retrieved March 8, 2012 from http://aea365.org/blog/?p=1214

Woodside, A., & Sakai, M. (2001). Metaevaluation of performance audits of government tourism marketing programs. *Journal of Travel Research, 39,* 369-379.

Wright, P. W. D., & Wright, P.D. (2007). *Wrightslaw: Special education law* (2nd ed.). In Wrightslaw. Hartfield, CT: Harbor House Law Press.

ABSTRACT

MiBLSi PROGRAM EVALUATION OF PARTICIPATORY ELEMENTARY SCHOOLS FROM 2003-2009

by

MARVIN GIBBS

August 2012

Advisor: Dr. Shlomo Sawilowsky

Major: Evaluation and Research

Degree: Doctor of Philosophy

This dissertation details the use of the Program Evaluations Metaevaluation Checklist (PEMC; Stufflebeam, 1999), which is for performing final, summative metaevaluations. It is organized according to the Joint Committee Program Evaluation Standards. For each of the 30 standards, the checklist includes six checkpoints drawn from the substance of the standard. It reports the use of the PEMC in evaluating the use of "So, How Are We Doing? A MiBLSi Evaluation Study." The study shows that the PEMC could be a functional tool for a metaevaluation if modified for a specific evaluation. The results of the Kruskal-Wallis one-way analysis show the H test found no differences among the domains (NS). The results of the Cronbach's alpha (CA) test for internal consistency show that item deletion via reanalysis of CA is effective (meaning if the item is deleted the reliability increases), and 26 standards were retained to conduct the CA, and the value obtained was .600.

AUTOBIOGRAPHICAL STATEMENT

I was born and raised in Detroit, Michigan. I grew up with my parents and older sister. For the most part, I liked school and always felt that I was good at math. I did well in most of my classes. However, I was bothered by the difference in the English I used at home and the English I studied in School. I strived to communicate by using the English I learned in school.

I was better than the average student throughout my academic life. I never participated in many sports because I was more interested in the academics. I was in the Spanish Club and completed eight semesters of Spanish during high school. After graduating from high school, I joined the military service and served in the U.S. Marine Corps for 8 years as an Aviation Radar Technician. I spent 36 months overseas including a tour in Vietnam. After my discharge from the service, I worked as a Communications Technician, Market Administrator, and Engineer for Michigan Bell Telephone Company (MBT). Then I retired after 15 years due to the break up of AT&T, the parent company of MBT. I continued to work in the communications field until 1988; at that time I stopped working and attended Marygrove College as a full-time student. I received my Bachelor of Arts (BA) in Social Science in about 18 months. After graduation I began a master's program in sociology at Wayne State University. I received my first master's degree in 1994 and my second master's degree in Community Counseling from Wayne State University in 1999. I began the Education Evaluation and Research doctoral program in 2005.

I am honored to have the opportunity to serve Combat Veterans with their reintegration to civilian life. I am thankful for the challenges I have faced over the years that have built my character and positive attitude. I am grateful for the professors, family, and friends who have encouraged me during my journey to obtain my Ph.D.