Using Citation Analysis Methods to Assess the Influence of STEM Education Evaluation

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Lija Ozols Greenseid

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Frances Lawrenz, Adviser

May 2008

ACKNOWLEDGEMENTS

Many people provided me with the support I needed to complete this dissertation. First and foremost I want to acknowledge my biggest supporter and best friend, my husband, Andrew. He always believed in me even on those tough days when I doubted I would finish. Additionally important, he provided me with the time and space needed to complete my writing by being a great dad. There were several hard evenings when I left our daughter, Arija, crying in his arms saying "no Mommy work" as I packed up my laptop to head for a coffee shop to write. His reassurance that she would be fine within minutes of my leaving made it possible for me to be productive and still feel like a good mom. Although neither can understand it at this point, I would be remiss if I did not also thank my adorably intense two-year old daughter Arija and her little brother- or sister-to-be for providing me with the motivation I needed to finish my dissertation so I could more fully enjoy life with them in the future. I love all of you very much.

While not quite family, my incredibly supportive committee deserves acknowledgement. My adviser, Frances Lawrenz, has been all things a doctoral student wants in an adviser. She is not only brilliant, productive, and politically-savvy, but has also been a terrific mentor and friend since I first took her qualitative research class as a Master's student. Working with her closely on evaluation and research projects over the last five years has helped me grow from a novice evaluator to one with confidence and a diverse toolkit of practices. Frances's work ethic and encouragement to be "productive not perfect" – although most of what she produces is pretty near perfect anyway – will stay with me in all my future projects.

Jean King has also been an incredible mentor as I have joined the field of evaluation. Jean has provided me with many rich opportunities and experiences. I will never forget (or forgive) her for asking me to be a panel respondent to a talk given by Michael Scriven – a talk that he composed on the plane and I did not have any advance copy to read. I will never be afraid of a public speaking engagement again after that experience!

My other two committee members – Michael Rodriguez and Joan Garfield – have also been supportive of me throughout my doctoral career and dissertation phase. Their insights and probing questions about my research topic have helped to strengthen this study in countless ways. Moreover, their collegiality and responsiveness have made this process both enjoyable and a model of what a committee should be.

To all the rest of my family and friends – a huge and heartfelt thank you for your encouragement even when I was not able to explain well what my study was all about. The text that follows is the elevator speech I never did perfect. I guess I do better when I have 200 pages rather than 2 minutes to explain my work. Thanks for listening anyway and celebrating with me on this important accomplishment.

ABSTRACT

This study explores the validity of using citation analysis methods as a way of assessing the influence of program evaluations conducted within the areas of science, technology, engineering, and mathematics (STEM). Interest in the broad influence of evaluations has caught the attention of evaluation theorists, practitioners, and funders recently. However, methods for measuring the influence of evaluations have yet to be developed and validated. Citation analysis is widely used within scientific research communities to measure the relative influence of scientific research and/or specific scientists. This study explores the applicability of citation analysis for understanding the broad impact of STEM education program evaluations.

Nine assumptions regarding the validity of using citation analysis methods to assess STEM education evaluation product influence are examined using data from four sources: (1) citation analysis data, (2) the opinions of an expert panel, (3) data from a survey of primary investigators and evaluators from local projects connected with four national program evaluations, and (4) a review of relevant literatures. The data collected for the validation study suggest that citation analysis methods provide data to help understand, to a limited extent, the influence of large-scale program evaluations on the fields of STEM education and evaluation. In particular, citation data can be used to understand and compare patterns of influence of multi-site STEM program evaluations.

Citations, however, are only one among many possible measures of one limited type of influence arising from the dissemination of evaluation products. Additionally, citation data do not appear to be useful for precisely quantifying the actual level of influence of any one evaluation. Moreover, the examination of the content of citations is

critical. Without understanding the content of the citations, judgments cannot be made about whether citations are actually measuring influence. Consequently, it is important to stress that citations are only one measure of one possible influence arising from an evaluation and are limited and should be interpreted as such.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Over the last decade, some evaluation sponsors became interested in how educational evaluation efforts contribute to building generalizable knowledge within the field of education. A number of large-scale, multi-site evaluations of science, technology, engineering, and mathematics (STEM) education programs funded by the National Science Foundation (NSF) since the mid-1990s exemplify this trend. NSF encouraged these evaluations to disseminate findings and lessons learned related to general topics of interest to the STEM education and STEM evaluation fields. The agency even provided additional years of funding to support these activities (Greenseid & Toal, 2006; Gullickson, Wingate, Lawrenz, & Coryn, 2006). Discussion of this phenomenon was recently described in an article about the Advanced Technological Education (ATE) program evaluation,

> Throughout the years… the environment at NSF changed… It came to view itself as a research and development enterprise and therefore one with responsibility for disseminating results broadly across the nation. More and more, evaluation was asked to play the role of informing the field rather than just informing NSF (Lawrenz, Gullickson, & Toal, 2007, p. 280).

There is debate in the field of evaluation about whether the generation and dissemination of generalizable knowledge is an appropriate primary purpose for evaluations. Many evaluation theorists include the generation of knowledge to be among the core purposes of evaluation. Patton states that the three primary purposes of

evaluation findings are rendering judgments, facilitating improvements, and/or generating knowledge. He discusses that while all three purposes may be found in some evaluations, most often one of the purposes is dominant in any particular evaluation (Patton, 1997, p. 65). The commonly used Rossi, Lipsey, and Freeman evaluation textbook also states that evaluations can "contribute to substantive and methodological social science knowledge" (2004, p. 2). Some evaluators, however, believe that the generation of knowledge is the purpose of research enterprises, while evaluation should be concerned primarily with judging the merit or worth of an evaluand. This perspective is present in the Fitzpatrick, Sanders, and Worthen textbook where the authors state that contributing to knowledge development is only a secondary concern of evaluation (2004, p. 6). Similarly, Alkin and Taut argue that there are different expectations for knowledge produced for research and evaluation, the former emphasizing scientific rigor and generalizability and the latter being measured by its usefulness (2003, p. 10).

In addition to debates about the purposes of evaluations, questions about the use and influence of evaluations have concerned evaluation practitioners and theorists for over thirty years. Since the 1970s, scholars have written numerous articles examining decision makers' use of evaluation data and processes. Now classic works written by Patton, Grimes, Guthrie, Brennan, French, and Blyth (1977), Alkin, Daillak, and White (1979), King and Pechman (1982), and Brown, Braskamp, and Newman (1978), among others, were all produced between the late 1970s to mid-1980s, a time period referred to as the "golden age" of evaluation use research (Henry & Mark, 2003b, p. 294).

In the last decade, the broad impact or "influence" of evaluations became a topic of interest within the evaluation community. Several scholars have advocated studying

the intangible influences of evaluations on programs, individuals, and society, in addition to examining the impact of evaluations on decision makers or stakeholders (Henry, 2000; Henry & Mark, 2003a; Kirkhart, 2000; Mark & Henry, 2004). Kirkhart (2000) presents an integrated model that describes three dimensions of evaluation influence: source, time, and intention. Kirkhart uses the term influence to describe all possible impacts of evaluations, while the term use is used to describe outcomes that are direct, intended, and tangible. Alkin and Taut (2003) adapt Kirkhart's model to distinguish between evaluation use and evaluation influence. They also add a dimension of awareness to Kirkhart's framework. Alkin and Taut define evaluation use as impacts that are aware/intended or aware/unintended and are immediate or shortly follow an evaluation. Evaluation influence they define as being those impacts that are unaware/unintended and/or arise after an evaluation's conclusion, the white areas on the cube in Figure 1.

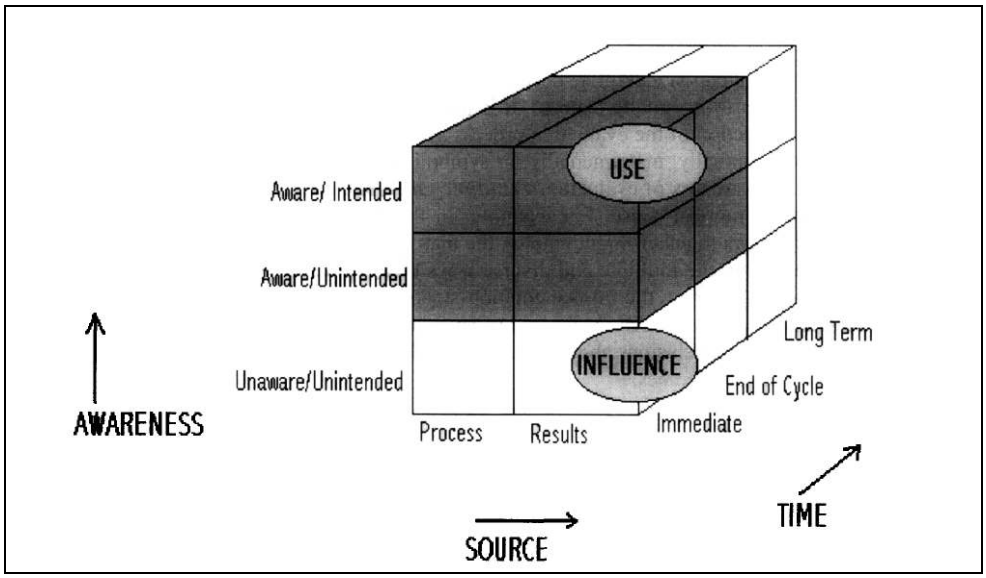

Figure 1. Evaluation Impact within the Program Context (Alkin & Taut, 2003, adapted from Kirkhart, 2000)

As illustrated above, the terms evaluation "use," "utilization," "influence," "impact," and occasionally "consequences" are all applied in the literature written about this subject. In this paper, the term evaluation use is used to describe the more immediate and direct impacts of evaluation processes and findings, usually on decision makers connected directly to an evaluation. The term evaluation influence describes the long-term and indirect impact of evaluations, primarily on individuals not connected directly to the evaluation. The term impact is used to describe any effects arising from an evaluation – encompassing both uses and influences.

While interest in the broad influence of evaluations caught the attention of evaluation theorists, practitioners, and funders during the past several years, methods for measuring the influence of evaluations have yet to be developed and validated. The purpose of this study is to assess the usefulness of one method – citation analysis – for

studying the influence of evaluations on one specific arena in which evaluations are conducted – science, technology, engineering, and mathematics (STEM) education. Using Alkin and Taut's modification of the Kirkhart framework, this study examines the influence sections of the Alkin and Taut cube. Specifically, influence arising from evaluation products (which is related to evaluation results, but also from instruments), primarily takes place in the long term and is both intended and unintended but is in the category of unaware influence. Using the logic-model framework of influence proposed by Henry and Mark (2003), this study contributes to the understanding of one possible pathway to influence. Figure 2 provides a description of the specific pathway under examination.

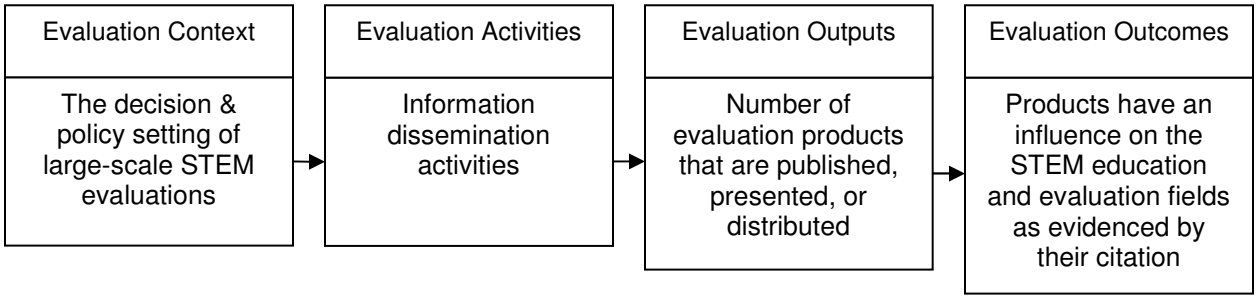

Figure 2. Logic Model of Evaluation Influence in Operation in this Study

Specifically, this study examines how to measure the impact of STEM education program evaluation products on the fields of STEM education and evaluation. Enumerating the number of products produced and describing the dissemination efforts undertaken by the program evaluators provides one measure of the extent of evaluation dissemination activities. It is necessary, however, to take the examination a step further to assess the actual impact of an evaluation's dissemination activities, in other words, to

make people aware of that which they are currently unaware. It is important to acknowledge that the generation of knowledge is just one of many possible pathways to influence. Not under consideration in this study are any of the possible uses or influences arising from the evaluation process; nor are influences related to changes at the interpersonal or collective levels being examined. While this study is examining the impact of evaluation products on the "fields," it is influence and use at the individual level, in terms of new knowledge making an impact on an individual's thinking or documentable uses of evaluation products, that are being measured here.

Purpose and Rationale of the Study

As stated above, the purpose of this study is to examine the usefulness of citation analysis methods for measuring the influence of knowledge-products created through STEM education evaluation efforts. Developing useful measures of evaluation product influence will help advance research on the topic and contribute to further understanding of the construct of evaluation influence. Calls for systematic, rigorous research on evaluation practices abound (Cousins, Goh, Clark, & Lee, 2004; Henry & Mark, 2003b). While several theoretical frameworks of influence have been developed, as discussed above, these frameworks serve to map out the concept and identify channels within which evaluations may have an impact. However, no specific measures of evaluation influence have yet been developed. The challenge is that it is difficult to measure the outcome of interest (i.e., the influence of evaluation knowledge products on fields of inquiry).

First, the broad influence of an evaluation cannot be measured or observed directly. Unlike decision makers' direct, instrumental use of evaluation results and

processes, the broad influence of an evaluation is indirect, usually unintended, and takes place at some time or distance from the actual conduct of the evaluation. Second, traditional social science research methods are not well-designed for measuring diffused influences. Survey methods may seem an obvious choice for gathering data on the extent to which evaluation findings influence individuals within a field. A survey is problematic, however, as it would be difficult to construct an accurate sampling frame for such a study, response rates would likely be low, results would be biased as individuals who respond to such a survey are likely to have been more highly influenced, and measurement error would be high as it would be difficult for individuals to remember and attribute influence to a particular study accurately.

So how can influence be measured? Logically, for influence to occur, an individual must have contact with an evaluation's products. Evaluation products are defined as any publications, presentations, instruments, or other related materials that were produced as part of an evaluation project. Often, evaluation products present information about the evaluation's findings, however, many products also share methods or are more theoretical in nature. Sometimes evaluation products are disseminated to an evaluation's stakeholders through mailings or presentations; however, many products are written publications that are accessible through evaluations' web sites or journals.

Enumerating the number of products an evaluation produces, however, is not sufficient for assessing its impact. Impact can only be measured using some indicator that an individual has actually read and been influenced by an evaluation product. One direct measure of whether an individual has been influenced by a particular publication is if that individual cites the work when writing his/her own publications. References

within publications are a convention used within scholarly communities to acknowledge the value of the influential work or ideas of other scholars (Merton, 1988). As argued by ISI/Thomson Scientific, the company that developed the Web of Science citation index,

> Citation is a direct measure of influence on the literature of a subject, and it is also a strong indicator of scientific contribution, since it is derived from patterns of interaction among millions of published articles. When one researcher cites another's work, he/she is acknowledging the relevance of that work to the current study. The interaction is both highly specific, and highly informed; it is a statement by an author of the scholarly relatedness of two works (ISI/Thomson Scientific, 2007).

A note on the terminology used in this paper. Throughout this paper, the term "citation" is used to describe the referencing of a document by a more recently published document. The document doing the citing is referred to as the "citing work" and the one receiving the citation is the "cited" work, which in this study is an evaluation product.

Methods of citation analysis, developed in the field of library informatics or bibliometrics, are widely used within scientific research communities to measure the relative influence of scientific research and/or specific scientists. Scholars in the natural sciences are often evaluated on the results of analyses of citations of their work in which their publications are ranked based on the frequency with which they are cited in other works. Certain fields even use indices of researcher productivity and influence to make high-stakes decisions about promotion and tenure within academic organizations. To this author's knowledge, however, the field of evaluation has not used citation analysis methods to assess the influence of program evaluation dissemination efforts.

Research Question and Method

The central research question guiding this thesis is to what extent are citation analysis methods useful for measuring the influence of evaluation products on the fields of STEM education and evaluation? At its core, this is a question about the validity of using citations to draw inferences about the influence of an evaluation product. Validation is a process of evaluating the extent to which interpretations of measures of a trait, behavior, construct, or any other inference are plausible and appropriate in a specific context and for specific uses (Kane, 2006). As Kane describes, validation is a process of collecting and presenting evidence of validity through the framework of an interpretive argument (Kane, 1992). While validation studies are often conducted in the context of inferences and decisions drawn from test scores, Messick states that the "principles of validity apply not just to… test scores… but also to inferences based on any means of observing of documenting consistent behaviors or attributes" (Messick, 1995, p. 741). Additionally, Messick argues that the process of construct validation does not mean that the measure is the operational definition of a construct. Instead, "the measure is viewed as just one of an extensible set of indicators of the construct" (1995, p. 742).

This study will use Kane's argument-based approach to validity as the framework for evaluating the usefulness of citation analysis methods for measuring evaluation product influence. Messick's descriptions of different categories of validity evidence, united under his unified theory of validity, also serve as a strong influence on this study. Additionally, this study is positioned within the ongoing developments in evaluation use and influence theory. In the end, the debate over whether knowledge development and dissemination should be a primary purpose of evaluation may be an academic discussion.

The reality is that some sponsoring agencies, such as NSF, are currently funding evaluation efforts with one goal of producing generalizable knowledge. In light of this new reality, research on how evaluations can contribute to growth of knowledge within a field is needed, and a first step in conducting such research is the development of methods that measure the influence of evaluations. This study will advance methodological knowledge in the field of evaluation by exploring the use of citation analysis methods for measuring the influence of evaluation products on the fields of STEM education and evaluation. Additionally, this paper contributes to building a theory of evaluation influence and adds to knowledge about evaluation practices related to use and influence.

The next chapter reviews the theoretical and empirical literature related to evaluation use and influence and provides an overview of the history and validity of citation analysis methodology. Chapter three describes the processes used to gather evidence to assess the usefulness of citation analysis methods for measuring STEM evaluation influence. Chapter four presents the findings from the validation study. The final chapter draws conclusions about the validity of using citation analysis methods for measuring evaluation influence, the applicability of citation analysis in STEM education evaluation and other contexts, and limitations. Finally, implications for theory and practice and areas for future research on this topic are discussed.

CHAPTER TWO

LITERATURE REVIEW

This chapter begins by describing the historical development of the evaluation use and influence literature over the last thirty years, exploring both theoretical and empirical developments and how these may be related to societal changes. Then, this chapter presents a short history of the development of citation indexes and analysis methods.

Evaluation Use and Influence

"In Search of Impact: An Analysis of the Utilization of Federal Health Evaluation Research" is the title of the classic paper published in 1977 by Patton and colleagues at the University of Minnesota (Patton et al., 1977). The work is one of the first empirical efforts to understand how decision makers use evaluations. It is not surprising that the use of evaluation processes and findings has been a major focus of empirical research within the United States' evaluation community[1] since evaluations became widespread with the proliferation of Great Society social programs in the mid-1960s. It can be argued that evaluation as an informal analytical approach has a history as long as that of humanity, however, formal program evaluation was not widespread until the mid-1960s. The Elementary and Secondary Education Act of 1965 mandated evaluations of Title I and Title III education programs, sparking a subsequent proliferation of educational

[1] A similar discussion has occurred in the field of knowledge utilization about the use of social science research information. For comparisons of the research and conceptualizations in the knowledge utilization and evaluation utilization fields, see Hofstetter and Alkin (2003) and Cousins and Shulha (2006). This paper, however, focuses only on the discourse and research on utilization within the field of program and policy evaluation within the United States of America.

program evaluations, evaluation training programs, and evaluators (Worthen, Sanders, & Fitzpatrick, 1997).

Many evaluators believe that program evaluation is a wasted expenditure if evaluation processes and findings have no impact on policies, programs, or, in some evaluators' minds, on society at large. Not all agree that the focus on use is worthwhile, however. Scriven in his *New Directions for Program Evaluation* volume entitled "Hard-Won Lessons in Program Evaluation" makes the distinction that "even utilization does not ensure utility" (Scriven, 1993, p. 75). Scriven argues that,

> even if an evaluation is used, this does not establish that it was useful (had utility), only that it was usable. *Useful* is an endorsement, a favorable (meta-)evaluation; *usable* is the property of qualifying for use and describes a minimal capability, an entry requirement. [emphasis in original] (p. 76)

He explains that the use of invalid or poor quality evaluations may lead to bad decisions (an example of usable but not useful) while other evaluations may justifiably not be used because they did not address decision makers' concerns, were not timely, or were of poor quality (therefore being neither usable nor useful). Rather than focusing on use, Scriven encourages studying the practice of evaluation to improve the quality of evaluations being conducted. Interestingly, it appears that Scriven may have quite recently reversed his earlier position where he argued against studying the use of evaluations but rather spending energy trying to improve evaluation practice (Scriven, 1993). In an editorial published in the summer of 2007 in the online "Journal of MultiDisciplinary Evaluation," Scriven argues it should be "standard operating practice for all continuing evaluation

relationships" to conduct "a minor study of the impact of… prior recommendations" (2007, p. iii).

Despite some disagreement as to the importance of evaluation utilization, the study of use has been a primary research focus of the field for the last 40 years. Like program managers who want to know if their activities are making a difference, the evaluation community has spent considerable time and energy collecting data to determine whether evaluation activities are used. This literature review examines three key questions that have implicitly or explicitly guided much of the last four decades of empirical research on evaluation use: (1) are evaluations used? (2) what can evaluators do to increase use? and (3) what is the effect of evaluations? This review uses a historical perspective to examine the evolution of these key questions by relating trends in the research to environmental forces within and outside of the evaluation community.

As shown in Figure 3, the distribution of studies of evaluation utilization over the last forty years follows almost a bell-shaped curve, peaking in frequency in the mid-1980s. Mirroring this distribution, this review focuses on work conducted during what has been called the "golden age" of evaluation use research (Henry & Mark, 2003b, p. 294): the decade from the mid-1970s to the mid-1980s. Additionally, however, this review highlights key studies conducted preceding and following this most productive decade to show how trends in the research were born and developed over time.

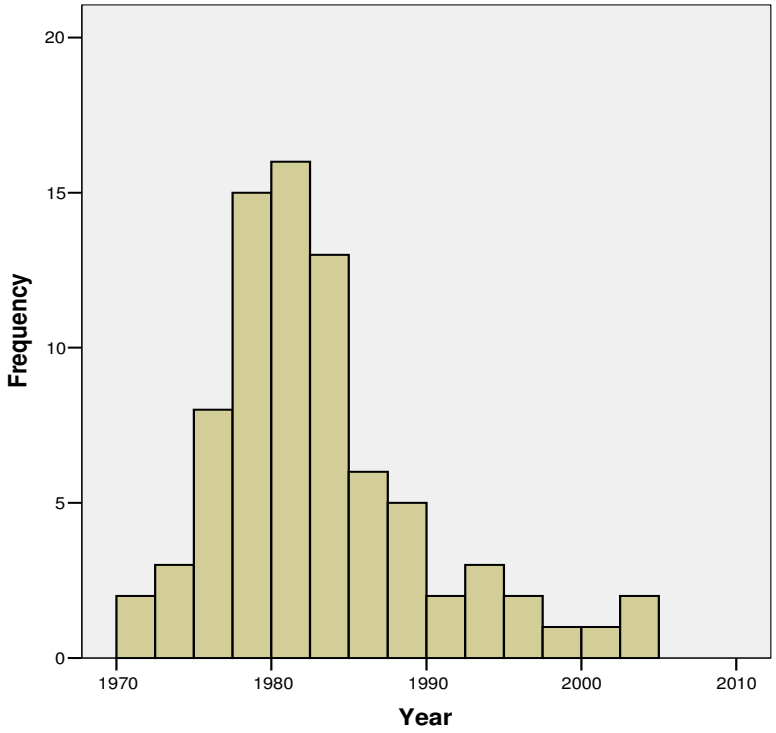

Figure 3. Number of utilization studies conducted per year (1970 and 2006) [Data from Cousins and Leithwood (1986) and Greenseid and Toal (2006)]

Evaluation Utilization Research during the 1970s

The question of the non-utilization and under-utilization of evaluations arose during the decade of the 1970s, a time of economic and political uncertainty resulting in questioning of the worth of the prior decade's large investment in social programs and related evaluation activities. Reflecting on this historical context, it is understandable that evaluators became increasingly concerned about the utility of their evaluations in light of economic uncertainty due to recessions and inflation, perceived failures of many Great Society programs in conquering societal ills, and the Watergate scandal that led to great mistrust of the federal government. As quoted by the then chairman of the

Committee on Labor and Human Resources in the foreword to a volume entitled "Evaluation in Legislation" published in 1979, "politics has gone from the age of 'Camelot' when all things were possible to the age of 'Watergate' where all things are suspect" (Williams, 1979, p. 8).

Several practicing evaluators of the time asserted that evaluation findings were rarely used to inform policy or decision making. As Alkin, Daillak, and White state in the first chapter of their 1979 study of evaluation use, the "game of collecting quotations on this [i.e., dissatisfaction with the usefulness of evaluations] is almost too easy to play" (Alkin et al., 1979, p. 15). They continue with the thought, "one almost wonders how evaluation still manages to survive given this clear consensus that it just does not influence programs" (p. 15). Examples of the concern over evaluation non- or under-utilization were widespread during the seventies. Weiss asserted that "a review of evaluation experience suggests that evaluation results have generally not exerted significant influence on program decisions" (Weiss, 1972a, pp. 10-11). Within the same edited volume Guba bemoaned the state of evaluation methodology, crying that extant evaluation procedures produced evaluation results "of little use to anyone" (Guba, 1972, p. 264). Elsewhere, Guba argues "useful evaluation information is not often produced; and even when it is, decision-makers and policy formulators sometimes see fit to disregard it" (Guba, 1978, p. 1). Rippey captures the ethos of the time: "at the moment, there seems to be no evidence that evaluation, although the law of the land, contributes anything to educational practice other than headaches for the research, threats for the innovators, and depressing articles for journals devoted to evaluation" (Rippey, 1973 as cited in Alkin et al., 1979).

During the early 1970s, assertions about the lack of use of evaluation information were based primarily on evaluators' perceptions rather than demonstrated through empirical study. Weiss is credited (Hofstetter & Alkin, 2003, p. 204) with being the first evaluation theorist to call for the empirical investigation of the use of evaluation information in her 1972 chapter entitled "Utilization of Evaluation: Toward Comparative Study" (Weiss, 1972b). Weiss's paper, first presented in 1966 at an American Sociological Association meeting, proposed empirical research of evaluation utilization to address the "frequent failure of decision-makers to use the conclusions of evaluation research in setting future directions for action programs" (p. 318). Weiss called for the empirical examination of a range of evaluation conditions theoretically believed to be associated with utilization, including the direction of the results (positive or negative), organizational issues, and evaluation procedural issues such as the identification of potential evaluation users and providing them with the information they need, involving program staff in the evaluation process, timely reporting of evaluation findings, and effective methods of dissemination. Her call was heard, and over the next decade empirical research on the aforementioned evaluation conditions and others was conducted to determine the extent to which evaluation information was being used.

Later in the decade, two influential studies were published that had a large impact on subsequent research and discussion. The studies – Patton, Grimes, Guthrie, Brennan, French, and Blyth's study on the use of evaluation information by national health administrators (Patton et al., 1977) and Alkin, Daillak, and White's study of the impact of evaluations of five Elementary and Secondary Education Act programs (Alkin et al., 1979) – are exemplars of research addressing the concern of underutilization during this

period. Both of these studies explicitly state that their purpose is to investigate whether evaluation activities were being used, and both advocate for expanding definitions of evaluation use and using naturalistic methods for uncovering uses. In these ways, both Patton and Alkin countered the then current charges of underutilization and mapped out pathways for future research on evaluation use.

Patton: The Origins of "Intended Uses for Intended Users"

In 1977, Patton and a team of graduate student researchers from the University of Minnesota's National Institutes of Mental Health evaluation methodology training program conducted 20 case studies of national health program evaluations to examine the extent to which the evaluations were used (defined broadly by the interviewees) and what factors interviewees believed were related to the use or non-use of evaluative information (Patton et al., 1977). Patton and colleagues interviewed 60 government decision makers to serve as key informants on the nature and degree of utilization of federal evaluations. Their interviews consisted of broad open-ended questions about the many ways in which decision makers used evaluation information to inform their thinking, judgments, and actions and allowed the decision makers to define utilization in their own terms. After gathering the interviewees' perceptions of evaluation utilization, they asked the decision makers to share why they believed the evaluations had the impact they had. Lastly, they presented the informants with a list of 12 factors culled from the evaluation literature as theoretically linked with evaluation utilization and asked them to comment on the importance of the factors.

Patton's study challenged prior conceptions of evaluation use in several ways. First, the researchers found that evaluations were being used, but not in ways specified in the evaluation literature. Rather than direct utilization, they found that evaluations often served to reduce uncertainty surrounding decision making. As they wrote, "the more typical impact was one where the evaluation findings provided additional pieces of information in the difficult puzzle of program action, thereby permitting some reduction in the uncertainty within which any federal decisionmaker inevitably operates" (p. 145). Patton concluded that conceptualizations of evaluation utilization had been focused too narrowly on instrumental uses, missing important uncertainty-reducing effects of evaluations and therefore underestimating the impact of evaluation activities. As written,

> Our findings, then, suggest that the predominant image of nonutilization that characterizes much of the commentary on evaluation research can be attributed in substantial degree to a definition of utilization that is too narrow in its emphasis on seeing immediate, direct, and concrete impact on program decisions. Such a narrow definition fails to take into account the nature of most actual program development processes. (p. 148)

Second, Patton and colleagues found that the two most important factors cited by interviewees as related to the use or non-use of evaluations were (1) political considerations and (2) what they termed the "personal factor." The first factor related to the political considerations such as budgetary fights, power struggles in Washington, DC, and internal debates about project merits. Of these, Patton found that budgetary issues appeared to be particularly salient. The second factor, the newly coined "personal factor," was not among the list of factors culled from published theoretical works. Instead,

decision makers frequently stated that the fact that a real person cared about and was committed to using the results of the evaluation was an important reason for the impact the evaluation had on their organization. Patton's later work show the influence of this study's findings. His focus on "intended uses for intended users" in his book *Utilization-Focused Evaluation*, soon to be in its fourth edition, provides a framework that demonstrates his commitment to attending to the personal factor when planning for evaluation utilization (Patton, 1997).

Alkin: "Alternative" Utilization and Factors Related to Use

At about the same time as Patton was conducting his study, Alkin, Daillak, and White began a multiple case study that also had a lasting impact on evaluation utilization knowledge and theory (Alkin et al., 1979). Alkin and colleagues conducted retrospective interviews with the operational staff and evaluators and document analyses to produce complete and accurate case studies of five Elementary and Secondary Education Act-funded program evaluations. Although the program evaluations were not randomly sampled (instead, the researchers used their contacts to secure access and find participants for the study), the programs that were included in the study represented a range of types of ESEA programs being funded and evaluations being conducted. In addition to multiple interviews with core informants to gather their detailed descriptions of the evaluations, the researchers were directed to additional interviewees who could provide perspectives on the evaluations and their impact. After draft case studies were written, they were circulated to the key informants who were provided with an opportunity to provide feedback on the cases. The informants' responses are published along with the final case studies to provide an additional degree of validation to the study.

Alkin, Daillak, and White's resulting publication, *Using Evaluations: Does Evaluation Make a Difference?* examines evaluation processes and the ways in which evaluations influenced decisions made about the programs. They address the question of whether utilization occurred by distinguishing between the "mainstream" definition of utilization that stressed immediacy and directness of evaluation impact and an "alternative" conception of utilization that includes impacts that are indirect and occur farther down the road. It is important to reemphasize that the researchers were examining evaluation utilization of ESEA-funded programs, the programs that gave birth to modern program evaluation and were the source of concern about the utility of evaluation efforts. In their choice of topic, therefore, Alkin, Daillak, and White are specifically addressing the field's concerns about the impact of high levels of federal government investment in program evaluation with the passage of the 1965 Congressional education act mandating evaluations. Like Patton, Alkin, Daillak, and White argue that it is important to look beyond narrow, direct, immediate instrumental uses of evaluation to examine broader, longer-ranging effects of evaluation efforts. Consequently, they advocate that naturalistic research methods, specifically case studies, participant observations, and field studies, should be used when conducting utilization studies (p. 32). As evidence of the affinity between Alkin and Patton, Alkin refers to Patton as "a 'kindred soul' in his views of evaluation utilization" (p. 8) in the Acknowledgments section of *Using Evaluations.*

Alkin and colleagues' cross-case analysis found instances of both mainstream and alternatives types of utilization. For example, they found that mainstream utilization occurred in one of their five cases – in this one case, a program was cancelled based directly on the evaluation's findings. In the other cases, they found some examples of

direct ways that evaluations had an impact on the programs, but uncovered many more instances that exemplify their alternative conceptualization of evaluation use. They state that one common mode of evaluation utilization is when evaluation information is one of multiple influences on subsequent actions. As they wrote, "evaluation utilization does often occur, though seldom in earth-shaking ways… the degree of utilization is not determined by mere chance but is, to a considerable extent, associated with certain characteristics of the evaluation situation" (pp. 232-233). Grounded in these data they advance a "theory" of evaluation utilization that outlines relevant characteristics of evaluation situations. Their analytic framework consists of eight categories: 1) preexisting evaluation bounds; 2) orientation of the users; 3) evaluator's approach; 4) evaluator credibility; 5) organizational factors; 6) extra-organizational factors; 7) information content and reporting; and 8) administrator style. These factors shaped future theorizing and research on evaluation use significantly, spinning off several studies in which the categories were tested in varying contexts.

A Separate Sphere: The Simulation Studies

At the same time that scholars were conducting studies of actual use, a robust research program of over a dozen simulation studies were conducted by Braskamp, Brown, Newman, Rivers, and colleagues between 1978 and 1987. These experiments examined hypotheses derived from applying communications theory (i.e., "who says what, how, to whom"), decision theory (i.e., how decision makers make decisions), and attribution theory (i.e., how individuals explain their and others' actions) to evaluation contexts. In each of the studies, simulated evaluation reports, varying only on the

independent variables of interest, were randomly distributed to education audiences accessible to the researchers such as teachers, administrators, parents, education students, and other education professionals. The studies' subjects read the reports and reported how they would respond if they were the decision maker described in the scenario.

This group of scholars examined a number of key independent variables posited to be related to the use of evaluation findings, including evaluation report style (use of jargon and data-based statements) (Brown et al., 1978), evaluator background (researcher, evaluator, or content specialist) and client's organizational role (teacher or administrator) (Braskamp, Brown, & Newman, 1978), gender (Newman, Brown, & Littman, 1979), clients' perceived need for evaluation (Brown, Newman, & Rivers, 1980), different data presentation modes (i.e., amount of data and advocacy vs. adversary argument styles) (Brown & Newman, 1982), and decision-making factors such as conflict levels, importance of the decision, and program setting (Newman, Brown, & Rivers, 1987).

Many of the findings from these experiments now seem self-evident or dated. For example, it is not surprising that audiences in the late 1970s more frequently agreed with the recommendations of male evaluators than female evaluators or that they rated "researchers" to be more objective than "evaluators" or "content specialists." Similarly, it does not take a randomized study to believe that jargon-filled reports without supporting data are perceived to be more difficult to understand than jargon-free, data-supported reports. Moreover, many of the significant findings from these studies were in the form of complicated statistical interactions, leading even the authors to make qualifications about the application of their results to the real world of evaluation. In

summarizing these studies for a *New Directions for Program Evaluation* publication, Newman, Brown, and Braskamp acknowledge that the studies'

> results may not apply equally as well to real-life decision-making contexts. In some instances, the patterns may be more pronounced, whereas in others the real-life situation may introduce intervening variables which might result in entirely different results. A good deal of evaluation feedback takes place in informal contexts, and the same variables may not be operating in the same fashion as they do for formal reports. (p. 33)

Additionally, the impact of these studies appears to be minimal because of an apparent distance between these research efforts and the rest of the evaluation utilization researchers. While Braskamp, Brown, and Newman mention the "under-utilization problem" in the introduction to their first published work (Brown et al., 1978), it is puzzling that none of their subsequent studies make reference to research being conducted by other evaluation theorists. Consequently, although they are researching evaluation utilization, these scholars appear to have been somewhat disconnected from what became the main currents of research in the field. The choice of simulation experiments and their focus on the use of findings without consideration of the influence of evaluation processes runs contrary to the viewpoints of other evaluation researchers. For example, Alkin in an early *New Directions for Program Evaluation* volume on the "Utilization of Evaluation Information" edited by Braskamp and Brown argued that naturalistic methods were important for capturing the "complex, evolving evaluation process" (Alkin, 1980, p. 22). He states that "naturalistic research techniques provide a procedure for concentrating precisely on the unfolding processes that result in observable

outcomes." Alkin's call for using naturalistic research methods may have been a reaction to these simulation studies as well as a reflection of the debates between qualitative and quantitative researchers at the time. Taken together, these issues may help explain part of the relative lack of influence this body of research has had in subsequent decades.

Later Studies: Patton and Alkin's Influence

Empirical research on evaluation utilization conducted after the publication of Patton and Alkin's studies shows the impact of their work, particularly their calls for broadening definitions of evaluation use. Dickey, one of Patton's collaborators, published results of a survey of the utilization of 47 ESEA program evaluations conducted in Minnesota (Dickey, 1980). Dickey coded the projects' final evaluation reports for key independent variables suggested in the literature as related to utilization. Then she surveyed ESEA project directors to determine whether they used the evaluation findings and, if so, in what ways. She stated that like in the Patton and Alkin studies, she chose not to define use or utilization for the project directors, but instead allowed them to share their own assessment of the evaluation's impact. As she stated, "The assumption underlying this inductive approach is that evaluations are likely to have influence on decision making in ways other than direct or immediate impact on program decision making" (p. 69). Additionally, Dickey specifically examined variables suggested by Patton's findings on the importance of the "personal factor" in evaluation utilization, creating decision maker attitude and decision maker involvement scales to measure levels of the personal factor.

Dickey found that despite the rather low methodological quality of the studies, project directors reported that the evaluation process was useful to them and that the evaluation was used most frequently to change or modify their practices (75%) or to reduce uncertainty (72%). Only two of the 47 respondents reported that the evaluations were not useful to them at all. Dickey's follow-up interviews with project directors reporting low levels of use found that these evaluations put an undue burden on project staff to design the evaluation and collect data, that inappropriate data collection instruments were used, and that the evaluation reports contained too much technical jargon. Additionally, political conflicts with the Minnesota Department of Education affected use. Utilization was found to be positively related to decision makers' positive attitudes towards evaluation, but not related to the level of involvement in the evaluation process, and attitude and involvement were not related to each other.

Another study that can be seen as influenced by Patton and Alkin's work was conducted between 1980 and 1982 by King, Thompson, and Pechman. The group conducted research on the utilization of evaluation products and processes by public schools in New Orleans under a grant from the National Institute of Education (King & Pechman, 1982). In the first year of the study, the team produced a comprehensive literature review and annotated bibliography on research on the local use of evaluations in school settings (King, Thompson, & Pechman, 1982). This grounding in the evaluation utilization literature clearly informed their subsequent research. In the second year of the grant, the team conducted in-depth case studies to learn how evaluation use operated on a local level. Citing Alkin's call to use naturalistic methods when studying evaluation utilization (1979), the team conducted year-long case studies using such

methods as interviews with 18 evaluation staff members and 70 school district evaluation users, observations of evaluation staff in meetings and during site visits, questionnaires for administrators, document review, and a series of "instance studies" – conversations about instances of significant evaluation events.

Although King, Thompson, and Pechman found only a few instances of direct, instrumental use of evaluation findings, there were ways that they observed that evaluation information was having an effect on decision makers. Consequently, like Alkin and Patton, they found it necessary to propose an expanded conceptualization of use to better reflect the reality of how evaluation information is used at a local level. Grounded in their case study data, King and Pechman proposed two new functions of evaluation information. The first, "signaling" use is the use of evaluation information and activities as signals from the local agency to funding agencies (p. 36). The second, "charged" use, refers to evaluation information that produces a reaction in the system (p. 40). They also discuss three additional factors that affect use: (1) the distinction between evaluation users' "espoused theories" and their "theories in action," (2) the self-confidence and openness of evaluation users – two new dimensions of Patton's "personal factor," and (3) the clout factor – the importance of key administrative support.

Other lesser-known studies of evaluation use conducted during the early 1980s include Kennedy, Apling, and Neumann's examination of the use of evaluation and testing information in Title I education programs (Kennedy, Apling, & Neumann, 1980) and Kennedy (1984). Kennedy and colleagues conducted a naturalistic case study of decision making in 16 public school districts. They examined qualitative data from interviews with policy makers, program managers, principals, and teachers and from

observations of group meetings for evidence of conceptual use of evaluation and testing information. Specifically, they examined linkages between factual evidence and individual prior knowledge and groups' shared understandings. Kennedy concluded that the process of conceptually using data is more complex than just accumulating findings and then making rational decisions based on the new knowledge. More often, factual evidence is synthesized with other sources of knowledge such as individuals' working knowledge and prior beliefs in the formulation of decisions. She also concluded that the decision making process is not static in time, but rather that ideas are formulated and reformulated based on new knowledge, new experiences, and the formulation of new beliefs. Additionally, decision makers' interpretations of evidence, not the actual evidence, are what are retained and applied in making decisions. These interpretations are often far abstractions at best, or erroneous understandings at worst, of the original evidence.

What overall conclusions can we draw from this golden age of research on evaluation use? Research conducted during the late 1970s and early 1980s found that evaluations were being used, however, scholars argued that "use" needed to be defined broadly and naturalistic methods needed to be used to find it. Additionally, we find that several of these studies began to outline factors related to specific contexts that determined whether evaluations were used. In the next phase of research, we see an increased focus on determining which general factors were related to use through the conduct of meta-syntheses of these individual studies.

The Rise of Meta-Syntheses in the 1980s

While evidence collected during the late 1970s suggested that most evaluations were being used, albeit often in indirect or unintended ways, the issue of how to increase the impact of evaluations by improving evaluation practice became increasingly important to evaluators during the 1980s. Three factors can be argued to have contributed to this new emphasis and to the rise of meta-syntheses of evaluation utilization research as an answer to this call. First, the 1980 election of Ronald Reagan, with his fiscally conservative political theory, presented both challenges and opportunities for evaluators. As written in the 1982 edition of Rossi and Freeman's popular evaluation textbook, *Evaluation: A Systematic Approach, 2nd Edition*, "the last few years have seen questioning of the continued expansion of government programs, resulting in increased requirements for effectiveness and efficiency… the evaluation enterprise must acknowledge the importance of the changing mood and times of the country" (p. 31). In their view, evaluators were uniquely positioned to provide vital information to decision makers attempting to determine which programs were worth continued funding and which should be cut.

Other evaluators, however, felt Reagan's election put increased pressure on the field to demonstrate the value of research and evaluation activities. This sentiment is expressed by King and Pechman, who wrote,

> national economic troubles, coupled with the election of Ronald Reagan as President, have put an end to those early, bright days, and evaluators today, as

> never before, must demonstrate the value of their wares to users who may be unable to afford such "luxuries" in any case (King & Pechman, 1982, p. 1)

A second factor at play during this time period was the birth of the movement toward professionalization of the field of evaluation. The appointment by a dozen leading educational organizations of a committee of educational evaluators and researchers in 1975 and the subsequent publication of the Joint Committee's *Standards for Evaluations of Educational Programs, Projects, and Materials* in 1981 can be seen as early steps in this movement. As stated in the introduction to the *Standards,* "the Committee was also guided in the belief that a set of professional standards could play a vital role in upgrading the practice of educational evaluation" (p. 5). The *Standards* present four sets of guiding principles that represent the attributes of high quality evaluations. The first set of the standards was "utility." The utility standards served to ensure that evaluations will be "informative, timely, and influential" (p. 13). The inclusion of utility standards demonstrates the importance of evaluation utilization to members of the Joint Committee and places on evaluators a responsibility for conducting evaluations in a manner in which the processes and results will be useful to evaluation clients and stakeholders.

Advances in social science methodology are a third factor that may have played a role in sparking new directions in research on evaluation use. In the early 1980s, social science researchers began to value integrative reviews and meta-analyses as forms of research that were complimentary and not just secondary to individual research studies. Moreover, the improvement of quantitative meta-analytic methods sparked a number of meta-syntheses to be conducted in a variety of fields during the early 1980s (see, for

example, Glass, McGaw, and Smith, 1981; Jackson, 1980; Cooper, 1982; and Light and Pillemer, 1984). Meta-synthetic methods, therefore, began to be viewed as holding promise for developing generalizable understandings about the relationships between variables across a variety of contexts, and social science researchers, including evaluation scholars, began to use such methods in their investigations.

Determining the precise effect of societal changes, the evolution of the evaluation field, or advances in social science research methodology on evaluation use research trends is not possible. It is clear, however, that by the early 1980s a number of evaluation researchers and graduate students began work on summarizing what had been learned from the flurry of research and theoretical activity on evaluation use that occurred during the previous decade. Several literature reviews and meta-syntheses describing factors important to the utilization of evaluation information and processes were conducted at approximately the same time and published within a few years of each other. The earliest located of these literature reviews, published in a 1979 edited volume, was conducted by Young and Comtois at the then U.S. General Accounting Office (GAO) during 1978 (Young & Comtois, 1979).

Young and Comtois reviewed the theoretical work and empirical research concerning the utilization of evaluation studies by the federal congress and derived, in a qualitative fashion, a list of six factors posited to be most likely related to the use of evaluation information by the federal government: (1) the political decision-making environment; (2) organizational aspects of the management environment; (3) commitment/involvement of decision makers and evaluators; (4) appropriateness of questions asked; (5) methodology; and (6) dissemination/reporting issues. The authors

go further and offer four recommendations to increase the likelihood that policy evaluations are utilized by Congress based on their review of literature and study of the actual uses of GAO evaluations. First, they assert that evaluations must be utilization-focused and that use must be planned from the beginning of an evaluation. Second, the standards used for assessing whether a study has been used should be negotiated and agreed upon early in a study. Third, the commitment and involvement of key decision makers in the evaluation process is important. Fourth and finally, other factors such as design and dissemination are important but secondary to the interaction factors previously outlined. Young and Comtois' prescriptive recommendations show the influence of Patton's utilization-focused evaluation approach as well as Weiss's understandings of how evaluations function in large policy settings.

Two years later, Leviton and Hughes published a literature review of factors affecting evaluation utilization (Leviton & Hughes, 1981). Although Leviton and Hughes examine evidence of utilization of program evaluations in addition to policy evaluation contexts, they more narrowly define utilization by stipulating that to count as utilization evaluation reports must be seriously discussed, not just read or cognitively processed. Even more restrictively, they state that in order for utilization to have occurred, "there must be evidence that in the absence of the research information, those engaged in policy or program activities would have thought or acted differently" (p. 527) based on the evaluation information. Leviton and Hughes identified five clusters of factors that empirical evidence suggested were related to use: relevance, communication, information processing, credibility, and user involvement and advocacy. In easy-to-read tables, the authors outline variables that inhibit and variables that encourage the

utilization of evaluations. For example, within the "information processing" factor the clear presentation of information is positively related to use, while the use of jargon is found to decrease use. Leviton and Hughes' presentation of the literature in this format provides readers an accessible picture of the complex relationship between a number of variables and evaluation utilization.

At approximately the same time that Leviton and Hughes were conducting their review of evaluation utilization research, Beyer and Trice conducted a synthesis of 27 empirical studies of the utilization of social science research efforts (Beyer & Trice, 1982). Although primarily focusing on the use of social science research, Beyer and Trice include five evaluation studies, including Alkin, Daillak, and White, and Patton et al. Like Young and Comtois, Beyer and Trice outline a number of factors that have empirical support for being positively related to use and then, drawing on the identified factors, advance a number of recommendations to increase the utilization of social science knowledge and evaluation. Interestingly, many of the recommendations they advance for researchers are the same as those discussed by evaluation theorists, such as the importance of attending to the needs of potential users and using mixed-methods to increase the validity and accessibility of results to users. Additionally, they argue that knowledge utilization researchers must expand their conceptions of use to capture the full range of uses of social science research, the same argument being advocated by evaluation utilization researchers.

In a 1985 publication directed at evaluation decision makers, Alkin and colleagues present a taxonomy of factors related to the use of evaluation information based on a synthesis of the evaluation use research literature conducted by Burry (Alkin,

1985). Three broad categories of factors were identified that affect the degree to which evaluation information is used: human factors, context factors, and evaluation factors. The human factors include factors related to the characteristics of the evaluator such as his or her commitment to use, willingness to involve evaluation users in the evaluation process, rapport, political sensitivity, and credibility. Additional considerations are the evaluators' personal characteristics such as gender, title, or position. There are additional human factors related to characteristics of the evaluation users, such as the users' organizational positions or professional experience levels, as well as their level of interest in evaluation, expectations for the evaluation, and their commitment to using the evaluation results.

Alkin's second category of use factors are contextual factors, including pre-existing evaluation constraints, features of the organization, and characteristics of the project. The contextual factors encompass such issues as contractual obligations and fiscal constraints, internal and external organizational features, and characteristics of the project being evaluated such as its age/maturity, innovativeness, and uniqueness. The third category of use factors are evaluation factors such as the methods used in the evaluation, the amount and quality of interaction between the evaluator and users, the relevance and specificity of the information, and methods of evaluation reporting. Alkin's experiences conducting rich case studies of evaluation use are evident in the identification of factors. Alkin places a greater emphasis on the multi-faceted human, contextual, and evaluation factors that interact to produce a unique evaluation decision-making environment rather than trying to identify a few key factors that appear to have

the greatest potential for increasing use in general. As he writes in the *Guide for Evaluation Decision Makers*,

> a key to understanding the role of the administrator in structuring the evaluation process is to appreciate the complex array of circumstances that surround an evaluation. Within each evaluation situation there are a multiplicity of factors that can have an effect on evaluation use. (p. 24)

The discussion of use presented to decision makers is framed by identifying steps that decision makers can take to structure the evaluation environment in ways to increase the potential for its use.

In 1986 Cousins and Leithwood synthesized the research on evaluation use through a novel, quantitative meta-synthetic process. Cousins and Leithwood analyzed 65 empirical studies of evaluation use conducted between 1971 and 1985 (Cousins & Leithwood, 1986), coding each according to the study's orientation toward dependent variables (i.e., the type of use examined, use as decision making, use as education, use as the processing information, or "potential" use, a catchall for other use types) and orientation toward independent variables (i.e., factors related to use categorized into two broad factors: evaluation implementation characteristics and decision or policy setting characteristics). Next Cousins and Leithwood examined the observed statistically significant relationships between the independent and dependent variables to assess the relative importance of identified factors to types of evaluation use. They conclude that evaluation quality and decision characteristics were the factors most commonly found to be related to use, followed by evaluation findings (e.g., positive findings), users' commitment and/or receptiveness to evaluation, and evaluation relevance.

In contrast to Alkin's review of the empirical literature, Cousins and Leithwood explicitly conducted their meta-synthesis to identify a handful of key variables related to use across a variety of evaluation contexts. Their publication falls somewhat short of this goal, however. When examined closely, it is apparent that their coding scheme identifies which relationships are most often studied rather than identifying the most prevalent relationships that actually exist in evaluation practice. Although some of the simulation studies were experimental in design, the majority of the studies included in the work were retrospective case studies or longitudinal field studies that preclude drawing conclusions about the relative importance of variables in statistical ways, as Cousins and Leithwood attempted to do in their analyses. Nevertheless, their development of an empirically grounded taxonomy of factors related to use addresses many of the same factors identified in earlier literature reviews and provides some measurement of the proclivity of the factors, at least in terms of the prevalence of their study.

In a follow-up study of sorts, Shulha and Cousins provide an updated qualitative review of the literature, both theoretical and empirical, published between 1986, when Cousins and Leithwood's study was conducted, and 1996 (Shulha & Cousins, 1997). Shulha and Cousins identified several themes in the evaluation utilization literature from the mid-80s to 90s. First, they state that attention to context, and, in particular, organizational or programmatic contexts, became increasingly important in the utilization literature. Second, the previous decade saw an increased focus on the uses arising from the evaluation process, as opposed to evaluation findings. In particular, the importance of stakeholder and evaluator collaboration was examined in empirical work by Ayers (1987), Cousins (1996), and Greene (1988; 1987).

Many of the studies included in Shulha and Cousins' 1996 review were theoretical or reflective case narratives as opposed to empirical research studies. In light of that limitation, Greenseid and Toal conducted a comprehensive synthesis of empirical research on evaluation use conducted between 1986 and 2006 taking off where Cousins and Leithwood's study ended (Greenseid & Toal, 2006). Greenseid and Toal distinguished between empirical research studies and "reflective case narratives" (a la Cousins & Shulha, 2006), using only those studies that were determined to be empirical research studies. After a rigorous screening process, 42 studies were included in the final review.

Greenseid and Toal found that research on evaluation use conducted during the last twenty years varied greatly. The studies used multiple bodies of literature to frame their investigations, such as communication theory, decision theory, and evaluation use theory. 86% of the studies referenced evaluation use theories in their framing sections, as compared to those conducted prior to 1986 (according to Cousins and Leithwood) of which only 10% of the studies referenced evaluation use theories. A few studies (approximately 20%) only examined only one type of evaluation use, predominately instrumental use, while the remaining strong majority (80%) examined more than one type of use. Generally, these studies examined the instrumental, conceptual, and symbolic uses of the evaluations under consideration.

In terms of their designs, Greenseid and Toal found that recent evaluation use studies used a wide range of methodological designs: approximately 17% of the studies were experimental or quasi-experimental quantitative studies, 22% of the studies were single case studies, 32% were multiple case studies, 30% of the studies were surveys, and

there was one document review. The methodological quality of the studies was also examined. Overall, two-thirds of the studies were assessed to be adequate or above adequate in terms of their conceptualizing of variables, appropriate match between research questions and designs, grounding in theory and literature, transparency in data collection and analysis, and adequate description and consideration of context. The one-third of studies assessed to be below adequate in nature were limited because of lack of a thorough description of the methods, weak sampling or data analysis methods, inappropriate generalizations, and concern over researcher bias.

Greenseid and Toal also found that there has been a shift over time in the types of factors being examined in relation to use. Since 1986 a significantly larger percentage of the empirical research has investigated the relationship between stakeholder participation/involvement and use than had been examined prior to 1986. Between 1971 and 1986 only 12% of studies included in Cousins and Leithwood's review examined the effects of stakeholder participation on use. In contrast, between 1986 and 2006 over 33% of the studies examined stakeholder participation or involvement. Almost all of these studies make reference to Patton's utilization-focused evaluation approach, another indication of the important impact of his work on the direction of the evaluation field.

Although varying in the specifics, the meta-syntheses conducted since the mid-1980s have developed a number of categorization schemes to better understand the variables affecting use, some within and some outside of the control of evaluators. Contextual factors such as the political decision-making context or organizational issues lie beyond an evaluator's control. However, evaluators can make decisions about the evaluation's design, reporting and dissemination strategies, and level of stakeholder

involvement, factors increasingly agreed upon as having an impact on the extent to which an evaluation is eventually used. Over time, support for planning for use early in an evaluation and involving intended users in this process, as Patton describes in his utilization-focused evaluation approach, began to dominate beliefs about how to increase the use of evaluation processes and findings. This is evident in the results of a survey conducted by the American Evaluation Association's Evaluation Use Topical Interest Group in 1997 in which 90% of respondents believed it was extremely or greatly important to plan for use at the beginning of an evaluation and approximately 83% believed it was important to identify and prioritize intended users and uses of the evaluation (Preskill & Caracelli, 1997).

Broadening the Research Agenda to Evaluation Influence in the New Millennium

It may be somewhat premature to make judgments about new directions for research on evaluation use in the new millennium. One issue, however, seems to have captured the most interest among evaluation use theorists and researchers. In recent years, calls for broadening the focus of evaluation use research have arisen from within the evaluation community – particularly by a number of scholars who have had leadership positions within the American Evaluation Association and evaluation journals. These scholars have advocated studying the intangible influences of evaluations on programs, individuals, and society, rather than focusing solely on the impact of evaluations of decision makers or stakeholders. While the concept of evaluation influence was advanced earlier (Merwin, 1983), it has only been in the last few years that

the term "evaluation influence" has been discussed broadly by evaluation theorists and explicitly examined by evaluation use researchers.

Evaluation Influence Literature

Kirkhart's publication in 2000 of an article entitled "Reconceptualizing evaluation use: An integrated theory of influence" has been credited as launching the recent focus on broad evaluation influence. Kirkhart defined influence as "the capacity or power of persons or things to produce effects on others by intangible or indirect means" (Kirkhart, 2000, p. 7). More a conceptual framework than a theory, Kirkhart maps out three dimensions of evaluation influence: Source, Intention, and Time, represented as the three sides of a cube-like figure (see Figure 4). Traditional instrumental use is placed on the cube as Results, either Intended or Unintended, and usually End-of-Cycle; while conceptual use could arise from any intersection in the cube. In this sense, use is viewed as those outcomes that are direct, intended, and tangible, while influence is a broader terms encompassing all possible impacts of an evaluation. Kirkhart argues that her framework should assist evaluation use researchers to conceptualize better and, therefore, conduct better studies of the broader impacts of evaluations.

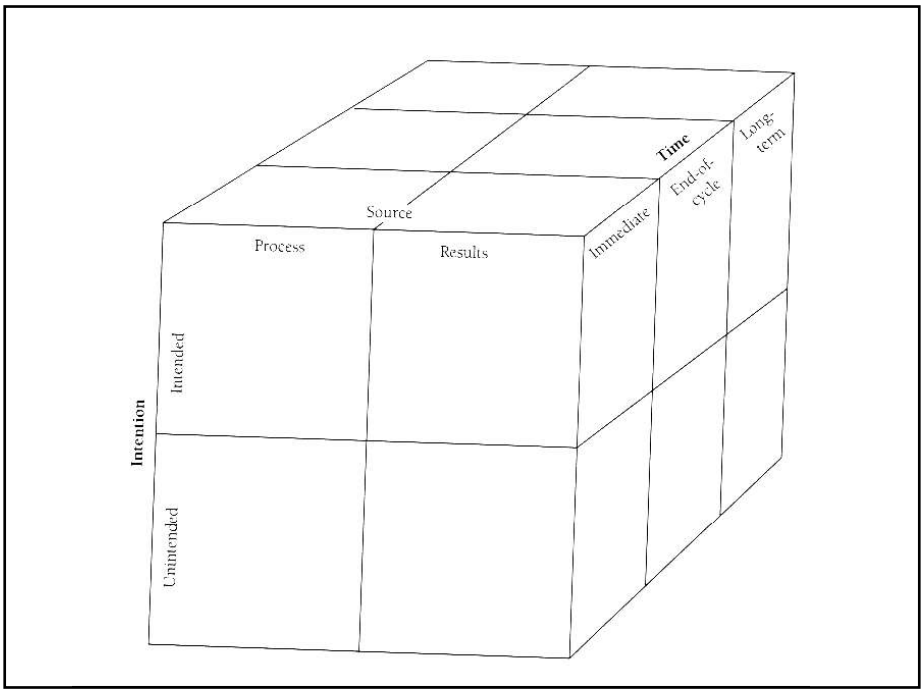

Figure 4. Kirkhart's "Integrated Theory of Influence" (2000)

Alkin and Taut (2003) challenge Kirkhart's assertion that shifting terminology from "use" to "influence" is helpful for building evaluation theory and instead believe that use and influence should be used to describe different types of impacts arising from evaluations. Alkin and Taut modify Kirkhart's influence cube, by adding a dimension of awareness in addition to intention (Figure 1, Chapter 1). Awareness is defined as those impacts that primary intended users can specify. Those impacts that primary users cannot name would be considered influence. Influence, therefore, is defined as those impacts that are unaware and unintended or those impacts arising in the long term. Both influence and use can come from the evaluation process or results. In addition to distinguishing between evaluation use and influence, Alkin and Taut clarify the distinctions between conceptualizations of "process use" and "findings (or results) use" that are found in the literature (e.g., Patton, 1997). They state that rather than being an

additional type of use, process use is better understood as an additional source where use can arise. They illustrate that both findings use and process use can be instrumental, conceptual, or legitimate/symbolic, see Figure 5.

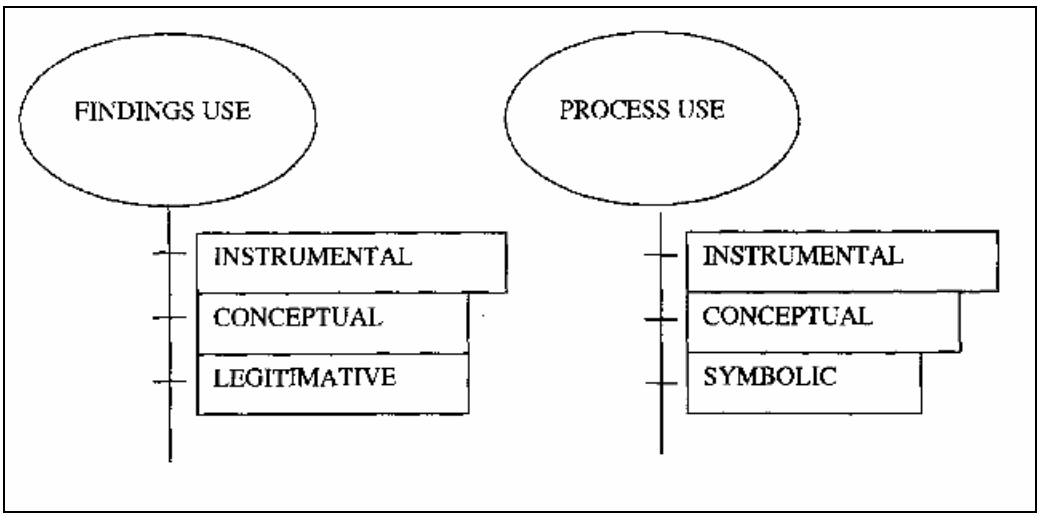

Figure 5. Types of Evaluation Use (Alkin & Taut, 2003, p. 7)

Alkin and Taut argue that while the influence of evaluations is important to study, the use of evaluation should be the primary concern of evaluators. As they write, unaware/unintended impacts

> …constitute an important aspect of an evaluator's work – that merits attention from evaluation research and practice. However, it seems not as essential to the evaluation profession as those impacts that are of a conscious (either intended or unintended) nature, in the eyes of the users, and hopefully the evaluator as well. (p. 9)

They continue in arguing that because influence, by definition, is unintended, "it is outside the domain of the evaluator to affect such possible influences" (p. 9).

Mark and Henry, in a series of related theoretical publications (Henry, 2000; Henry & Mark, 2003a; Mark & Henry, 2004), join Kirkhart in advocating for a broadening conceptualizations of the consequences of evaluations. Unlike Alkin and Taut, however, Mark and Henry argue that the goal of evaluation is broad social betterment, not merely use. Evaluators, they state, must look beyond use to identify the long-term impacts of their work. As Henry stated in a recent interview,

> Ultimately, we should be concerned with an evaluation's influence on the beneficiaries of a program or policy, and look at whether people are better off as a result of the evaluation… By changing the term from use to influence, it becomes easier to think about evaluation as an intervention and to seek the broader ways in which evaluations, or the evaluation process itself, influences social betterment in the long term (Henry, 2005, p. 10).

Henry and Mark propose an evaluation logic model that includes a number of possible influence mechanisms operating at three levels of influence (individual, interpersonal, and collective) that produce changes leading to social betterment. Mark and Henry derived their model by reviewing a variety of social science research literatures to identify promising mechanisms or processes that they believe could help explain how evaluations have an impact. They believe that their framework, although neither exhaustive nor parsimonious enough in its current rendition, should be beneficial to future evaluation influence researchers who are searching for ways of studying the broad consequences of evaluations. As program theory is used by evaluators who want to understand the complex outcomes of a program intervention, they believe evaluation

use researchers can use their evaluation influence logic model to guide the difficult work of identifying the consequences of evaluations.

Empirical Research on Evaluation Influence

As with many topics in the field of evaluation, theory has preceded empirical investigation of the concept of evaluation influence. Despite the amount of interest in the topic of evaluation influence, there have not yet been specific methods developed and validated for assessing the broad impact of evaluations on fields of practice. Current research on the topic has used traditional social science research methods with only mixed success. The first published research study to apply Mark and Henry's influence framework was a case study of the influence of D.A.R.E. (Birkeland, Murphy-Graham, & Weiss, 2005; Weiss, Murphy-Graham, & Birkeland, 2005). As the researchers began the study in 2000, prior to Mark and Henry's publications, their work was not developed with the pathways to influence framework in mind. Instead, as noted in a footnote of their article, *American Journal of Evaluation* reviewers and then editor, Mel Mark, suggested that they consider the Henry and Mark framework when they submitted their work for publication.

Weiss, Murphy-Graham, and Birkeland followed their reviewers' advice and attempted to trace the influence pathways evidenced by their data on the D.A.R.E. evaluations. As they describe it, however, they encountered "tangles of interaction" operating in their cases and found the exercise of tracing pathways to be less promising than envisioned by their creators. First, they found that the pathway to influence in each district was different and the task of tracing such pathways was arduous and challenging.

Second, they were concerned about the different perceptions held by each interviewee about the intervening variables that affected their decisions. Moreover, their respondents' memories could have been faulty as they were being asked to report on events occurring several years in the past. Together these issues prevented the researchers from satisfactorily utilizing Mark and Henry's framework to draw strong conclusions about how the DARE evaluations were influential.

Although Weiss and colleagues struggled to apply the Mark and Henry framework, their study is helpful in advancing a new type of evaluation use: "imposed use," which occurs when program stakeholders are required to pay attention to evaluation results (p. 16). This manifested itself in the DARE evaluations as programs were required by their funding agencies to use interventions scientifically proven to be effective, which the DARE program was not. The researchers state that they believe this type of use will become more prevalent in the future as mandates to use evidence-based interventions increase in popularity.

A second research study to consider the Mark and Henry framework was published in early 2007 by Christina Christie (Christie, 2007). Unlike Weiss's study, Christie states that the Mark and Henry framework served as the conceptual foundation and inspiration for her study. She chose to examine just the individual level of their framework and just one of the mechanisms proposed to operate at that level: behavioral change. Using a simulation study, Christie investigates the types of information (large-scale data, case study data, or anecdotal accounts) that decision makers report as having an influence on their decisions. She finds that while all three types of data are reported as having some influence on decision making, in general, large-scale and case study data

tend to be more influential than anecdotal accounts. Additionally, individuals working in educational settings or with educational backgrounds were less likely to report using large-scale evaluation data – a finding Christie believes may be attributable to a backlash against the No Child Left Behind Act's (2001) preference of randomized control trials over other evaluation designs.

Why Christie chooses to link this study to evaluation influence as opposed to the evaluation use literature is not addressed. Her blanket statement in the first paragraph that "use is a central outcome of any evaluation, because without it, evaluation cannot contribute to one of its primary objectives, social betterment" (p. 8) indicates her orientation toward Mark and Henry's beliefs. Her study, however, could easily be reframed as examining the types of information decision makers use in making judgments. Her study could be seen as a revival and extension, 25 years later, of Brown and Newman's 1982 study of the effect of data presentation types on decision makers' decisions in which they considered presenting no data, percentages only, graphs only, and both percentages and graphs. Interestingly, however, she makes reference to that study only in her discussion of simulation study methodology, not in her section on relevant literature.

In addition to the two journal articles mentioned above, Cheng (Cheng, 2006) conducted a case study of evaluation influence in a doctoral dissertation study. Cheng conducted a qualitative case study using retrospective, semi-structured interviews and document review to assess the influence of program evaluations on literacy instruction in two elementary schools within one school district. Cheng found that evaluation data were only one source of information that influenced individual, interpersonal, and

collective understandings and practices. Although Cheng specifically attempted to apply the Henry and Mark (2003b) and Mark and Henry (2004) frameworks, she, like Weiss et al., found this difficult to accomplish. Like Weiss's "tangled webs of influence," Cheng found that the "various types of evaluation process and outcomes were interrelated with one another" (p. 203) and had "difficulty identifying the step-by-step pathways that led to behavioral changes" (p. 207). She found that evaluation use and influence at one level affected use and influence at other levels in sometimes complex ways. In the end, she was able to identify three factors that appeared to be related to evaluation use/influence: human factors, structures/resources, and external factors, similar to those described in earlier studies of evaluation use.

The three studies discussed above use traditional social science research methods, such as retrospective case studies and experimental simulation studies, to gather data on evaluation influence. There are two other studies that use alternative methods for assessing influence. The first study was conducted by the World Bank in 2004 (Bamberger, 2004). This study presented case studies of eight World Bank projects deemed to be "highly influential" in that they were able to make a large impact on governmental policies or practices. In addition to presenting evidence of the impact of the projects, the projects were also assessed for their cost-effectiveness by comparing the cost of the evaluation to the savings recovered by the governments as a result of following the evaluation's recommendations. The report presents lessons learned through their study of these eight exemplary evaluations. They conclude that evaluation influence is greatest when (1) the evaluation addresses current concerns and there is a commitment by decision makers to use results; (2) the evaluation was only launched after

establishing clearly defined information needs; (3) the evaluator understands that evaluation is only one source of data within a decision-making context; (4) there is a strong relationship with the client and findings are effectively communicated; and (5) the evaluation is conducted by either the evaluation unit of the managing or funding agency or by an outside agency, or jointly, as the context dictates.

The second study, conducted by the Editorial Projects in Education Research Center (Swanson & Barlage, 2006), took a different approach to studying factors that influenced national educational policy. While not a study of program evaluation influence exclusively, the study's innovative methodology warrants inclusion in this discussion. The study combined surveys of educational policy experts with citation analyses to create an "influence" index to rank influential studies, organizations, people, and information sources. Most of the top ranked influential studies are research studies or even long-term data collection initiatives rather than program evaluations (with the exception, perhaps, of the Tennessee STAR class-size experiment). Nevertheless, the methodology for measuring educational studies' influence contributes a new approach to studying influence. The Editorial Projects in Education study constructed a three-dimensional "influence index" combining expert ratings collected through surveys with citation analyses of both the news media and scholarly literature. This influence index is useful for producing a list of the most highly influential studies; however, expert surveys cannot feasibly be conducted every time a single program evaluation's influence is to be measured.

From the above discussion it is clear that definitions of evaluation influence are still under discussion within the field of evaluation. At the same time that theoretical

agreement needs to be reached in the field, work is needed to develop and validate measures and methods for assessing an evaluation's influence. As Kane explained, "Empirical evaluations of competing theories are possible because a community can agree over definitions of relevant descriptive attributes and on acceptable measures of these attributes" (2006, p. 46). This thesis will contribute to the methodological side of the investigation by proposing descriptive attributes and validating methods and measures for gathering data on the one aspect of evaluation influence – the dissemination of generalizable knowledge derived from evaluation findings within the context of STEM education. The method being explored in this study is that of citation analysis. The following section describes the development of citation analysis methods.

Citation Analysis

Citation analysis is the best known method used within the field of bibliometrics, which is the formal study of scholarly communication (Borgman, 1990, p. 13). This section will describe the development and growth of citation indexes, different types of citation analysis methods, applications of citation analysis, and the validity of citation analysis methods for measuring research and researcher impact. Further issues concerning citation analysis methodology are presented in Chapter Four, including citation content categorization schemes, the coverage of citation databases, and the merits of existing citation index metrics.

Development and Growth of Citation Indexes

Some scholars trace the idea of citation indexing back to ancient times, arguing that it was used in the Talmud beginning in the twelfth century and in Anglo-Saxon legal texts as early as the mid-18th century (Wouters, 2000, p. 66). The modern science of citation analysis is credited to Eugene Garfield, who has been called the "undisputed patriarch of citation indexing" (Cronin & Atkins, 2000, p. 5). Garfield began work on developing citation indexes during his doctoral studies at the University of Pennsylvania in the mid-1950s. The invention of computer database technologies allowed for the systematic collection and retrieval of citation data to be feasibly used for assessing patterns in scholarship and relationships between scholars. In 1961, he founded the Institute of Scientific Information (ISI), which published in 1963 the first *Science Citation Index* (now Web of Science database).

Garfield's unique personal history clearly influenced the development of modern citation indexing. The concept and mechanisms of citation indexing was built over many years and drew upon a variety of Garfield's experiences ranging from being a child of the Great Depression, two aborted attempts at undergraduate degrees in medicine and chemistry, being discharged from the army, and working in a chemistry lab exploring the use of newly emerging punch card technologies to manage medical literature (Thackray & Brock, 2000). As Garfield himself describes in a historical review of citation indexing written in 1979, he became "personally obsessed" with the intellectual challenge of how to develop a mechanized way of indexing relationships among scientific publications (Garfield, 1979b) resulting in years of dedication to developing and refining citation indexing databases.

The resulting database, the *Web of Science*, has grown enormously since the mid-60s. As of January 2008, the *Web of Science* index contained approximately 38 million citation records from over 9,300 research journals from around the world (Thomson Scientific, 2006). Its *Social Sciences Citation Index*, one of several content-specific databases, catalogs journal citations from 1956 until the present. *Web of Science* is a subscription-based indexing tool, costing institutions approximately $100,000 per year for access (N. Herther, personal communication, August 13, 2007). A competitor to the *Web of Science* is Elsevier's *Scopus* citation index, which was developed in 2004. *Scopus*, also a subscription-based product, claims to be the largest citation database, covering over 15,000 peer-reviewed journals from more than 4,000 publishers worldwide. *Scopus* states that it indexes 2,850 journal titles in the social sciences (Elsevier, 2008).

A recently developed competitor to these pricey subscription services is the no-cost Google Scholar search engine (http://scholar.google.com). Although Google Scholar is widely available, free, and easy to use, it does not provide a public listing of what is indexed in the database. A number of empirical studies have found that it indexes a greater amount of grey literature (i.e., writings not found in academic journals, such as conference presentations, preprint or unpublished manuscripts, books, newsletters, etc.), however, Google Scholar's coverage of academic sources has been found to be less complete than its subscription competitors (Schroeder, 2007; Yang & Meho, 2006) The Google search engine is a third mechanism for uncovering references, although it is not a formal citation index. Its search function can be used to find references embedded in the grey literature as well as references posted in even less formal ways such as on organizational or personal websites. As citation indexing databases become more comprehensive and accessible, interest in methods for analyzing citation data have grown.

Citation Analysis Methods

Citation analysis is a term used to describe the general methodology of using citation counts as data for examining and evaluating scholarly impact. Moed (2005) defines citation analysis as the "construction and application of a series of indicators of the 'impact,' 'influence' or 'quality' of scholarly work, derived from citation information" (p. ix). Citation analysis consists of collecting citation counts on a specific object of interest, and then using the data to describe the relationships between the object of interest and other objects that are linked through citations. Data obtained from citation

analyses are interpreted using a number of citation indexes (described briefly below and in greater detail in Chapter Four) or are sometimes mapped visually using network analysis software to describe the relationships among variables of interest.

Citation analyses usually focus on the citations to and/or references from a specific person or subject of interest. The person or subject of interest may be a specific researcher, research team, university, or scientific field, theory, or model. There are two sub-types of citation analysis that take different approaches to collecting data: co-citation analyses, and bibliographic coupling analyses. Co-citation analysis consists of collecting information about papers (or authors) that are jointly cited by a particular publication (or author) in order to establish a degree of similarity between the two (White, 1990). For example, a co-citation analysis has been conducted on how frequently two particular scientific papers are cited together in other papers. The more frequently the two papers are cited together, the more closely related they are determined to be. Bibliographic coupling is a similar yet distinctly different approach to collecting and analyzing citation information. In bibliographic coupling studies, papers are analyzed to see which papers and/or authors cite the same papers, even though they do not cite each other. The more of the same publications two papers and/or authors cite, the more closely related they are deemed.

Applications of Citation Analysis

Citation analysis methods have been used in a variety of studies for different applications. Moed makes a distinction between two contexts of the use of citation analyses: the scholarly research context, and the policy context (2005, p. 14). In the

scholarly research context, citation analysis is used to examine relationships among variables in a theoretical framework. In the policy context, by contrast, citation indexes are used to make decisions about individuals or groups based on the results of the studies. These distinctions are sometimes referred to as citation analysis for research versus evaluative purposes.

Citation analysis for research purposes. There are several different types of scholarly studies that have been conducted using citation analysis methods. Citations have been analyzed in a wide variety of fields to answer questions about scientific inquiry, scholars, and knowledge dissemination. For example, bibliometricians have used citation data to document the development of research on a particular topic or theory (Culnan, 1986; Small & Greenlee, 1999), to examine the relatedness of authors within a field (McCain, 1999; Peters & Van Raan, 1991; Stokes & Hartley, 1989), or to study the growth of scientific cooperation internationally or inter-disciplinarily (Moed & Bruin, 1999; Rogers & Cottrill, 1999).

Citation analysis for evaluative purposes. In addition to scholarly citation analysis, there is a field of evaluative citation analysis in which citation counts are used in policy settings to measure, evaluate, and make judgments about the performance of journals, scientists, and research groups. The most simplistic analysis counts the raw number of citations to a particular scientist, research group, or journal to measure its impact on a scientific field. More sophisticated measures have been developed as well. In the evaluation of journals, the most frequently used measure is the "journal impact factor," which is a calculation of the average number of citations to a particular journal's recent articles within a journal over a particular time frame. This measure is used as an

indicator of that journal's influence on its field. Citation analysis methods have been widely used to study the influence of top journals in a field or to analyze characteristics of journal editors or contributors (see, for example, Brooks, 1999 and Zsindely, 1999).

The most controversial application of citation analysis is the use of citation counts or indexes in the evaluation of the scientific performance of individual scholars or research groups such as academic departments. The practice of using citation counts or other citation indexes to evaluate university faculty and others for decisions about promotion and tenure or awards and recognitions appears to be growing (Reed, 1995) and is perceived to be ubiquitous in the sciences (Kelly & Jennions, 2006). Some researchers argue that the use of citation information provides unbiased and comparable metrics for evaluating the impact of researchers (L. D. Brown & Gardner, 1985; Holden, Rosenberg, & Barker, 2005; Meho & Sonnenwald, 2000), and others have found that citation counts correlated highly with peer-reviewed assessments of scholarly impact (Goldberger, Maher, & Flattau, 1995; Rinia, van Leeuwen, van Vuren, & van Raan, 1998; Smith & Eysenck, 2002). Others, as discussed below, challenge the validity of using such metrics in an evaluative way.

Validity of Citation Analysis

The use of citations by researchers within the sciences and other fields such as accounting and information sciences has been well-studied since the development of citation indexes in the 1960s. Along with empirical studies using citation information as data, discussed above, there is a rich body of theoretical literature discussing the validity of using citations to make judgments about research impact. The validity argument

advanced by citation scholars has been examined from a variety of perspectives, but many of the discussions rest on the sociological argument that citations are indicators of scholarly influence because they are a norm established within scientific communities to "give credit where credit is due."

Sociologist Robert Merton is a key proponent of the sociological perspective of citations. Merton describes the paradoxical situation of the academic who establishes intellectual property rights by making his or her private contributions publicly available (1988). Merton argues that the academic system is supported by the unstated agreement that work will be provided freely assuming it is cited appropriately by those who are influenced by it in some way. His sociological perspective on citations is informed by the above argument. He argues that citations are measures of research influence as they are the agreed upon manifestation of the scientific system of knowledge dissemination. As he explains,

> This system of open publication that makes for the advancement of scientific knowledge requires normatively guided reciprocities. It can operate effectively only if the practice of making one's work communally accessible is supported by the correlative practice in which scientists who make use of that work acknowledge having done so. In effect, they thus reaffirm the property rights of the scientist to whom they are then and there indebted. (p. 621)

Merton describes that plagiarism has been acknowledged since at least the seventeenth century as a crime against intellectual property. Contemporary academic systems continue to treat plagiarism as a serious violation of misconduct resulting in often severe consequences such as failing classes or being expelled from universities.

Merton's students Cole and Cole extend his sociological perspective of citing behavior by arguing that citations are indicators of "socially defined quality" of scholarly papers and thus are indicators of scholarly influence (Cole, 1989). As they state, "recall the assumption underlying citation analysis: authors give credit where credit is due, that is, they cite their influences; by citing them, authors are said to 'reward' the colleagues whose work they use" (p. 10). They argue to be cautious in the use of citation indicators, however, as they are imperfect measures of socially constructed perspectives on research quality and influence. Therefore, they stress that citations may be used in comparative research studies, but should not be used to assess quality or influence levels of individual scientists out of the context within which that researcher works. They conclude with the caution that "citing is *not* simply giving credit where credit is due or listing influences but, instead, is a complex response in which many motives, especially the desire to persuade, play a part" (p. 12).

The perspective that citations play a role in persuasion is reflected by Gilbert (1977) when he argues that scientific papers are "tools of persuasion" (p. 115). As tools of persuasion, references contained within works are used to advance the author's argument. Citations are instrumental for persuading the scientific community that a new paper is built on solid theoretical grounds, has attended to relevant issues in the literature, and is based upon but does not duplicate other relevant studies. There are other views of the meaning of citations. For example, Davenport and Cronin discuss citations in terms of the trust that one scholar puts into another's work. When one scholar cites another, he or she is trusting the claims presented in the work. Davenport and Cronin liken citations to votes for trust with highly cited authors being those the scholarly community has

decided are trustworthy (Davenport & Cronin, 2000). While there are a number of competing conceptualizations of how citations are used, what they symbolize, and what they measure, there is no one comprehensive "theory of citation" although many have called for such (e.g., Moed, 2005; Wouters, 1999).

Empirical studies of citation validity. Empirical studies of the validity of using citations as measures of scientific influence have primarily been conducted as correlational studies in which measures of citation counts are correlated with other measures of influence such as awards and peer judgments. Cole and Cole's study of influential physicists (1967) is one of the early formative studies of this genre. Cole and Cole examined the scientific output of 120 physicists including the number of papers they published, the number of citations to their papers, the rankings of their academic departments, and their awards. Additionally they surveyed over 2,000 physicists in academic institutions to ask about influential individuals in their field. Cole and Cole found positive correlations of .33 to .67 between the number of citations to their papers and the other measures of recognition of researcher influence and quality. Other classic correlational studies in the field found similar relationships. Clark (1954) found that citation counts were the most highly correlated measure with psychologists' assessments of the most influential members of their field. Other studies include Russell (1999), Rinia et al. (1998), van Raan (2006), and Nederhof and van Raan (1993) among others.

Critics of citation analysis. While many scholars are cautiously supportive of using citation analysis methods, others within the field of informatics are more critical of citation analysis methods. Even Garfield, inventor of the modern citation indexes, has cautioned against using them without careful consideration of how to make appropriate

comparisons between journals or individual, as fields and even sub-fields can vary widely in citation conventions (E. Garfield, 1998). Seglen (1997) argued against the use of the journal impact factor because he found the measure to be skewed with a small number of highly cited articles artificially inflating a journal's impact factor, he did not see a link between the impact factor and the true scientific quality of the journal's articles, and he argued that journal impact factors vary greatly based on characteristics of the research field. Likewise Moed (2005), although generally supportive of the use of citation analysis, argued that it is more appropriate in the evaluation of individuals, groups, and institutions when it is (1) formal, (2) open, (3) scholarly founded, (4) supplemented with expert background knowledge, (5) conducted in a clear policy environment, (6) explicitly states notions of scholarly quality, and (7) used for enlightenment rather than in a formulaic application (p. 2). Similarly, Cole and Cole (1989) stress that "citations are a very good measure of the quality of scientific work for use in sociological studies of science; but because the measure is far from perfect it would be an error to reify it and use it to make individual decisions (p. 12).

Arguably the greatest criticism of citation analysis comes from a series of articles written by MacRoberts and MacRoberts (1986; 1989; 1996). In their first study (MacRoberts & MacRoberts, 1986) MacRoberts and MacRoberts examine the issue of whether what is evident as being the influences on a paper are actually cited in that paper's reference list. They examine the way citations are or are not used in a selection of seminal works, such as papers by Einstein and Mendel, among others. In a piece of their analysis, they randomly selected 15 studies on the history of genetics and examined the number of citations compared to the number of influences they assessed should have

been cited in the work (the method for determining how many references there "should" be is unspecified). They found that, on average across the 15 studies, citations only captured 30% of the total influences in the papers.

MacRoberts and MacRoberts present a number of ways in which citations do not capture influences, for example, when an author does not cite influences on background knowledge and assumptions, lifts a reference without actually reading it, or is just ignorant of the literature, among others. MacRoberts and MacRoberts expand on these problems in their subsequent publications, including such issues as citation bias, self-citation, preferring secondary sources over primary sources, and variation in citation rates across fields (1989; 1996). They challenge two basic assumptions: first, that citations are a valid indicator of the influences on the work of scientists, and second, that writers cite in order to give credit where credit is due. In contrast, they argue that the evidence suggests that scientists do not cite the majority of their influences, especially the lion's share of influences that are informal interactions between scientists. They argue that rather than turning to citations as a measure of influences on scientists, one should instead "head for the lab bench, stick close to the scientist as he works and interacts with colleagues, examine his lab notebooks, pay close attention to what he reads, and consider carefully his cultural milieu" (p. 442).

While the concerns discussed above are important, many defenders of citation analysis have picked apart the claims and argue that the method remains a valid, albeit limited, measure of scholarly influence. Additionally, the method is valuable because it can be feasibly applied in a variety of settings and the findings from studies can be verified and replicated. Garfield defends the validity of using citation counts to evaluate

research and researchers by addressing three areas: what citation counts measure, what they do not measure, and their accuracy. In terms of the first concern, Garfield is careful to caution against ascribing too high a level of precision to citation counts. As he states, "what they [citation counts] are is a very general measure of the level of contribution an individual makes in the practice of science" (Garfield, 1979a, p. 362). As such, concerns about negative citations, self-citations, or high citations received by methodological papers are overstated, in his opinion.

Second, Garfield addresses concerns that citation counts do not measure important aspects of scholarly influence, including the importance of premature discoveries, theories that are no longer cited because they have been incorporated into the general discourse on a topic ("obliteration by incorporation"), or the prestige of the journal in which a citation is contained. Garfield addresses each of these concerns and in the end concludes that citation counts are interpretive tools that call for "thoughtful and subtle judgments on the part of those who employ them" (p. 367).

Third, in terms of their accuracy, Garfield addresses concerns about the mechanics of compiling reliable data by presenting a variety of technical solutions to the issues such as the presence of multiple authors and searching for authors with similar names. He again stresses that while there is some imprecision on collecting citation data, empirical studies of correlations with citation counts and other measures of influence provide further support for their utility. He argues that on practical grounds, as the scientific enterprise grows it becomes more difficult and costly to identify areas in which the greatest contributions are being made. Citations, Garfield believes, can be a helpful tool in that endeavor.

In sum, citation analysis indexes and methods have been used widely for over fifty years despite some controversy surrounding their proper application. While the validity of citation analysis to study and evaluate the impact of scientists and scientific endeavors has been examined for some time, the use of citation analysis within the context of STEM program evaluation efforts has not yet been studied. In the next chapter, this paper will present criteria and methods used to examine the applicability of citation analysis methods to the study of STEM education evaluation influence.

CHAPTER THREE

METHODS

This chapter describes the process used to evaluate the validity of the application of citation analysis methods in the study of STEM education evaluation influence. As stated earlier, the research question guiding this study is:

To what extent is citation analysis a useful method for measuring the influence of evaluation products on the field of STEM education evaluation?

This chapter begins with a conceptual overview of validity and then describes this study's sample, data sources, analysis methods, and limitations.

Validity

Understandings of the concept of validity are debated within the social science community. As described in the fourth edition of *Educational Measurement,* the conversation about validity began in the 1950s with publication of the first editions of *Educational Measurement* and the APA, AERA, NCME *Standards for Educational and Psychological Testing* (Brennan, 2006). Both publications stressed the relationship between validity and prediction in terms of correlations of test scores with "true" criterion scores. The *Standards* outlined four types of validity that continued to be referred to in discussions of the topic: content, predictive, concurrent, and construct (American Psychological Association, 1954). Until the late 1980s, most work on validity consisted of descriptions and debates of these types of validity and methods for assessing the validity of test scores. Over time, the emphasis changed from being about "test

validity" to discussions of validation process necessary to evaluate the appropriateness of the inferences drawn from test scores to serve particular applications (Brennan, 2006).

In 1989, Messick authored the chapter on validity for the third edition of *Educational Measurement* (Messick, 1989b). In his lengthy chapter, Messick defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13). Rather than different types of validity, as previously posited, Messick emphasizes that all validity evidence is to be integrated to form an evaluative judgment about both the inferences and consequences of the use of an assessment. In subsequent publications, Messick refines his description of the evaluative process of validation, emphasizing that all validity evidence is related to construct validity and cautioning that there are two major threats to construct validity: construct underrepresentation and construct-irrelevant variance (1989a; 1994; 1995). Messick's view has become known as the unified theory of validity and has had great impact on thinking about validity over the last twenty years.

Like Messick, Kane's work on validity emphasizes the integrated nature of validity evidence, although Kane emphasizes the conceptualization of the validation process as an interpretive argument. Following on earlier work (Kane, 1992), Kane's chapter on validation for the fourth edition of *Educational Measurement* defines the validation process as "an evaluation of the extent to which proposed interpretations and uses are plausible and appropriate" (Kane, 2006, p. 17). Validity, therefore, is the "extent to which the evidence supports or refutes the proposed interpretations and uses" (p. 17). The interpretive argument presented in a validation study must be presented clearly,

coherently, and plausibly and, like scientific theories, interpretive arguments cannot be "proven." Instead, the validity arguments can only be supported by making a strong case for proposed interpretations and uses of scores, providing adequate evidence to support such claims, and ruling out possible alternative interpretations (p. 29).

The debate about validity continues. The fall 2007 edition of the *Educational Researcher* newsletter published by the American Educational Research Association was devoted to presenting a "Dialogue on Validity" in which Lissitz and Samuelsen present the most recent challenge to the unified theory of validity (Lissitz & Samuelsen, 2007). Lissitz and Samuelsen argue that the focus on construct validity, as stressed by Messick, is unhelpful in the realm of educational testing in which, they believe, content validity is the primary concern. They challenge the notion that validity lies, not in the test itself, but in its inferences and instead argue that definitions of the test, the test's development process, and psychometric theories on which the test is based determine the test's validity (p. 442). They emphasize the importance of studying the test's construction process, including related psychometric theories, in developing valid educational tests. Many dissenting opinions on Lissitz and Samuelsen's arguments were presented even within the same volume, (see, Embretsen, 2007 and Gorin, 2007) and in Kane's commentary on the article published in a subsequent volume (Kane, 2008), reinforcing support for the Messick and Kane approaches to validity.

Validation Method

This study examines the extent to which citation analysis methods are useful for measuring the influence of evaluation products on the fields of STEM education and

evaluation. This study is framed around Kane's interpretive argument approach to validity, as it is the broadest approach and one that is applicable to evaluating the validity of an empirical method such as is the case in this study. First, the sample used in the study is described. Then the study's data sources and analysis methods are presented. Finally, the validation method, including interpretive argument and assumptions, are detailed.

Sample

This study examines the extent to which STEM education evaluation influence can be measured using citation analysis methods by examining citations to products of a sample of four National Science Foundation evaluation initiatives. The sample includes the evaluations of three multi-site programs: (1) Advanced Technological Education program (ATE); (2) Collaboratives for Excellence in Teacher Preparation program (CETP); and (3) Local Systemic Change through Teacher Enhancement program (LSC), and one example of an evaluation technical assistance project: the Math Science Partnership - Research, Evaluation and Technical Assistance (MSP-RETA) project run by Utah State University. These four evaluation initiatives were originally purposefully selected for a study of involvement, use, and influence in multi-site STEM evaluations for the NSF-funded grant project, *"Beyond Evaluation Use: Determining the Effect of Project Participation on the Influence of NSF Program Evaluations."* The four evaluations were selected as they varied in terms of the level of project-staff members' involvement in the evaluation process. One of the key research questions for the *Beyond Evaluation Use* study was how influential the studies were on the fields of STEM

education and evaluation. After exploring numerous options for assessing influence, the research team decided to conduct a citation analysis as one of several ways to measure influence. This validation study is an extension of the *Beyond Evaluation Use* project.

The four evaluations[2] in this study vary not only in the forms of evaluation, types of initiatives, and process by which they were initiated. They also differ in terms of the levels of project staff involvement, timeframes, funding, purposes, data collection methods, and dissemination efforts. The following describes each of these four evaluations in more detail, while Table 1 presents a summary of how these evaluations compare along these key aspects.

Advanced Technological Education (ATE) Program and Evaluation[3]

The Advanced Technological Education program (ATE) was founded by a Congressional mandate in 1993 to improve US advanced technological education leading to an increase in the number and quality of skilled technicians in the workforce. The program primarily targeted students and teachers in two-year colleges, although secondary schools and four-year colleges were involved as collaborating institutions.

---

[2] While the MSP-RETA technical assistance project is not an evaluation per se, for the sake of convenience the term evaluation will be used throughout this paper to refer to the four evaluation initiatives comprising the sample.

[3] Information for this summary was collected from a variety of sources synthesized for two reports produced for the Beyond Evaluation Use grant: the *Advanced Technological Education (ATE) Evaluation Summary* (Johnson, 2006), and Toal and Johnson, (2007, September 6), *Beyond Evaluation Use: Determining the Effect of Project Participation on the Influence of NSF Program Evaluations Report 5: A Case Study of the Impact of the Advanced Technological Education (ATE) Program Evaluation.* University of Minnesota: College of Education and Human Development.

Since 1993, approximately 345 sites were funded for a program cost of over $350 million.

The ATE program evaluation contract was awarded to the Evaluation Center at Western Michigan University in 1999, with Arlen Gullickson as the primary investigator. NSF funded the program evaluation with grants totaling approximately $3.1 million between 1999 and 2005. The ATE evaluation was developed by evaluators at WMU with input from NSF program officers, however, little feedback from ATE projects or centers was solicited during the program planning process. The evaluation was primarily summative in nature and used mixed-methods including a quantitative annual web-based survey, multiple site visits, and targeted studies. Dissemination of evaluation findings was a focus of the evaluation in the later years of its funding. In addition to traditional technical evaluation reports, journal articles, and conference presentations, the evaluation commissioned nine issue papers and targeted studies intended to make the program evaluation findings useful to STEM educators and evaluations.

Collaboratives for Excellence in Teacher Preparation (CETP) Program and Evaluation[4]

The National Science Foundation funded the Collaboratives for Excellence in Teacher Preparation program between 1993 and 2000 for a total investment of $350 million. In total, 19 CETPs, each involving 3-15 higher education institutions and several school districts, were founded across the nation. The program sought to increase the

[4] Information for this summary was collected from a variety of sources synthesized for two reports produced for the Beyond Evaluation Use grant: the *Collaboratives for Excellence in Teacher Preparation (CETP) Evaluation Summary* (Johnson, 2006), and Greenseid and Johnson, (2007, August 8), *Beyond Evaluation Use: Determining the Effect of Project Participation on the Influence of NSF Program Evaluations Report 2: A Case Study of the Impact of the Collaboratives for Excellence in Teacher Preparation (CETP) Core Evaluation.* University of Minnesota: College of Education and Human Development.

number of well-qualified pre-K-12 teachers of mathematics and science by improving teacher preparation and increasing collaboration within and between K-12 and higher education institutions.

The CETP program evaluation (known as the "CETP core evaluation") was developed several years after the CETP program was founded. It arose out of national discussions between the CETP project PIs about the need for comparable data across project sites and the challenges of collecting comparable data. Unlike the ATE evaluation, which was planned primarily by the program evaluators and NSF personnel, the CETP evaluation involved a significant number of CETP projects in the development and refinement of instruments and development of data collection methods. The evaluation was conducted between 1999 and 2004 for a total cost of $990,000. Dissemination of evaluation findings was less of a focus of the CETP evaluation than it was for the ATE evaluation. The involvement of project evaluators and PIs in collecting evaluation data was a greater concern.

Local Systemic Change through Teacher Enhancement (LSC) Core Evaluation[5]

The Local Systemic Change program existed between 1995 and 2005 as an effort to improve K-12 mathematics and science teachers' content and pedagogical knowledge through professional development delivered through building- or district-wide initiatives.

---

[5] Information for this summary is from the case study produced for the Beyond Evaluation Use grant: Johnson and Greenseid, (2007, October), "A Case Study of the Impact of the Local Systemic Change through Teacher Enhancement (LSC) Core Evaluation (Beyond Evaluation Use Project Report 3)." Beyond Evaluation Use Project. University of Minnesota. College of Education and Human Development, Minneapolis, MN.

Over the decade the $250 million program funded 88 projects across 31 states reaching approximately 70,000 teachers and over two million students.

The LSC projects were informed in the program solicitation materials that they would be required to participate in a national, cross-project evaluation. NSF contracted with Horizon Research soon after the program's establishment to conduct the national "core evaluation" for the program. This was the first attempt by NSF to conduct an overarching national program evaluation; the CETP core evaluation was substantially informed by this experience. The LSC core evaluation designed standardized data collection instruments to collect comparable data across the sites and report outcomes to NSF with input from LSC projects about the design of the evaluation and the classroom and professional development observation protocols. At the end of the initial contract period, NSF project officers requested that the core evaluation use their two-year no-cost extension period to disseminate lessons learned to the STEM education and evaluation fields.

Math Science Partnership – Research, Evaluation, and Technical Assistance (MSP-RETA) Evaluation Capacity Building Efforts[6]

The NSF Math and Science Partnership program is a research and development initiative authorized by the US Congress in 2002 to improve K-12 math and science instruction, improve student achievement, and address the achievement gap. Since 2002, 80 awards totaling over $500 million have been granted to MSP partnership projects that

[6] Information for this summary is from the case study produced for the Beyond Evaluation Use grant: Volkov and Johnson, (2007, September 19), *Report 4: A Case Study of the Impact of the Utah State Math Science Partnership (MSP) – Research, Evaluation, and Technical Assistance (RETA) Project.* University of Minnesota: College of Education and Human Development.

bring together institutions of higher education, K-12 school districts, and local business partners. In addition to partnership projects, Research, Evaluation, and Technical Assistance (RETA) projects were funded to build the research and evaluation capacity of the projects.

The Utah State RETA was funded in 2002 with a grant of $1.8 million to provide such technical assistance and build the evaluation capacity of the MSPs. The RETA developed a network of evaluation experts to provide one-on-one technical support and consulting to MSP projects who wanted such assistance. The project also organized annual national conferences with an evaluation focus for MSP PIs and local evaluators. A result of one such conference was the development of the Design-Implementation-Outcomes (DIO) Cycle model, which provided a framework for MSP projects to use when conducting their evaluations. The Utah State RETA is different from the three evaluations described previously in its focus on evaluation capacity building as opposed to a coordinated national program evaluation effort. This shift in emphasis reflected NSF's interest in building the capacity of local projects to conduct rigorous evaluation, with a de-emphasis on obtaining comparable data across the projects.

Table 1. Description of the four NSF evaluation initiatives

Evaluation	Evaluation Timeline/ Budget	Purpose	Data Collection Methods	Dissemination Methods
ATE program evaluation Western Michigan University PI: Arlen Gullickson	1999-2006 $3.1 million	Primarily summative evaluation; provide evidence of quality of ATE program; inform program improvement	Mixed methods: • Online survey • Site visits	Dissemination to fields prioritized by NSF • Issue papers; • Targeted studies; • Over 50 reports and publications; • Brochures intended for ATE project audiences
CETP core evaluation University of Minnesota PI: Frances Lawrenz	1999-2004 $990,000	Both formative and summative evaluation; extent of improvement in teacher preparation	Mixed methods: • Surveys of deans/department chairs, principal investigators/evaluators, faculty, college students, students in grades 6-12, teachers, and principals • Classroom observations of teaching practices	NSF primary audience for reporting efforts • Evaluation reports and technical reports • Instrument handbooks and training manuals
LSC core evaluation Horizon Research, Inc. PI: Iris Weiss	1995-2005 $6.5 million	Both formative and summative evaluation; extent of system-wide implementation and improvement in teacher content and pedagogical content knowledge	Mixed methods: • Observations of professional development sessions; • Classroom observations; • Interviews with teachers and project staff; • Questionnaires administered to teachers and principals	Dissemination to field focus of last 2-years of funding: • Emphasis on dissemination to field through journal publications, presentations, and training in protocols

MSP-RETA project Utah State University PI: Cathy Callow-Heusser	2002 – present $1.8 million	Evaluation Capacity Building; provide evaluation technical assistance to strengthen evaluations of local MSPs	N/A; data are collected locally. The Utah State RETA developed of DIO Cycle of Evidence model for evaluation to provide an evaluation model for local efforts.	Transfer of knowledge through direct technical assistance provided to the local projects and through hosting national evaluation conferences for MSPs

Comparison of the Four Sampled Program Evaluations to other NSF Evaluations

The program monitoring-type model used in the ATE evaluation, the core evaluation models of the LSC and CETP evaluations, and the evaluation capacity building model in the MSP-RETA initiative reflect three distinct approaches to national NSF STEM evaluation initiatives. Although similar in that they are multi-year, large-scale evaluation projects, the evaluation designs, methods, and dissemination strategies reflect a variety of best practices currently in use in multi-site evaluations. In terms of their representativeness to other NSF program evaluations, the ATE, CETP, LSC, and MSP-RETA evaluations represent a wide range of multi-site program evaluation efforts supported by the National Science Foundation over the past 15 years.

An article in the *New Directions for Evaluation* volume on "Critical issues in STEM evaluation" presents a history and overview of NSF STEM program evaluations (Katzenmeyer & Lawrenz, 2006). First, Katzenmeyer and Lawrenz distinguish among three major types of evaluations: (1) "status" studies in which national or international surveys of STEM education are conducted, (2) program evaluations in which all the projects related to a particular funding initiative are evaluated, and (3) project evaluations

which are site-specific, local evaluations. Three of the four evaluations represented in this study would be considered traditional program evaluations, while the fourth, the MSP-RETA project, represents a program-level evaluation initiative that does not actually evaluate all of the projects, but instead is intended to provide technical assistance to them.

Second, Katzenmeyer and Lawrenz outlined different types of initiatives that have been the focus of NSF funding efforts since the 1990s. These included systemic initiatives, monitoring systems, partnerships, and technical assistance programs. The four program evaluations in this sample encompass all of these areas to varying degrees and at various times. The LSC and ATE evaluations began primarily as monitoring systems, although they became more evaluative in nature over time (p. 14) with LSC focusing on systemic initiatives and ATE addressing many issues including partnerships. The CETP evaluation was primarily an evaluation of systemic initiatives and partnerships, and the MSP-RETA project represents an attempt at building evaluation capacity of the MSP project evaluations through technical assistance.

Third, Katzenmeyer and Lawrenz discuss reasons NSF initiates program evaluations. They examine both who initiated the evaluation and what was the intended purpose of the evaluation. The four evaluations in this sample range in terms of how they were initiated. The ATE evaluation was funded six years after the ATE program had been Congressionally mandated and was initially conceived as a program-level monitoring effort. The CETP core evaluation, in contrast, was funded at the request of the PIs and evaluators from the local CETP projects in an effort to have comparable findings. In contrast to both of the ATE and CETP evaluations, the LSC core evaluation

was commissioned at the same time the national initiative was funded and was NSF's first attempt at establishing a core program evaluation. The MSP-RETA project was conceived as a technical assistance project to build the capacity of the local MSPs. Three years later, large grants were made to conduct a separate monitoring study and a program-level evaluation.

While it is not possible to quantify the extent to which the four evaluation initiatives in this sample represent all evaluations funded by NSF, the discussion above suggests that the evaluations represent a range of evaluations of a certain scale commissioned by the agency. Other examples of program evaluations of this level include the program evaluations for the Centers for Learning and Teaching, Systemic Initiative, the Mathematics and Science Partnerships, Program for Women and Girls, Preparing Future Faculty, Graduate Fellows in K-12 Classrooms, and Integrated Graduate Education and Research Traineeships programs, among others. One experienced NSF evaluator assessed that over years there have been approximately 50 program evaluations of the types that the four included in this sample represent (F. Lawrenz, personal communication, October 30, 2007).

Data Sources and Analysis Methods

This study uses data from four sources to inform an overall evaluative judgment about the extent to which citation analysis methods are valid for measuring evaluation influence: (1) citation analysis data, (2) the opinions of an expert panel, (3) data from a survey of primary investigators and evaluators from the local projects connected with the four national program evaluations, and (4) a review of relevant literatures. The

University of Minnesota Institutional Review Board stated that the data from the citation counts, expert panel, and literature reviews did not meet the definitions of data from human subjects and were thus not necessary to be reviewed (C. McGill, personal communication, January 17, 2008). Use of existing survey data in this study was exempt from full IRB review under category 4, under study # 0801E25484.

Citation Analysis Process

In this study, Web of Science, Google Scholar, and Google were used to collect citation information from lists of evaluation products produced by the four program evaluations of interest. The lists of program evaluation products were compiled by reviewing the principal investigators' curricula vitae, program evaluation archives, and through Internet web searches. Products were defined as any publications, presentations, instruments, or other related materials that were produced as part of the NSF-funded evaluation project. Additional publications that were substantially informed by the principal investigator's experiences conducting the program evaluation were also included; the principal investigators themselves made judgments about which of these additional products should be included and were provided with the opportunity to review the final lists to ensure their completeness and accuracy. For more details and a full list of the evaluation products, please see Greenseid, Johnson, and Lawrenz (2008).

Searching was completed between the following dates for each evaluation:

Table 2. Citation search dates

Evaluation	Web of Science Searching	Google Scholar Searching	Google Searching
ATE	August 27-30, 2007	September 4-6, 2007	September 6 and 10, 2007
CETP	August 16 and 18, 2007	August 19-23, 2007	December 3-4, 2007
LSC	October 23-25, 2007	October 25-November 13, 2007	November 13-December 2, 2007
MSP-RETA	September 25, 2007	September 25-26, 2007	September 26, 2007

Data were collected manually by entering full reference information about the citations into an Excel spreadsheet that was later checked for accuracy and compliance with APA style and converted into an SPSS datafile.

Web of Science search process. Access to the ISI Web of Science was granted through the University of Minnesota library system. The database was searched using the cited reference search function, which searches for matches in the reference lists of all the journal articles indexed in the database. The database was searched for the names of all of the authors of the evaluation products within time periods appropriate to each specific program evaluation. Additionally, searches were conducted for works authored by the National Science Foundation, Advanced Technological Education, Collaboratives for Excellence in Teacher Preparation, Evaluation Center, Horizon Research, Local Systemic Change, Math Science Partnership, Utah State University, and related variants.

The Web of Science search returns results in abbreviated form and consequently often did not contain sufficient information to determine whether a particular reference was actually the product of interest. The database might report that there was a citation

to an article by “F Lawrenz” published in 2004, but the title of the work was abbreviated as “sci ed eval” or something similarly too vague to ascertain whether the work was a CETP product or another article by Lawrenz. In these cases, the researcher obtained and reviewed reference lists from the original sources (e.g., original journal articles) to ensure accuracy in collecting citations.

Google Scholar search process. The Google Scholar searches were conducted on the same lists of author names as were used in the Web of Science search, as well as by searching for exact product titles and likely variants. Combinations of author names and titles were also used to refine searches as necessary. As with the Web of Science searches, it was necessary for the researcher to obtain access to original reference lists to verify the accuracy of citations in many cases. Additionally, the researcher had to conduct additional research to verify the reference information of the citing works as it was found that Google Scholar’s reporting of title, author names, and sources was sometimes inaccurate.

Google Search process. In many cases, author name searches were unmanageable using the Google search engine. For example, searching for “Iris Weiss” (in quotations) returned 14,500 results. Consequently, while author name searches were attempted, title searching was used as the primary method for culling through the numerous possible results for actual citations. As mentioned above, much time was needed to verify whether the returned results were true citations to the evaluation products of interest. Again, original reference lists were consulted when available to verify whether the returned results were actual citations as well as to collect the necessary data to write a full reference of the citing work.

Expert Panel Survey

Additional validity evidence was gathered from an expert panel consisting of three evaluation theorists and four evaluation practitioners. The evaluation theorists are considered to be among the foremost experts in the topic of evaluation use and influence and, consequently, are highly respected for their understandings of the construct of evaluation influence. The evaluation practitioners are the primary investigators of the four sample evaluations used in the citation analyses. They each have an insider's understanding of the program evaluation's impact that will allow for judgments to be made as to whether the data from the citation analysis fits their impressions of the evaluation's influence.

Description of expert panelists

Evaluation Theorists

- **Karen Kirkhart** is Professor of Social Work at Syracuse University. Her research interests focus on social work research; clinical practice evaluation and program evaluation, and multicultural validity. She has published an evaluation influence model and maintains an interest in meta-evaluation and program evaluation standards. Kirkhart is the 2007 winner of the American Evaluation Association's Robert Ingle Service Award and Paul F. Lazarsfeld Award for Evaluation Theory.

- **Michael Patton** is an independent organizational development consultant and has written five major books on the art and science of program evaluation, including the influential "Utilization-Focused Evaluation." He is also the lead author of an influential early study of evaluation use. Patton has made a significant contribution to the ever-developing methodology of evaluation, facilitating clarity about interventions and their potential impact using logic modeling, systems thinking, complexity theory, and concept mapping. He is the recipient of the American Evaluation Association's Lazarsfeld Award for Evaluation Theory and Myrdal award for outstanding evaluation practice.

- **Marvin Alkin** is Professor and Chair of the Social Research Methods Division in the Graduate School of Education and Information Studies at the University of California – Los Angeles. He has written extensively on evaluation utilization and on comparative evaluation theory. Alkin has been a consultant to six national governments and has conducted more than 75 evaluations of a variety of educational, governmental, and foundation programs. Alkin is a winner of the American Evaluation Association's Lazarsfeld Award for Evaluation Theory.

NSF Evaluation Primary Investigators

- **Catherine Callow-Heusser** is the director and owner of EndVision Research & Evaluation, with which she directs projects including the external program evaluation of the Bureau of Indian Affairs (BIA) Reading First Grant, DIBELS assessment of K-3 students enrolled in BIA Reading First schools. Catherine is currently the Principal Investigator of the NSF Math Science Partnership Research, Evaluation, and Technical Assistance project (MSP-RETA) at Utah State University to build evaluation capacity and provide technical assistance to MSP projects. She formerly directed Utah State's Early Head Start Research project as well as numerous other research, evaluation, and development projects.

- **Arlen Gullickson** has been affiliated with Western Michigan University's (WMU) Evaluation Center since 1991, most recently serving as its director. He also is a professor of counselor education. Prior to coming to WMU, Gullickson had been a faculty member at the University of South Dakota, and he served as coordinator of the South Dakota Rural Science and Math School without Walls Project. While at WMU, he has directed a number of major evaluation research projects, including the NSF's Advanced Technological Education (ATE) program. Gullickson is the 2007 winner of the American Evaluation Association's Alva and Gunnar Myrdal award for outstanding evaluation practice.

- **Frances Lawrenz** is Associate Vice-President for Research and Professor of Educational Psychology at the University of Minnesota. Her major research focus is science and mathematics program evaluation. She has been the Primary Investigator for numerous NSF program evaluations, including the Collaboratives for Excellence in Teacher Preparation program and has served on the assessment working group for the National Standards of Science Education. Lawrenz is also a winner of the American Evaluation Association's Alva and Gunnar Myrdal award for outstanding evaluation practice.

- **Iris Weiss** is President of Horizon Research, Inc. (HRI), a contract research firm in Chapel Hill, NC specializing in science and mathematics education research and evaluation. Dr. Weiss was the Principal Investigator of the NSF Program Local Systemic Change through Teacher Enhancement (LSC) program. She has also provided consultation to the NSF, the US Department of Education, the

> National Science Teachers Association, the Council of Chief State School Officers, and many others. She participated in the evaluation of NSF's model middle school mathematics and science teacher preparation and Triad curriculum programs. She has also served on the assessment working group for the National Standards of Science Education.

The four evaluation PIs were sent copies of citation analysis reports prepared for their evaluations and were asked to comment on the accuracy and representativeness of the report and their impressions of the usefulness of citation analysis for understanding the influence of their evaluations on the STEM education evaluation field. The three evaluation theorists were provided with a copy of the overall citation analysis report that compared citation patterns across the four program evaluations. The reports provided some overview of the citation analysis methodology employed in the study, however, they were not as comprehensive as the data presented in this dissertation. For instance, the reports did not present the data from the findings about the content of the citations.

Evaluation Project Survey Data

A third source for gathering validity evidence was a survey conducted with project evaluation primary investigators and evaluators from the four multi-site evaluations mentioned above (ATE, CETP, LSC, and MSP-RETA) for the Beyond Evaluation Use project. As shown in Table 3, a total of 369 individuals responded to the survey for an overall response rate of 46%.

Table 3. Respondents to the Beyond Evaluation Use Project Survey

	ATE	CETP	LSC	MSP-RETA	Total Sample
Evaluators	11	19	30	3	63
Non-Evaluators (PIs, project staff, project-related consultants, etc.)	175	35	43	53	306
Total	186	54	73	56	369

The representatives were asked about their levels of involvement in and use of the four program evaluations. Additionally, they were asked to judge the influence of the evaluations on the STEM education and evaluation communities. These three questions from the survey regarding the use and influence of the evaluations by the fields are used as data in this validation study.

Caution should be used when interpreting the data from the MSP-RETA survey respondents. The MSP-RETA survey respondents indicated confusion regarding exactly which evaluation they were supposed to be considering when responding to the survey questions. Additionally, follow-up interviews with some respondents also indicated that some survey respondents were confused and may have been thinking about different evaluations connected with the Math Science Partnership program, rather than the evaluation project conducted by Utah State University.

Literature Review

The final source of data for the study is a comprehensive literature review regarding validity issues surrounding citation analysis. Citation analysis has been used widely as a way of measuring the impact of scientific research for the last 40 years and

the literature is rich with discussions of its validity, in particular with reference to using citation analysis for making high-stakes decisions such as tenure and promotion of academic faculty.

Validation Method

Using Kane's approach, the following interpretive argument was developed to guide the validation process:

> **Argument:** *Citations are interpreted as indicators of the impact of STEM education evaluation products, and, as such, citation analysis has utility as a method for measuring the influence of program evaluations on the STEM education and evaluation fields.*

Nine assumptions follow from this argument. Each is listed below, along with the data sources and analyses that were used to evaluate the assumption. Chapter Four presents the findings from the evaluations of each of these assumptions.

***Assumption 1**: Citation analysis is an established method for measuring the impact of STEM education evaluations or research efforts in related fields.*

This assumption examines the extent to which citation analysis is currently being used in STEM education evaluation contexts and related fields. While the existence of the use of citation analysis is not sufficient for establishing the validity of the method for drawing inferences about evaluation influence, evidence of its acceptance within STEM education evaluation and related fields provides some support for using the method. The validity of this assumption was evaluated with findings from the literature review and

with the judgment of the evaluation theorists. First, literature within the fields of STEM, education, evaluation, and the intersection of the three fields was reviewed with attention to the different levels of analysis at which citation analyses have been conducted. Second, the three evaluation experts were asked to make a judgment about the extent to which references are a convention used within STEM education and evaluation papers to give credit to influential ideas or methods. The basic use of references within the context of STEM education evaluation must first be established before examining how they are used, what they mean, and what inferences can be drawn from them.

***Assumption 2**: The content of citations to STEM education evaluation products suggests they are used to give credit where credit is due or represent other indicators of influence.*

Following from assumption one, this assumption states that not only do authors cite each others' work, but they do so in meaningful ways that indicate the work is relevant or influential to their own. This assumption was examined through a review of relevant literature and a content analysis of the citations gathered from the four program evaluations comprising this study's sample. The literature review focused on empirical studies of the content of citations to scholarly research publications, gathering evidence on the extent to which citations are relevant and/or an indicator of influence versus being irrelevant (e.g., non-related). Next, a content analysis of a random selection of 30 of the collected citations was conducted, coding the citations based on a taxonomy derived from the literature review and informed by theories of evaluation use and influence.

Assumption 3*: Citation databases exist that provide adequate coverage of the STEM education and evaluation fields.*

This assumption is a question about the coverage of existing citation analysis search engines and whether they index sources that would likely contain references to evaluation-related works. Specifically, articles in peer-reviewed journals are only one of many types of products produced by evaluations; others include evaluation reports, books and monographs, conference presentations, and evaluation instruments. A search engine that only captures references to peer-reviewed articles would clearly not provide adequate coverage of citations to evaluation-related products. This assumption was evaluated by first comparing the databases' coverage with lists of prominent STEM education and evaluation journals. Then the publication types actually captured in the citation searching (for Google Scholar and Google) were analyzed for representation of differing types of publications. Particular consideration was given to concerns over the under-representation of practitioner-oriented sources as opposed to more traditional scholarly sources.

Assumption 4*: The process of gathering STEM education evaluation product citation data can be conducted accurately.*

The evaluation of this assumption asks whether the data collection methodology used in this study results in data that are reliable across repeated collection attempts and different data collectors. The primary researcher and one outside researcher, trained to use the same process as the primary researcher, attempted to replicate the citation results of a random selection of 25 evaluation products. The results were analyzed in terms of

the total number of citations found and the intra-rater and inter-rater percent agreement among the three data collection trials, as well as discussing possible factors affecting the differences among the trials.

Assumption 5*: Citation data can be transformed into meaningful indexes for comparing levels of STEM education evaluation product influence.*

Evaluating this assumption provides evidence regarding procedures for drawing inferences about evaluation product impact from citation counts. First, a review of the literature on existing citation indexes was explored. Then a comparison of six relevant citation indexes was conducted with data from the four sample program evaluations.

Assumption 6*: Citation indexes are related to other measures of STEM education evaluation influence.*

The first five assumptions were concerned with development and measurement of the descriptive attributes of evaluation product impact. This sixth assumption relates to the descriptive attributes of the theoretical construct of evaluation influence. Evaluation product citation indexes were compared with the rankings of other measures of influence (convergent patterns). Unfortunately, there is no established measure of evaluation influence. The first source of evidence for convergent patterns of relationships, therefore, comes from literature on the relationships between citation indexes and other measures of research influence/impact. Second, data from a survey of project-level representatives from the four multi-site evaluations used in this study provides a measure of influence. In the survey, project-level representatives were asked how influential the evaluations

were on the STEM education and evaluation communities and how they used the evaluation instruments in subsequent evaluations. Rankings of the evaluations' influence levels will be compared with citation indexes.

Assumption 7*: Citation analysis is useful for understanding differences in patterns of use and influence within and across STEM education program evaluations.*

This assumption evaluates the usefulness of citation measures to understanding influence patterns within and across STEM education evaluation contexts. First, the study used multiple regression to examine the relationship between and among four variables of interest (program evaluation, product type, product field, and product content area) and influence (mean citations per product). Second, a visual analysis of the data was conducted on the citation network for the four evaluations together as well as each evaluation's individual citation network. Finally, the expert panel was asked to comment on the results of the analyses to see if the results confirmed their expectations according to their personal knowledge of their program evaluations and dissemination efforts.

Assumption 8*: Citation analysis is useful for understanding the influence of STEM program evaluations of different sizes.*

As the previous assumptions were all examined with data from four multi-site STEM education program evaluations, this analysis assesses the transferability of the developed indexes to STEM evaluations of different sizes. The first evaluation examined represents that of a small-scale, single-site program evaluation (the Wisconsin Academy Staff Development Initiatives' Retention and Renewal program evaluation) and

the second represents a large-scale, international status study (Trends in International Mathematics and Science study). It was expected that the TIMSS study would have high levels of influence, as it was ranked among the top ten most influential studies in a recent study (Swanson & Barlage, 2006), the multi-site evaluations having a moderate level of impact, and the single-site evaluation receiving few or no citations.

Assumption 9: *The consequences of using evaluation product impact indicators as measures of STEM evaluation product influence are more beneficial than detrimental.*

This assumption explores the possible positive and negative consequences of using citation analysis results as measures of evaluation influence. Data to examine this assumption consisted of the judgments of the evaluation experts as to the likely consequences of using citation counts within research endeavors on STEM education evaluation influence and their overall assessments of the validity and utility of using citation analysis to measure evaluation influence.

In sum, nine assumptions are examined to provide evidence to the extent of validity of using citation analysis methods for assessing STEM education evaluation influence. Table 4, below, presents an overview of the assumptions, data sources, and analyses conducted in the study.

Table 4. Assumptions, data sources, and analyses

Research question: To what extent are citation analysis methods useful for measuring the influence of evaluation products on the fields of STEM education and evaluation?		
Assumptions	**Data source**	**Analyses**
Assumption 1: Citation analysis is an established method for measuring the impact of STEM education evaluations or research efforts in related fields.	Literature Review	Review of literature regarding use of citation analysis within STEM, education, and evaluation fields
	Survey of evaluation theorists	Analysis of expert panel's judgment about use of references as ways of giving credit within the field of STEM education evaluation
Assumption 2: The content of citations to STEM education evaluation products suggests they are used to give credit where credit is due or represent other indicators of influence.	Literature Review	Review of empirical literature regarding the content of citations
	Content analysis of citations from CETP, ATE, LSC, and MSP-RETA program evaluations	Random selection of 30 citations; content coding of citations as used within the original works
Assumption 3: Citation databases exist that provide adequate coverage of the STEM education and evaluation fields.	Search of coverage of key journals; Citation analysis results from the CETP, ATE, LSC, and MSP-RETA program evaluations	Comparison of the coverage of evaluation literature (journals, reports, presentation sources) of three citation analysis search engines: ISI Web of Science, Google Scholar, and Google
Assumption 4: The process of gathering STEM education evaluation product citation data can be conducted accurately.	Citation analysis results from CETP, ATE, LSC, and MSP-RETA program evaluations.	Random selection of 25 citations and repeated data searching by both the primary researcher and an outsider. Percentages of intra-rater and inter-rater agreement and overall citation counts are reported.
Assumption 5: Citation data can be transformed into meaningful indexes	Literature review	Description of existing citation indexes used to measure research impact (e.g., ways of calculating citation impact scores)

for comparing levels of STEM education evaluation product influence.	Citation analysis results from the CETP, ATE, LSC, and MSP-RETA program evaluations	Examination of the usefulness of existing citation indexes for measuring evaluation influence
Assumption 6: Citation indexes are related to other measures of STEM education evaluation influence.	Literature review	Review of empirical studies regarding the correlation between citation counts and other measures of research influence or impact
	Citation indexes from the CETP, ATE, LSC, and MSP-RETA program evaluations and survey responses from local evaluation projects' PIs and evaluators	Agreement between rankings of program evaluation citation indexes and evaluation project leader and evaluator survey respondents
Assumption 7: Citation analysis is useful for understanding differences in patterns of use and influence within and across STEM education program evaluations.	Citation analysis results from the CETP, ATE, LSC, and MSP-RETA multi-site program evaluations.	Comparisons of mean citations per product for: • evaluations with different dissemination purposes • product types • product fields • product content areas
	Survey of primary investigators	Analysis of primary investigators reflections on the patterns of citations found for their products
Assumption 8: Citation analysis is useful for understanding the influence of STEM program evaluations of different sizes.	Citation counts from the WASDI R2 single-site program evaluation, and the TIMSS large-scale status study, compared to four evaluations in sample.	Comparison of citation levels for single-site, multi-site, and status study evaluations
Assumption 9: The consequences of using citation analysis to measure STEM evaluation product influence are more beneficial than detrimental.	Survey of evaluation theorists and primary investigators	Analysis of evaluation theorists' beliefs as to the possible positive and negative consequences, validity, and utility of using citation analysis to measure evaluation influence

Limitations

There are several limitations inherent in this study. First, the primary data used in testing the majority of the assumptions presented above come from a sample of just four large-scale NSF STEM evaluations. The extent to which these four evaluations are similar to STEM evaluations conducted outside of the NSF context or of different sizes is unknown. While some evidence will be considered comparing the citation patterns of these four evaluations to two evaluations differing in scale, clearly a more thorough analysis of a variety of contexts should be conducted before strong extrapolations can be made to other STEM evaluation contexts and to other educational or evaluation contexts.

Second, it is important to reemphasize that this study only examines one small type of influence, that arising from the dissemination of findings, and that the measure examined here (citations) captures only a small portion of the construct. Valid indicators of a wide variety of possible evaluation influences need to be developed. Moreover, once additional indicators are advanced, the collection of further convergent correlation evidence will help to further validate this study. Additionally, this is only one possible way of measuring the influence of knowledge generation and dissemination efforts. Some other possible ways of tracing influence would be to follow-up with recipients who were mailed evaluation reports to survey them about their use or to track downloads of reports and instruments from program evaluation websites, among others.

Third, the argument-approach to validation process does not result in "yes/no" or "valid/invalid" conclusions to be drawn about the posited interpretive argument. Rather, the process attempts to define the boundaries within which valid inferences can be drawn. While not a limitation per se, the importance of the careful and clear presentation of the

findings from this study must be stressed so that invalid extensions of the interpretations do not occur. Despite the aforementioned limitations, this study should contribute to the development of methods useful for assessing aspects of the construct of evaluation influence, thereby furthering theoretical understandings in the field and enabling future research efforts on the topic.

CHAPTER FOUR

FINDINGS

This chapter presents the findings associated with the nine assumptions used to evaluate the interpretive argument framing this study. This interpretive argument is:

> **Argument:** *Citations are interpreted as indicators of the impact of STEM education evaluation products, and, as such, citation analysis has utility as a method for measuring the influence of program evaluations on the STEM education and evaluation fields.*

The empirical findings evaluating each of the assumptions are discussed below. Chapter Five weighs the strengths and weaknesses of the validity evidence presented here to make a final judgment about the validity of citation analysis methods for measuring the influence of STEM education evaluations.

Assumption 1: Citation analysis is an established method for measuring the impact of STEM education evaluations or research efforts in related fields.

The first step in evaluating the extent to which citation analysis is a useful method for assessing the influence of STEM education evaluation products is to examine the existing literature related to the use of citation analysis within STEM education evaluation and related fields (see Figure 6). There are two levels at which this assumption is evaluated. First, literature is reviewed to assess the extent to which citation analysis has been used in the fields of STEM, education, evaluation and the conjunction of these three fields. Second, the literature is reviewed with attention to the levels of analysis at which citation analysis has been applied, namely at the levels of individuals,

research centers, and program evaluations. Following the literature review, the panel of expert evaluation theorists' assessment of the use of references and citation analysis within STEM education evaluation is presented.

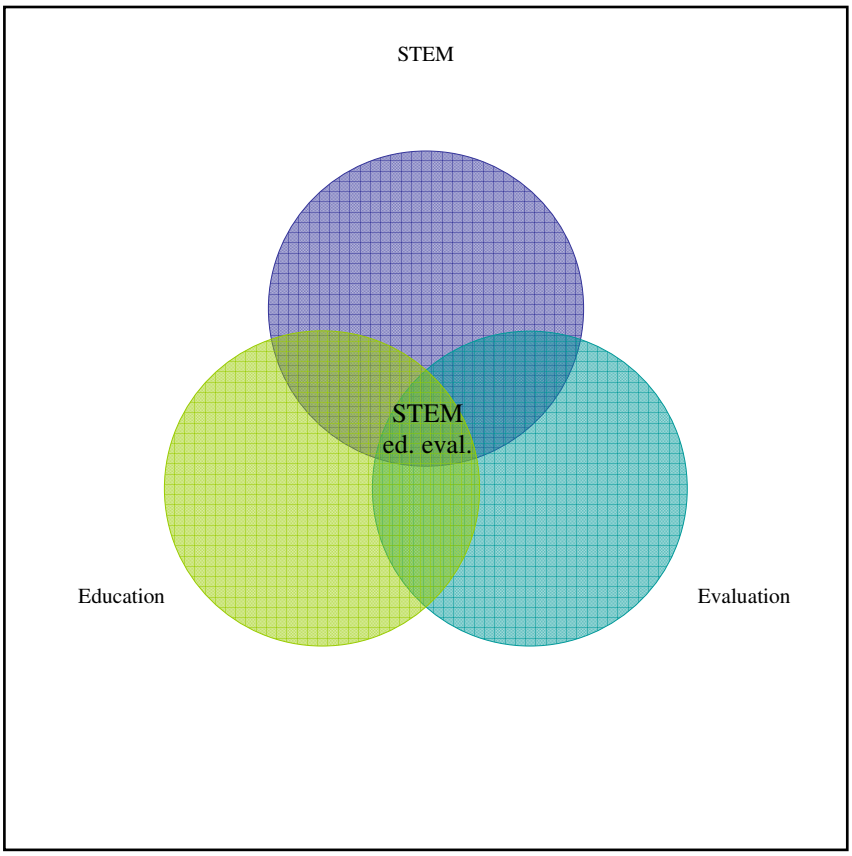

Figure 6. Relationship among fields of STEM, education, evaluation, and STEM education evaluation

Use of citation analysis within STEM fields. Citation analysis has a long history of usage within the sciences. In fact, as presented in Chapter Two, Garfield's earliest work on developing a citation index was to trace relationships within the field of chemistry. The *Science Citation Index* developed by Garfield's Institute for Scientific Information in 1963 today provides citation indexing on a wide variety of scientific disciplines included in over 9,000 scientific journals. Additionally, the *Web of Science*

includes specialized databases for specific STEM areas such as the life sciences (Biological Abstracts and BIOSIS Previews), agriculture and environmental sciences (CAB Abstracts), food science, food technology, and nutrition (Food Science and Technology Abstracts), physics, electrical/electronic technology, computing, control engineering, and information technology (Inspec), biomedicine, bioengineering, public health, clinical care, and plant and animal science (MEDLINE), and animal biology (Zoological Record) (Thomson Scientific, nd). Numerous citation studies have been conducted from these data over the past forty years.

In terms of the levels of analysis, citation data have been used widely to conduct research within the sciences (examples provided previously), as well as extensively in the evaluation of individual scientists, groups of scientists, and scientific research centers. In addition to many studies published in journals, a book was recently published that specifically addresses the use of citation analysis in the evaluation of research (Moed, 2005). As discussed previously, the use of citation data for evaluative purposes has been the most controversial application of citation analysis within the sciences, with calls for abandoning the use of citations for evaluative purposes, in particular in making decisions about promotion and tenure of academic faculty (Seglen, 1997; Walter, Bloch, Hunt, & Fisher, 2003).

Use of citation analysis within education in general and STEM education specifically. While the use of citation analysis is widespread within STEM areas, the application of citation analysis within educational fields is not nearly as ubiquitous. To find literature on the use of citation analysis within education, the *Education Resources Information Center (ERIC)* database was searched using the keyword “citation analysis.”

This keyword search found 519 matches. Examination of the first 100 of the 519 titles and abstracts documented that the vast majority of publications addressed the use of citation analysis within fields other than education, for example, in library and information sciences, psychology, management, or economics. In fact, only six of the first 100 references appeared related to citation analysis within education-related fields. Therefore the search process was refined using the descriptor "citation analysis" and the keyword "education." This combination identified 154 possibly relevant studies. The titles and abstracts of these studies were examined, and it was found that many of the studies were coded in ERIC as "higher education" but were really related to academic disciplines other than education. In the end, 44 studies were selected as related to citation analysis within the field of education.

These 44 studies examined a wide-range of educational topics using citation analysis methods. Of the 44 studies, ten studies addressed issues in higher education, the educational area in which citation analysis has been applied to the greatest extent. Among the remaining 34 studies, three studies each were related to gifted education, instructional systems and design, and agricultural studies. Two studies were conducted in each of the following areas: educational technology, learning styles, adult education, reading, medical education, and special education. One study was conducted on each of the following topics: curriculum development, doctoral scholarship, bilingual education, school psychology, educational administration, music education, education of the hearing impaired, education sciences, elementary and early childhood education, teacher evaluation, moral education, vocational education, and science education.

Note that as stated above, only one study was discovered that used citation analysis methods within any STEM education area (Tamir, 1996). Tamir, in a short guest editorial paper in the *Journal of Research in Science Teaching,* states that "the Citation Index [capitalization in original] is an important source often used to evaluate the scholarly impact of academicians" (p. 690). Tamir stated that he was interested in examining influential science education researchers and consequently decided that a citation analysis comparing the lists of authors cited in two versions of the *Handbook of Research on Science Teaching* (Gabel, 1994; Shulman & Tamir, 1973) would provide him with a picture of trends in the field of science education. Tamir's particular conclusions about trends in science education researchers are not important to this study, however, his decision to use citation analysis as a method of assessing the relative influence of science education researchers provides one example of a belief in the merit of the method for discussing questions of scholarly influence within the field.

Most of the educational studies employing citation analysis used citations to examine patterns of scholarship within sets of journals or scholarly papers. Half of the 44 studies (n=22) conducted citation analyses of journals, articles, or dissertations. Following journals, studies of relationships within specific fields or theories were frequently conducted (n=10). Using citations to evaluate or compare research levels of colleges or universities was the third most frequent level on which citation analysis was used (n=7). Similarly, one study was also conducted of university departments, one examined scholarship presented at a professional conference, and one used citation analysis to examine an educational policy. Two of the studies examined educational issues on a national or international level, and two studies were conducted on individual

authors. None of the 44 studies that were found used citations to examine educational research centers or educational program evaluation impacts.

The wide diversity of educational areas in which citation analyses are being conducted and the range of types of studies being conducted suggest that some educational researchers agree that using citations as measures of individual, group, institutional, or publication impact is valid. While the use of citation analysis within education is not nearly as common as it is within other fields, there is a small and growing body of literature supporting such uses of citation information.

Citation analysis within the field of evaluation. While citation analysis has been used as a method to evaluate individuals and research centers, particularly within STEM areas, there was little evidence found in the ERIC search of its use within education. There has been some recent interest in citation analysis within the field of evaluation. A recent article in the *American Journal of Evaluation* discusses a participatory, mixed-methods approach to evaluate a large-scale initiative sponsored by the National Cancer Institute (Trochim, Marcus, Masse, Moser, & Weld, 2008). Bibliometric analysis was one of several analyses used to evaluate the research initiative. Citation data were used to create index variables to measure one key area of impact: the effectiveness of communication efforts through publications. In another *American Journal of Evaluation* article, a comparison of sixteen national models for evaluating government-funded research was conducted. Citation indexes were found to be used prominently in one of four models of research evaluation (Coryn, Hattie, Scriven, & Hartmann, 2007). A recent article in the *Journal of MultiDisciplinary Evaluation* examines more deeply the application of citation counts within the context of the evaluation of research, presenting

information about the strengths and weaknesses of the methodology (Coryn, 2006). Positive audience reaction to a presentation about this dissertation provides another indicator of interest within the field (Greenseid, 2008). Although there has been some very recent interest in citation analysis methods within the field of evaluation, its application as a measure of evaluation influence has not previously been advocated.

The discussion above suggests that while there is a long history of using citation analysis within STEM areas, there is currently only a small body of studies using citation analysis within education, and only one example using citation analysis within STEM education. No examples were located that used citation analysis as a way of assessing the impact of program evaluations within STEM fields.

Judgments of evaluation theorist panel. The three evaluation theorists were asked to judge the use of citations within STEM education evaluation; their responses are presented below. The theorists were asked whether they agreed, somewhat agreed, somewhat disagreed, or disagreed with the statement: *References are a convention used within STEM education and evaluation papers to give credit to influential ideas or methods.*

Table 5. Evaluation Theorist Panel Responses

	Strongly Agree	Agree	Disagree	Strongly disagree
References are a convention used within STEM education and evaluation papers to give credit to influential ideas or methods.	0	3	0	0

As shown in Table 5, all three of the theorists agreed that references are used within STEM education and evaluation papers to give credit to influential ideas or methods. Patton cautioned however, that "citation alone does not tell us the nature,

quality, and purpose of the use/influence" (M. Patton, personal correspondence, March 8, 2008). Similarly, Kirkhart stressed that while giving credit to influential ideas or methods is one plausible interpretation, examining the actual meaning of the citation within the context of the document is important. As she stressed, "if *influential* is taken to mean something that stimulates a strong critical reaction – positive *or* negative – then I would certainly agree, but if *influential* implies emulation or replication, then I think it is too narrow" (K. Kirkhart, personal correspondence, March 30, 2008). Alkin also expressed concerns about what he called "throwaway" references. As he elaborated, "authors feel that a good paper should cite certain authors in order to be considered thorough" (M. Alkin, personal correspondence, April 2, 2008).

Assumption 2: The content of citations to STEM education evaluation products suggests they are used to give credit where credit is due or represent other indicators of influence.

As stressed in Kirkhart's comments above, a key assumption underlying the validity of using citations as measures of influence on the STEM education and evaluation communities is that individuals cite evaluation products in ways that indicate that they have been influenced by that product. The converse would be if individuals provided references to evaluation products but these references were not meaningful indicators of influence because they were misattributed, just included as window dressing, or the like. This assumption was explored first by examining the theoretical and empirical literature within the field of citation analysis regarding the content of citations and people's motivations for citing. Then, a content analysis was conducted on a random

sample of 30 citations to the evaluation products used in this study to determine the extent to which the citations supported the notion that they were used in ways that indicated influence.

Literature review. While Chapter Two presented a number of theories on what citations mean, most scholars in the field of citation analysis agree that at its most basic level a citation is an indicator that a cited work was used in some way (Garfield, 1979a; Shadish, Tolliver, Gray, & Gupta, 1995; Small, 1982). As explained by Shadish et al,

> So what do citation counts mean? The literal view is that higher citation counts mean that a work was *used* more often, so had more impact. In a strictly operational sense, this has to be true. But many authors attribute additional meanings to highly cited works, including importance, creativity, quality, eminence and persuasiveness (Shadish et al., 1995, pp. 477-478).

Content analysis studies conducted by information scientists have developed a number of citation categorization schemes to better understanding how citations are used within citing works or to describe the rhetorical nature of the citations. One commonly discussed and applied scheme was developed by Moravcsik and Murugesan (1975). Their classification scheme includes four dimensions regarding the nature of citations: 1) conceptual (e.g., theoretical) vs. operational (e.g., connected to a tool or method); 2) organic (e.g., necessary for understanding the paper) or perfunctory (e.g., a general acknowledgment); 3) evolutionary vs. juxtapositional (e.g., built on the work's foundation or in opposition to it); and 4) confirmative or negative (e.g., disputing a paper or confirming it). Moravcsik and Murugesan analyzed citations to 30 articles published within the field of high energy physics and concluded that the citations were somewhat

balanced, but tended to be slightly more conceptual than operational and evolutionary as opposed to juxtapositional. They also found that although a majority of citations were organic almost two-fifths were perfunctory and about 14% were negational. They concluded that their findings raise questions about the validity of using citations as measures of quality as a number of citations were critical of the quality of the works they cited.

The assertion that a troubling percentage of citations are negative in nature has been challenged in a number of empirical follow-up studies, including, most importantly, a meta-analysis in which the levels of negational citations was found to be smaller, as low as 1% in some studies, than in the original study (Small, 1982). There are also a number of works that challenge the conclusion that negative citations are problematic for citation validity on theoretical grounds. This line of reasoning argues that while negative citations may challenge the use of citations for assessing the quality of a paper, they remain valid indicators of the influence of that paper on discourse within a field (Cole & Cole, 1971; Garfield, 1979a; White, 2004). As argued by Cole and Cole, "If a paper presents an error that is important enough to elicit frequent criticism, the paper, though erroneous, is probably a significant contribution. The significance of a paper is not necessarily determined by its correctness" (p. 25).

There are other existing coding schemes that complement the Moravcsik and Murugesan scheme. Some of these are derived by conducting content analyses of citations, along the lines of Moravcsik and Murugesan (Chubin & Moitra, 1975; Swales, 2001; White, 2001). Other studies derive taxonomies of citation motivations from

interviews or surveys about scholars' intentions in citing particular works (Leydesdorff & Amsterdamska, 1990; Peritz, 1983; Shadish et al., 1995).

The content coding scheme used in this study was informed by two previous studies: the first (Spiegel-Rosing, 1977) represents the first tradition in citation content analysis studies, while the second (Shadish et al., 1995) comes out of the second tradition of surveying citers. Spiegel-Rosing's coded citations to 66 articles published within the *Science Studies* journals between 1971 and 1974. She derived 13 content categories describing the ways in which cited works were used by the citing works. Of the 13 categories, the most prominent were (1) "cited source substantiates a statement or assumption, or points to further information," accounting for 80% of the citations, (2) "cited source was mentioned in the introduction or discussion as part of the history and state of the art of the research question under investigation," accounting for 5.8% of the citations, and (3) "cited source contains the data… which are used for comparative purposes in tables and statistics," which accounted for another 5.3% of the citations. The remaining 9% of citations were divided among the other 10 categories. One category that was not frequently represented in Spiegel-Rosing's work, but is applicable in this study, was "cited source contains the method used" in the study.

Shadish et al. conducted a survey of individuals who had cited works published in psychology journals asking them which of 28 possible reasons was most important to why they cited a particular work. The top four items accounted for over half of the responses: (1) "this reference supports an assertion in the sentence in which it occurred," (2) "this reference documents the source of the method or design feature used in your

study," (3) "this is a classic reference in the field," and (4) "this reference is a 'concept marker' – it represents a genre of studies, or a particular concept in the field."

Content of citations – empirical examination. While the previous literature review presents a number of ways of classifying the content of citations, no one classification scheme was discovered that explicitly tried to judge the type of influence of the cited work on the citing work. The categorization scheme used in this study, therefore, was informed by the theoretical and empirical literature about evaluation use and influence to assess the types of uses and influences evident in the content of the collected citations. The 30 randomly selected citations were coded according to how the cited product was used within the citing work, using a categorization scheme derived from Spiegel-Rosing (1977) and Shadish et al. (1995), but tailored to fit an evaluation context. As some of the sampled works contained citations to the product within more than one section, the total number of citations that were analyzed was 35.

The following eight categories were used to code the 35 selected citations:

1. Classic - The cited work is described as a classic or exemplar in the field;
2. Concept Marker - The cited work is a 'concept marker' – it represents a genre of studies, or a particular concept in the field;
3. Factual Statement - The cited work is used to substantiate a factual statement or assumption;
4. Further Information - The cited work provides further information about the topic in the sentence;
5. Empirical Findings - The cited work's empirical findings are presented;

6. Instrument/Method - The cited work provides information about the instrument or method used, or considered to be used, in the study or evaluation;
7. Validity - The cited work provides information about the validity or reliability of a method or instrument;
8. Resource - The cited work is included among a list of resources on a particular topic.

The citations were coded within the context of the paragraph or paragraphs in which they were contained, rather than narrowly examining their use within the particular sentence in which they fell. For example, while a reference may have been embedded in a sentence that presented a basic fact, in following sentences the detailed findings for the study was presented. This type of citation was coded as "empirical finding" rather than "factual statement."

As shown in Table 6, almost half of all of the citations (45.7%) were references to instruments the citing works either used, or considered using, in their own study. Presentations of empirical findings were the next most frequent type of citations (17.1%), with factual statements, references to evidence of validity for chosen instruments, and the inclusion of evaluation products in resource lists encompassed approximately 9% each. Two of the remaining four citations were examples of concept markers (5.7%), where the evaluation product was included within a list of studies addressing the same topic or using the same method. One citation was a reference to more information on the topic (e.g., "see XXX for details"), and one of the selected citing works was determined not to represent a citation in the sense of the study. This was an NSF official announcement,

part of the Federal Register, requesting comments regarding the proposed continuation of collection of data using instruments developed by the LSC core evaluation. While it does refer to the LSC instruments, and hence is understandable how it was captured as a citation during the data collection phase of this study, it is not citing the LSC instruments but rather requesting information and feedback about their use. None of the citations selected for this coding study was found to be used as citing a "classic work" in the field.

Table 6. Content codes for selected citations

	N	%
Instrument/Method	16	45.7
Empirical Findings	6	17.1
Factual Statement	3	8.6
Validity	3	8.6
Resource	3	8.6
Concept Marker	2	5.7
Further Information	1	2.9
Not citation	1	2.9
Classic	0	0.0
Total	35	100.0

As nearly half of the citations were references to evaluation instruments, further examination was warranted regarding exactly how and why the authors stated the instruments were used. The first finding is that two of the sixteen instrument citations were references to instruments that were collected and reviewed but determined not to fit the needs of the researchers and therefore were not used in their studies. Both studies stated that they considered but rejected the instruments as they did not meet their needs. One paper explained that the researchers intended to use an existing instrument, however, the instruments they found were all developed to evaluate inquiry-based approaches to mathematics teaching and their particular study was of a different approach. As stated in

the second paper, “In considering these observational tools, we found that they were too broad to be of use for the mentor teachers” (Zubrowski, 2007, p. 863).

A second finding is that almost half of the studies citing evaluation instruments or methods stated that they modified the instruments or methods, or only used particular sections or items from the instruments, in the development of their own measures. Seven of the fourteen citations to instruments that were used stated that they adapted the instrument or method in some way. One study’s researchers specified that they only chose items that were predictive of their particular research focus, another stated that they removed numerical rating scales and used the instruments in more qualitative ways, and the remainder just stated that they adapted or modified the instrument or approach.

A third finding is that four of the citations to instruments were from NSF evaluation projects for which the instruments or models were specifically developed and, in some cases, were required to be used. In two cases, these were program evaluation self-citations, referencing the instruments used within the program-level evaluations, and two were citations from projects: one from an LSC project, and one from an MSP project. These examples appear to be classic examples of instrumental use of an evaluation.

There were also three cases in which the instrument developed for one of the program evaluations under study here was used as originally constructed within another context. One example was from a Center for Learning and Teaching proposal in which an LSC evaluation instrument was cited as selected for use. The two remaining examples did not have any direct connection to the evaluation that produced the instrument. Instead, one study stated that they utilized the instrument because it was one of very few they found that existed, and the other provided a number of reasons for the instrument’s

use. As the researcher wrote in an electronic listserv, "The benefit of this instrument is that it has been approved by a federal agemcy (sic), it's comprehensive, and it takes qualitative data and codes it into quantitative data" (Danin, 2000).

Table 7. Findings regarding instrument/method citation

	N	%
Modified instrument	7	43.8
Instrument was developed for use in the evaluation	4	25.0
Used instrument in another context	3	18.8
Did not use	2	12.5
Total	16	100.0

Assumption 3: Citation databases exist that provide adequate coverage of the STEM education and evaluation fields.

Sources must exist to obtain representative citation information in order for citation analysis to be used validly to assess the influence of STEM education evaluation products. Each of the three databases utilized in this study has its advantages and disadvantages. While Web of Science is the preferred method of conducting large-scale citation analyses as its controlled vocabulary allows for greater refinement of search results, it is known that disciplinary coverage varies within the database. For example, Moed (2005) found that coverage of the Web of Science was excellent in several areas of science, including physics, chemistry, molecular biology, and biochemistry, was good but not excellent in coverage of engineering, biological sciences, mathematics, and psychology, and moderate in other social sciences including sociology, political science, educational sciences, and the humanities (p. 3).

Several empirical studies have been conducted comparing the citation results of Google Scholar and Web of Science. A review of 10 comparative empirical studies conducted between 2005 and 2007 found that the two databases were overlapping and complementary (Schroeder, 2007). The review concluded that while Google Scholar indexes a greater amount of grey literature (i.e., writings not found in academic journals such as conference presentations, preprint or unpublished manuscripts, books, newsletters, etc.) than Web of Science, its coverage of peer-reviewed journals has been found to be less complete for certain disciplines. Web of Science was found to provide access to older articles, while Google Scholar had greater coverage of recently published works. Overall, the article concludes that Google Scholar's strength is its open access, ease of use for novice users, and broad coverage, however, it does not provide an easy way to retrieve data for further analyses and manipulation, instead requiring tedious manual data collection and cleaning.

The Google search engine is not a formal citation database that indexes reference lists within specific journals. Google catalogs any print material accessible through the internet, including grey literature as well as references posted in even less formal ways such as on organizational or personal websites. In addition to formal citations, Google also captures online bibliographies or other electronic sites that catalog resources without formally citing them. In this study, these are referred to as "electronic resources" although in other studies the term "Web hyperlinks" has been used (Vaughan & Shaw, 2003). While there is some controversy over including electronic resource references along with formal citations in citation analyses (Egghe, 2000; van Raan, 2001), others have argued that including electronic resource references gives a fuller picture of

intellectual influence. For example, Sloan argued that citations within course syllabi and reading lists accessible on the Web may actually be a greater indicator of influence than research citations as they demonstrate the transfer of knowledge to a new generation of scholars (Sloan, 2001, as quoted in Vaughan & Shaw, 2003). Based on these arguments, this study attempted to capture as many references to the products as possible to get the broadest picture of influence on the fields. Consequently, the choice was made to use all three databases, and citations found in peer-reviewed journals, grey literature, and electronic resources were all collected and analyzed in this study.

Coverage. Coverage of the STEM education and evaluation literature was assessed by examining published coverage lists and conducting database queries for Web of Science and Google Scholar. As Google does not index reference lists within journals in the same way as the other citation databases, it is not possible to do the same sort of research on "coverage" by Google. First, two lists were compiled of the top approximately one dozen journals in the fields of evaluation and STEM education. These lists encompassed both more research-oriented and practitioner-oriented publications, as well as prominent international journals in addition to US publications and some content-area specific journals to try to capture key areas within the evaluation and STEM education fields. Then, the coverage of the journals on the lists was checked against the Web of Science's published covered journal list. Google Scholar does not publish what it indexes; therefore its coverage was assessed through searching the database manually.

Table 8. Database coverage of a selection of journals in the field of program evaluation

Journal	Web of Science (as of 8/16/07)	Google Scholar
American Journal of Evaluation	Yes	Yes
Educational Evaluation and Policy Analysis	Yes	Yes
Educational Research and Evaluation (Netherlands)	No	Yes
Evaluation and Program Planning	Yes	Yes
Evaluation Review	Yes	Yes
Evaluation: The international journal of theory, research, and practice (UK)	No	Yes
Journal of MultiDisciplinary Evaluation (online journal)	No	Yes
Journal of Personnel Evaluation in Education	No	Yes
New Directions for Evaluation	No	Yes
Research Evaluation (UK)	Yes	Yes
Studies in Educational Evaluation (UK)	No	Yes
The Canadian Journal of Program Evaluation (Canada)	No	Yes

Table 9. Database coverage of a selection of prominent journals in STEM education

Journal	Web of Science (as of 2/29/08)	Google Scholar
International Journal of Science Education (UK) – research oriented	Yes	Yes
Journal for Research in Mathematics Education (NCTM) – research oriented	Yes	Yes
Journal of Chemical Education – content area specific	Yes	Yes
Journal of College Science Teaching – practitioner oriented	No	Yes
Journal of Computers in Mathematics and Science Teaching – content area specific	No	Yes
Journal of Engineering Education – content area specific	Yes	Yes
Journal of Research in Science Teaching (US, NARST) – research oriented	Yes	Yes
School Science and Mathematics – research oriented	No	Yes
Science Education – research oriented	Yes	Yes
Science Scope – practitioner oriented	No	Yes
Teaching Children Mathematics – practitioner oriented	No	Yes
The American Biology Teacher – content area specific	Yes	Yes
American Journal of Physics – content area specific	Yes	Yes
The Science Teacher – practitioner oriented	No	Yes

The tables above clearly show that Google Scholar provides indexing for more key journals in the STEM education and evaluation fields than does Web of Science. While Google Scholar covers all of the selected journals in both fields, Web of Science only covers five of twelve (41.6%) of the evaluation journals and eight of fourteen (57.1%) of the STEM education journals. Additionally, Web of Science covers primarily research-oriented and US-based journals. The journals that are practitioner-oriented, internationally-originating, or content-area focused are not well represented within Web of Science.

While journals are one important source of citations to the evaluation products, a better measure of influence on the fields of evaluation and STEM education includes references from grey literature and informal sources. The fields of STEM education and evaluation are heavily practitioner-oriented so examining just journal citations would bias the study toward only assessing the influence of the products on the more academic and research-oriented areas within the fields. To attempt to examine this question, data from the citation analysis of the four program evaluations were used as a sample to study differences in the types of citations found by the three databases. Although it is possible to describe differences in the types of citations, the degree to which the citations represent the true numbers of citations contained within these sources is unknown.

Number of citations. First, the question of whether the databases found significantly different numbers of citations per product was addressed. As shown in Table 10, the total number of citations found for each of the three databases was different. Web of Science found 24 citations, while Google Scholar and Google found nearly ten times that number.

Table 10. Citations found by Web of Science, Google Scholar, and Google

Search Engine	Total Citations	Mean Citations per Product[7]	SD
Web of Science	24	.10	0.526
Google Scholar	205	.83	2.754
Google	216	.88	3.474

Differences within the mean number of citations per product found by the three databases were tested using paired t-tests. Both Google Scholar and Google found significantly higher mean citations per product than did Web of Science (GS-WoS: t=4.525, df=245, p<.001; Google-WoS: t=3.647, df=245, p<.001). There was no significant difference between the mean number of citations found by Google Scholar and Google, however (t=-.334, df=245, p=.739).

Reference sources. Next, this study examined differences in the sources of the 376 references[8] found by the three databases, meaning whether the citation was found in a reference list printed within a publication, report, presentation, instrument/tool, electronic resource, or other source. As shown in Table 11, the Web of Science found only references contained within publications as would be expected as the database only indexes a select number of journal reference lists. Google Scholar and Google found references within a diversity of sources.

[7] The distribution of the mean citations per product variable was positively skewed, with a large number of products receiving zero or one citation only. Therefore p-values reported in this study should be interpreted cautiously.

[8] There were 376 total references found to the 246 evaluation products used as the sample in this study. These 376 references were contained with 280 unique citing works, however, many works cited one or more products.

Table 11. Sources of the citations found by the three databases

Citation Sources	Web of Science	Google Scholar	Google	Total Unique
Electronic Resource	0	2	49	49
Instrument/Tool	0	7	3	9
Other	0	2	15	16
Presentation	0	23	32	48
Publication	24	94	68	139
Report	0	77	49	115
Total	24	205	216	376

A chi-square analysis was used to test differences within the distributions of the reference sources found by Google Scholar and Google. The types of sources found by Google Scholar and Google was found to be significantly different[9] (χ^2= 66.481, *df* = 5, *p*<.001). Post-hoc tests found that Google captured a significantly higher proportion of electronic resources and items coded as "other", while Google Scholar found a greater number of evaluation reports. Interestingly, the two databases did not differ significantly in terms of the number of citing publications, instruments, or presentations.

Unique citations. In addition to the number of citations and the sources of the citations, information about the contribution of each database to finding the 376 total citations was examined. As shown in Table 12, Web of Science alone accounted for less than 7% of the total number of citations found using all three databases. While Google Scholar and Google independently each found a little over half of the total, together they captured 98.4% of the total. Another way to look at the data is to highlight that Web of Science contributed six unique citations of the 376 total citations found using all three databases.

[9] Note: one cell had an expected count of 4.87, however, as this represents approximately 8% of the cells it should not affect the overall outcome of the test.

Table 12. Number of citations found for combinations of databases

Database	Citations	% of total
Web of Science	24	6.3
Google Scholar	205	54.5
Google	216	57.4
WoS + Google Scholar	215	57.1
WoS + Google	234	62.2
Google Scholar + Google	370	98.4
Total (WoS + GS + Google)	376	100.0

Assumption 4: The process of gathering STEM education evaluation product citation data can be conducted accurately.

An important concern as to the validity of citation analysis methods for assessing STEM evaluation influence is whether the method of collecting citation data can be done in a way that is accurate – in other words, in a way that is consistent across replications. The term reliability is used to describe the consistency of data obtained through a measure. The reliability of an instrument can be calculated through a variety of procedures, however, these calculations depend on the ability to estimate which part of the variance of a measurement is due to random error and which is due to true differences within the individuals being measured.

In this study, it is not possible to distinguish between differences in "true" citation counts and the observed differences within the citation counts as it is impossible to replicate exactly the process of collecting citation count data. To replicate collecting of citation information either the same researcher needs to collect the data again at a future date, which introduces the variable of time, or a second researcher can collect data at the same time, which introduces the variable of differences within the researchers. These

types of "systematic errors" preclude calculation of traditional measures of reliability. Instead, this study will examine patterns in the citation data resulting from replication studies conducted at future time points and with different researchers to try to tease out some of the effects of these variables on the collection of accurate citation information.

The accuracy of the citation data and method were examined by having two researchers try to replicate the original citation data results at a second data collection time point with a random selection of 25 evaluation products from the four program evaluations. The researchers were the original primary researcher and one other member of the Beyond Evaluation Use project team who was trained by the original researcher. The researchers were blind to the original findings when collecting the citation information. The replication by the primary researcher was conducted between January 14 and 16, 2008. The second researcher conducted her replication between January 14 and February 19, 2008. Searching was completed for the original study between August 27 and December 4, 2007 (see Table 2, Chapter 3 for specific dates for each evaluations). The original data collection took place, therefore, between five months and six weeks prior to the replication study.

As shown in Table 13, the total number of citations found for the 25 products using each search engine during the collection trials was the same for Web of Science, but greater for both Google Scholar and Google for the replication trials conducted at later time points.

Table 13. Number of citations found for original and replication studies

Database	Original Study	Replication by Primary Researcher	Replication by Secondary Researcher	Total (combining all three trials)
Web of Science	3	3	3	4
Google Scholar	19	25	26	36
Google	10	21	15	27
Total (Unique Citations)	27	37	34	47

Analyzing the data statistically, a significant difference was found only in the mean number of citations per products between the original data and the replication by the primary researcher (t=-2.449, df=24, p=.022). Neither the differences observed between the original and secondary researcher replication nor the differences between the primary and secondary researchers during the replication study were found to be statistically significant. While somewhat counterintuitive, this finding perhaps suggests that researcher experience is an important factor in the quality of data found in citation studies. During their first collection trials the primary and secondary researcher collected similar amounts of data, despite differences in the timing of the collection periods. However, the primary researcher maybe have gotten "better" at collecting information and therefore found more data during her second data collection attempt.

In addition to examining overall citation counts, the percent agreement of the citations among the three trials was calculated. The percent agreement examines whether the databases found exactly the same citations, not just the number of citations found as examined above. As shown in Table 14, the percent agreement for the Web of Science data was extremely high between each of the trials, while the Google Scholar and Google

results have lower agreement levels due to the greater number of citations found during the replication studies.

Table 14. Percent agreement of citation information

Search Engine	Original and Replication by Primary Researcher	Original and Replication by Secondary Researcher	Replications by Primary and Secondary Researchers	All Three Trials
Web of Science	96.6%	96.6%	100.0%	96.6%
Google Scholar	72.9%	64.4%	78.0%	57.6%
Google	67.8%	78%	76.3%	61.0%

Assumption 5: Citation data can be transformed into meaningful indexes for comparing levels of STEM education evaluation product influence.

This assumption addresses the question of whether there are meaningful ways of transforming raw citation counts into indexes for comparing product influence levels across different evaluations. While raw citation counts provide one indication of the absolute number of citations to the products for a particular evaluation, in many cases it is important to be able to compare different evaluations that vary in terms of their funding periods and levels, and therefore in the expectations of the number of products they would produce. In the case of the sample data, for example, at the time the citation data were collected, the MSP-RETA evaluation was funded for five years and for $1.8 million, CETP was funded for five years and $990,000, ATE was funded for seven years and $3.1 million, and LSC was funded for ten years and $6.5 million. Clearly the expectations for the number of products these evaluations would produce and the amount

of influence these evaluations would have in terms of citations should be different. The question remains as to whether there are meaningful indexes that can be calculated to control for the different sizes and scopes of evaluations. A review of the literature in this topic will be followed by an examination of the utility of existing citation indexes with the data collected for this study in the following sections.

Review of the literature. The citation analysis literature has presented a few solutions to the problem addressed above. Arguably, the most influential work in this area was written by Hirsch in 2005. Hirsch reviews several existing indexes for measuring, comparing, and evaluating the scientific output of researchers and then proposes a new index, the "*h* index," which he believes overcomes the weaknesses of alternative approaches (summarized in Table 15). The *h* index is calculated as follows:

> A scientist has index *h* if *h* of his or her *Np* [number of papers published over n years] papers have at least *h* citations each and the other (*Np*-*h*) papers have ≤ *h* citations each. (Hirsch, 2005, p. 16569)

For example, a scientist who, over the course of 7 years, published 15 papers, 10 of which had 10 citations or more with the remaining 5 having fewer than 10 citations, has an *h* index of 10. Hirsch states that the *h* index provides an estimate of a scientist's (or group of scientists') "importance, significance, and broad impact" (p. 16572), while overcoming the shortcomings of the other indexes described below. As Hirsch argues, "two individuals with similar *h*s are comparable in terms of their overall scientific impact, even if their total number of papers or total number of citations is very different" (p. 16569). Hirsch cautions that while the *h* index is a meaningful metric for comparing

researcher's impact, the comparisons must only be done between scholars within the same field, as publication and citation rates vary greatly between disciplines.

Table 15. Comparison of exiting citation indexes (adapted from Hirsch, 2005)

Index	What measures?	Limitations
Total number of papers	Scientific productivity	Does not account for the importance or actual impact of the papers
Total number of citations	Total impact	Difficult to accurately find total number of citations; index easily inflated by one or two highly-cited works
Citations per paper	Comparative impact controlling for the authors' publishing lifespan	Biased toward the publication of a few, highly-cited works; disadvantages steady research contributions
Number of "significant papers" (defined according to some criterion)	Broad and sustained impact	Criterion needs to be clearly defined and defended; biased against more junior researchers or projects funded for shorter periods
Number of citations to each of the top X most-cited papers	Broad and sustained impact controlling for lifespans	X is defined arbitrarily – will favor some authors over others; produces multiple numbers, hard to compare

Several studies have been conducted to examine the convergent validity of the *h* index with other measures of scholarly impact (e.g., Bornmann & Daniel, 2006; van Raan, 2006), finding that the *h* index correlates highly with peer review judgments of influential researchers and research groups. Critics of the measure argue that it is too simplistic to reduce researcher impact to a single number, that the index advantages more established researchers who have published a greater number of papers over younger scientists without long publications to their record or individuals who wrote only a few highly influential papers, and that the measure lacks accuracy and precision to be used at

the individual scientist level (Bornmann & Daniel, 2007; Egghe, 2006; Lehmann, Jackson, & Lautrup, 2005). Despite some critics, in just the few years since its proposal, the *h* index has been used in a number of studies on researchers' productivity, including in a comparison of the productivity of chemistry research groups (van Raan, 2006), high-energy physicists (Lehmann, Jackson, & Lautrup, 2006), and scholars in the *Journal of the American Society of Information Science* (Rousseau, 2006), among others.

Recently an alternative to the *h* index, the *g* index, was created by Egghe (Egghe, 2006). The *g* index was created to make a measure more sensitive to highly cited papers. Egghe argues that two authors with the same *h* indexes may have vastly different number of citations to their top articles; the *h* index is insensitive to these highly cited works and thus equates authors who most individuals would agree have not had the same levels of influence. The *g* index is defined as "the highest number *g* of papers that together received g^2 or more citations" (p. 8). This index can be calculated from the same data needed to calculate the *h* index, but Egghe argues it is more helpful in distinguishing between scholars' true levels of impact.

Comparison of citation indexes using the NSF program evaluation data. Six of the quantitative indexes presented above were compared to assess whether different conclusions about the influence of the program evaluations are obtained using different indexes. As shown in Table 16, the total number of products (a measure of productivity and not influence) was a dichotomized measure with the sample data, ATE and LSC having the same number of products (n=98) and CETP and MSP-RETA having approximately the same number (n=24 and n=25 respectively). Ranking the evaluations by their total impact, measured by the total number of citations, finds that the LSC

evaluation had 247 citations, 3.8 times as many citations as the next highest, which was the ATE evaluation with 64 citations. CETP had 42 citations and MSP-RETA had ten citations. Comparisons of these numbers suggest that the LSC had a much greater overall impact than the other evaluations. How do the other indexes compare?

Table 16. Comparison of citation indexes

Index	What measures?	ATE	CETP	LSC	MSP-RETA
Total number of products	Scientific productivity	98	24	98	25
Total number of citations	Total impact	64	42	247	10
Citations per product	Comparative impact controlling for differences in evaluation output	.65	1.75	2.53	.40
Number of citations to each of the top X most-cited papers [here X = number of products within top 5% of cited products]	Broad and sustained impact controlling for lifespans	0	1	10	1
H index	Importance, significance, and broad impact	4	3	8	1
G index	Importance, significance, and broad impact; more sensitive to highly-cited papers than the *h* index	4	6	14	3

Figure 7 illustrates that in all of the other four citation indexes (number of citations per product, number of highly cited products, *h* index, and *g* index), the LSC evaluation products had the greatest levels of broad impact, importance, or significance among the four evaluations in this sample. Rankings of the remaining three evaluations vary among the three, however. Both the citations per product and *g* index conclude that

the ranking is LSC, CETP, ATE, then MSP-RETA, however, the *h* index finds that the ATE evaluation's products had a greater impact than CETP or MSP-RETA. The highly-cited product index appears to do little to distinguish between the evaluations, with both CETP and MSP-RETA having a score of "1" and ATE being a "0."

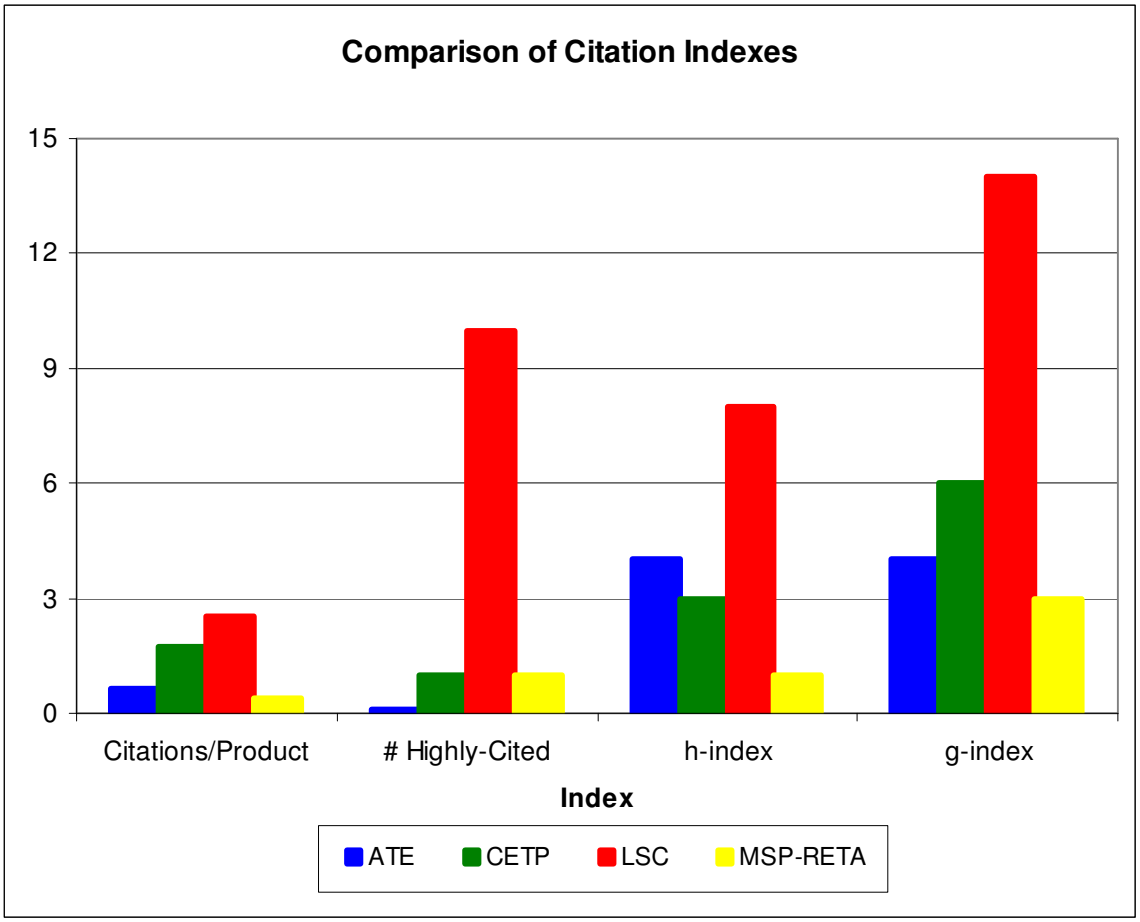

Figure 7. Comparisons of Citation Indexes

In addition to the relative rankings produced by the indexes, the magnitude of differences suggested by the indexes is illuminating. As discussed above, the total citation counts suggest that the LSC products had a much greater impact than the products from the other evaluations. The citations per product index suggests much more similar levels of impact across the four evaluations, which is confirmed by examining the bivariate correlations between program evaluation (ATE, CETP, LSC, MSP-RETA) and number of citations. Significant differences were found in the average number of citations per products produced by the four program evaluations ($F=2.985$, $df=3$, $p=.032$). However, these differences account for only 3.6% of the variance in mean citation counts, and, when included in a multiple-regression model that includes other variables related to variation in citation levels, program evaluation is found not to be significant.

Assumption 6: Citation indexes are related to other measures of STEM education evaluation influence.

As presented in the literature review chapter, a number of studies have found citation indicators to be related to other measures of research impact such as peer review. This assumption examines whether there are relationships between citation indexes and other measures of influence for the four STEM education evaluations examined in this study. The ideal design to gather data for this question would be to correlate citation indexes with responses from a survey of a broad sample of STEM educators and evaluators who were asked to assess the influence of each these program evaluations.

Unfortunately, at the time of the writing of this paper no such survey has been conducted, although one is currently being designed for the Beyond Evaluation Use project.

Instead, the data used in the following analyses come from a survey and interviews conducted by the Beyond Evaluation Use project of the PIs and evaluators who were involved at the project level with the four program evaluations. These data are problematic in several ways. The survey respondents and interviewees were more closely connected to the program evaluations than the "average" person in the STEM education field. Moreover, the respondents were only commenting on the use and influence of the one specific evaluation with which they were affiliated. Differences in the responses to the questions, consequently, may be linked to individual differences in the respondents and respondent groups as opposed to reflecting true differences in use and influence levels. In other words, it is possible that lower perceived levels of influence for one of the evaluations is due to the respondents in that group being generally more critical than one of the other groups of respondents.

Additionally, as mentioned previously, the data from the MSP-RETA survey should be particularly interpreted with caution. It is clear from both the survey responses and follow-up interviews that some proportion of respondents to the MSP-RETA survey were confused about exactly which evaluation they were supposed to be considering when answering the questions. Therefore it is possible that some respondents were thinking about an evaluation other than the MSP-RETA project conducted by Utah State when answering the questions. With those caveats in mind, the following analyses

compare the relative rankings of the influence of the four program evaluations between the citation indexes and survey data.

Comparison of the Citation Indexes and Survey Questions. Three questions related to the use of evaluation products and the influence of the evaluation were identified from the survey of project PIs and evaluators:

- Q48 - I used data collection instruments from the [ATE, CETP, LSC, or MSP-RETA] program evaluation in another evaluation. Rated on 1-4 scale: No; Yes - a little; Yes – some; Yes – extensively. *Note – this question was only asked of individuals who responded that they had participated in another evaluation since being part of one of the four evaluations, not on the full sample.*
- Q61 - How influential do you feel the [ATE, CETP, LSC, or MSP-RETA] program evaluation was on the STEM education community? Rated on 1-5 scale: Not influential at all; Marginally influential; Somewhat influential; Influential; Very influential.
- Q62 - How influential do you feel the [ATE, CETP, LSC, or MSP-RETA] program evaluation was on the evaluation community? Rated on 1-5 scale: Not influential at all; Marginally influential; Somewhat influential; Influential; Very influential.

Analysis of variance was used to examine differences in reported mean levels of use and influence across the four groups, followed by post hoc tests using the Tukey HSD adjustment. Significant differences were found in the reported use of the data collection instruments in other evaluations (F=7.404, df=3, p<.001) and for the perceptions of

influence of the program evaluations on the STEM education community (F=3.460, df=3, p=.017). No significant differences were found in the perceptions of influence levels of the program evaluations on the evaluation community.

Observed means for the question about the use of instruments in other evaluations are presented in Table 17. Post hoc tests found that mean for LSC was greater than both the means of ATE and MSP-RETA, and the CETP mean was greater than the MSP-RETA mean. Consequently, the ranking of the use of products according to these data is (from high to low): LSC, CETP, ATE, MSP-RETA.

Table 17. Q48 - I used data collection instruments from the [ATE, CETP, LSC, or MSP-RETA] program evaluation in another evaluation.

Evaluation	N	Mean	SD
ATE	68	2.10	1.053
CETP	31	2.48	1.208
LSC	54	2.78	0.925
MSP-RETA	21	1.67	1.065
Total	174	2.33	1.103

* Rated on 1-4 scale: No; Yes - a little; Yes – some; Yes – extensively.

Observed means for the question about the influence of the program evaluations on the STEM education community are presented in Table 18. Post hoc tests found significant differences only between the mean level of influence of the LSC evaluation and MSP-RETA evaluation. Consequently, the ranking (high to low) of the influence of the program evaluations is: LSC, CETP/ATE (tie), MSP-RETA.

Table 18. Q61 - How influential do you feel the [ATE, CETP, LSC, or MSP-RETA] program evaluation was on the STEM education community?

Evaluation	N	Mean	SD
ATE	177	2.82	1.108
CETP	51	2.86	1.059
LSC	71	3.11	1.178
MSP-RETA	46	2.43	1.109
Total	345	2.83	1.128

*Rated on 1-5 scale: Not influential at all; Marginally influential; Somewhat influential; Influential; Very influential.

As shown in Table 19, six of the seven different indicators of influence (four citation indexes and three questions from the surveys) show differences among the influence levels of the four program evaluations. In the six indicators that had differences, all six agreed that the LSC core evaluation had the greatest influence and the MSP-RETA evaluation had the least influence. The ranking of the influence of the ATE and CETP evaluations is split between the indicators. However, if the total number of citations was included in the analysis, the pendulum would swing toward the ATE evaluation having the greater influence of the two. Considering that the ATE evaluation was funded for a longer period, at greater expense, and had the explicit purpose of disseminating findings for part of its funding, it is noteworthy that the CETP evaluation achieved almost the same level of impact as the ATE evaluation.

Table 19. Rankings of program evaluations using citation indexes and survey data

Evaluation	Number of citations	Citations/ product	*H* index	*G* index	Survey – Instr. Use	Survey – STEM Ed Influence	Survey – Eval. Influence
ATE	2	3	2	3	3	2	N/S
CETP	3	2	3	2	2	2	N/S
LSC	1	1	1	1	1	1	N/S
MSP-RETA	4	4	4	4	4	4	N/S

Assumption 7: Citation analysis is useful for understanding differences in patterns of use and influence within and across STEM education program evaluations.

The previous assumption assessed the ability of citation indexes to distinguish between different levels of influence of STEM program evaluations. This assumption

examines whether citation data are useful for discerning differences in patterns of use and influence across STEM evaluation contexts. First, citation patterns are assessed quantitatively by examining relationships between citations and other factors that are expected to be related to the use and influence of the products. Second, patterns in the citation data are explored visually through citation maps. Third, the primary investigators of the evaluations were sent reports on the citation patterns for each of their program evaluations. They were asked questions regarding the extent to which they felt the citation data presented in the report were representative and accurate descriptions of their perceptions of the influence of their evaluation products.

Quantitative Analysis of the Data

This section examines relationships between citations and evaluation product types, fields, and content areas. Among the four evaluations a total of 245 evaluation products were produced through the end of 2006. A descriptive analysis of the evaluation types, fields, and content areas, along with definitions of the categories used in the analysis, are presented in Appendix A. Multiple regression was used to examine factors related to citations per product. For the analysis, the one product coded in the "education-general" field was recoded as STEM education, because despite having addressed general issues in recruitment and retention, it is also applicable to STEM. The regression model examined the relationship between a product's type, field, content area, and program affiliation, and the number of citations the product received. This regression model is only one of many that could be used to analyze these data, but it was chosen as it reflected the interest in the main effects for the factors.

The overall regression model was found to be significant (F=2.949, df=15, p<.001), however the total amount of variance in citations accounted for by the model was relatively small (r^2=16.2). (See Table 20.)

Table 20. SPSS output for regression model

Dependent Variable: Number of Citations per Product

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	931.395(a)	15	62.093	2.949	.000
Intercept	108.230	1	108.230	5.140	.024
Program	43.371	3	14.457	0.687	.561
Product Type	450.583	3	150.194	7.133	.000
Product Field	1.375	2	0.688	0.033	.968
Product Content Area	19.506	7	2.787	0.132	.996
Error	4821.805	229	21.056		
Total	6294.000	245			
Corrected Total	5753.200	244			

a R Squared = .162 (Adjusted R Squared = .107)

Among the variables entered in the model, only product type was found to be related to citation levels. As shown in Table 21, evaluation instruments and tools were cited most frequently, averaging 7.25 citations per product. Post hoc analyses found evaluation instruments had significantly higher mean citations per product than the other three product types, which ranged between zero citations per presentation to 1.72 citations per report.

Table 21. Citations per product for different product types

Evaluation Product Type	Number of Products	Number of Citations	Citations per Product
Instrument/Tool	20	145	7.25
Report	103	177	1.72
Publication	58	55	.95
Presentation	65	0	.00
Total	246	377	1.53

The program evaluation producing the product, the product's field, and the product's content area were not found to be related to citation levels. Although there were significant differences between mean citations of products produced by different program evaluations in bivariate correlations, as discussed previously, program evaluation did not contribute to explaining the variance in citation counts above and beyond that explained by the product's type.

This quantitative analysis is just one of many analyses that could be conducted using the collected citation data. For the Beyond Evaluation Use project, additional analyses were conducted including descriptions of relationships between product and citing work authors, patterns of self-citation, and examinations of the characteristics of highly-cited products, defined as those within the top 5% of all cited products (see: Greenseid et al., 2008). Together these types of analyses provide insights into what factors affect the number of citations of individual evaluation products and of overall evaluation knowledge production and dissemination efforts. Moreover, they can be used to explore the fields, content areas, and other researchers where the products have been found to have had an influence – as indicated through their citations. A second way of seeing the impact of the evaluations on the fields is presented next.

Visual Analysis of the Data

The visual analysis of citation relationships has been used to show the growth of concepts over time, relationships between journals or scientists, or patterns of co-citation between clusters of authors within specific disciplines (Cawkell, 2000). Using a network

analysis approach with citation data is still being developed, however. As Hargens states, "compared to the proliferation of many other types of citation-based studies of scholarship… the comparative study of reference networks is still relatively underdeveloped" (2000, p. 497). In the fall of 2007, Garfield released a beta-version of a citation visualization software program, *HistCite*, which facilitates the creation of network maps, such as those described above, with data from Web of Science. The program is not compatible with data obtained through other citation databases, however. In this study, therefore, the network analysis program *NetDraw* was used to create maps of citation relationships between and within the four program evaluations.

Networks of products and citing works. Figure 8 presents the network map for all the evaluation products and citations. In each of the maps presented below, arrows point from the citing work to the product it is citing. In the diagram, the squares represent products produced by one of the four program evaluations or the multi-evaluation product: ATE is blue, LSC is yellow, CETP is green, MSP-RETA is purple, and one product that was produced jointly by the ATE, CETP, and LSC evaluations is orange. The red circles represent works authored by individuals who were not members of any of the program evaluation teams.

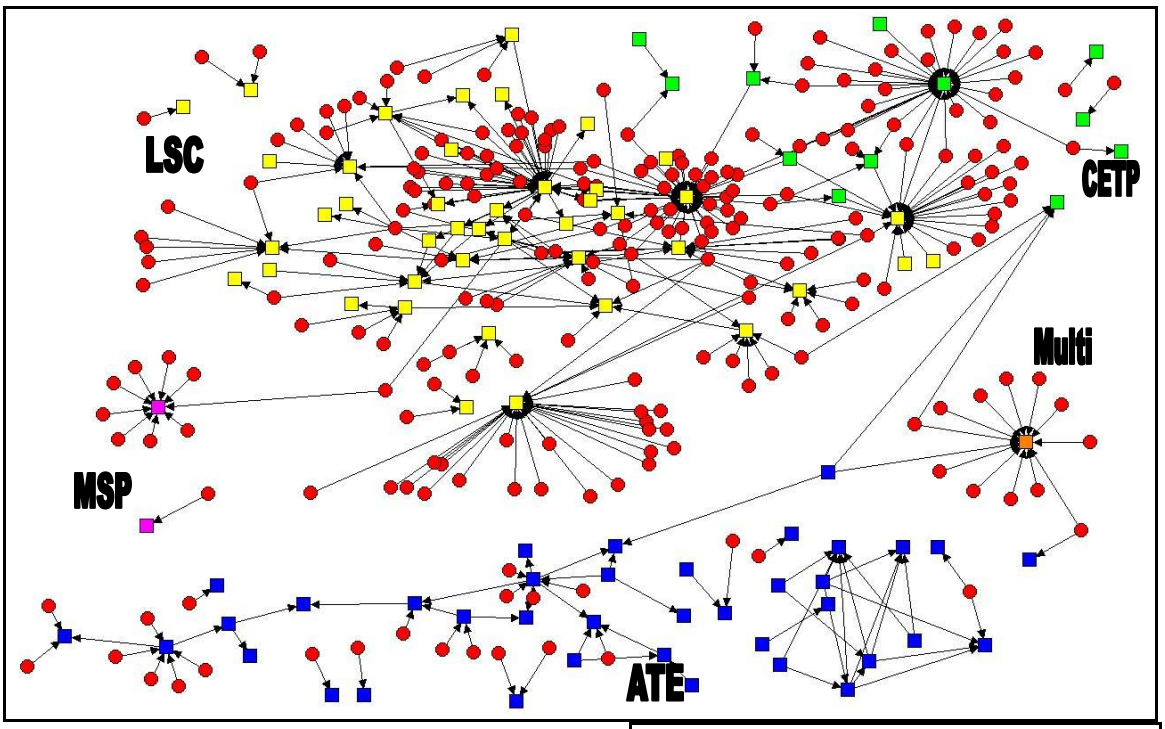

Figure 8. Citation network for the program evaluations

Color Key			
	ATE evaluation		MSP-RETA
	CETP core evaluation		Multi Evaluation
	LSC core evaluation		Other

As shown, there is an interconnected web of citations between and among the four program evaluations and outside sources. This web is most tightly connected between and within the LSC evaluation products, which is not surprising given that the LSC products obtained the highest number of citations. Viewed more closely, however, several connections between the LSC cluster and CETP cluster are also apparent. In addition to citations from CETP evaluation products to LSC evaluation products, there are several citations by outside sources to products produced by both evaluations. These connects show the influence of the LSC core evaluation, primarily through its instruments, on the CETP core evaluation, as well as the two evaluations' joint influence in the areas of teacher preparation and professional development. The ATE products and

MSP-RETA products are also connected to the LSC/CETP web, although more tangentially.

Individual program evaluation networks. Examining the citation networks for the ATE, CETP, and LSC program evaluations separately reveal some additional insights. The MSP-RETA data are not presented as only two products were cited and there were no connections between the citations or products.

ATE evaluation citation network. Figure 9 presents the overall citation network of ATE products (blue squares) and works authored by individuals outside of the ATE evaluation team (red circles). As shown on left side of the figure, there are a number of products that were cited by only one or two internal or external works. On the right side of the figure, there is a mass of ATE products that refer to each other, along with one external citation. Six of the products in this mass were published as part of the "Advanced Technological Education Program Evaluation Briefing Paper Series," and, as shown, commonly one paper in the series cited one or more of the other papers in the series. The one external citation was from an ATE Centers Impact report that was not affiliated with the ATE program evaluation.

In the center of the figure, there are a number of ATE products that are linked together. Many of the products in this citation chain were articles published together in the "The ATE Program: Issues for Consideration" monograph. The works in the monograph were cited by both internal and external sources in subsequent years. In general, the citation patterns presented below highlight that the citations to ATE products were primarily self-citations from other ATE products, although a number were from outside sources. This suggests that the evaluation products were used by individuals

connected to the evaluation, rather than having a large amount of documentable influence on external publications.

This does not mean, however, that the knowledge produced by the ATE evaluation was limited in its influence to a small group of evaluators conducting its day to day work. In fact, there were 24 separate individuals who were listed as authors on one or more of the ATE evaluation products, demonstrating that the ATE evaluation engaged a variety of internal staff and external consultants in its work. That said, these citation levels and patterns do suggest that the greatest influence of the ATE evaluation products was on those affiliated with the ATE national evaluation in some way, as opposed to the field more broadly.

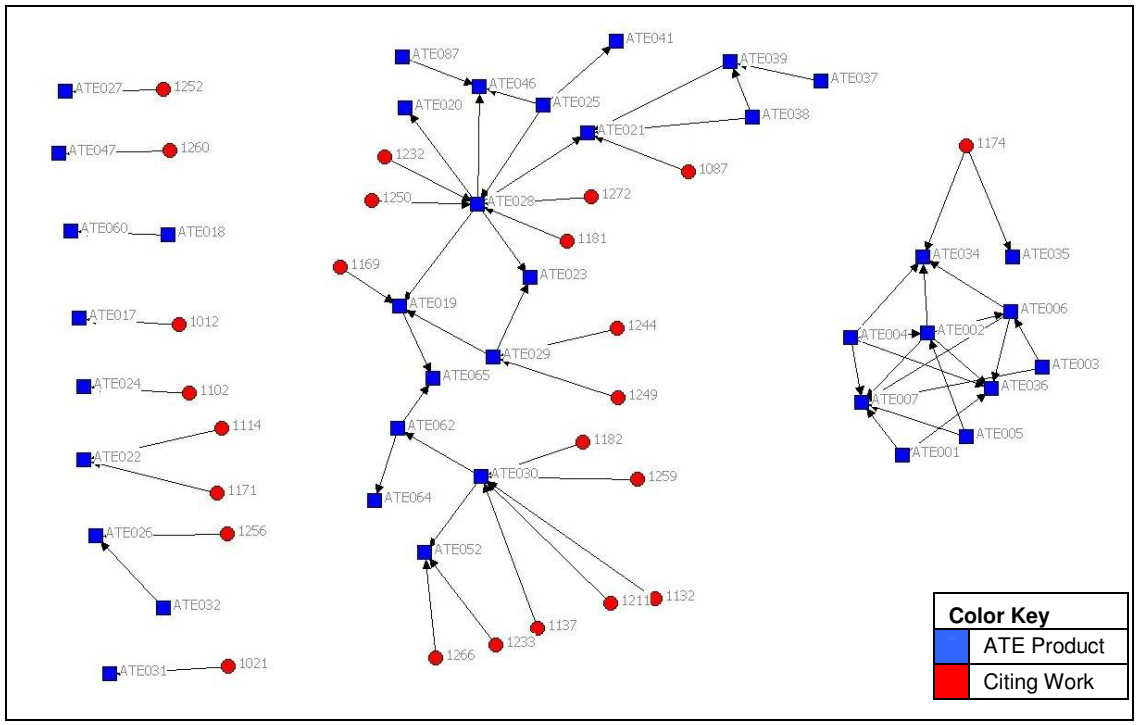

Figure 9. ATE products and citations

ATE evaluation author patterns. In Figure 10, the color coding scheme is changed to illustrate the different authors of the external citing works. ATE products are again represented as blue squares. The citing works are colored as to whether they were affiliated with one of the other three program evaluations under consideration in this study (i.e., CETP, LSC, and MSP-RETA), other NSF sources, or from sources not affiliated with NSF or any of the program evaluations (see color key).

A number of the citing works to ATE products were by authors not affiliated with the four program evaluations or NSF (represented by red circles). The works from these non-affiliated sources primarily cited five ATE publications:

- ATE022: Owens, T. (2002). Dissemination: A key element of the ATE program. In A. R. Gullickson, F. P. Lawrenz, & N. Keiser (Eds.), *The ATE program: Issues for consideration: A monograph* (pp.26-44). Washington, DC: National Science Foundation.
- ATE028: Zinser, R., & Lawrenz, F. P. (2004). New roles to meet industry needs: A look at the Advanced Technological Education program. *Journal of Vocational Education Research, 29*(2), 85-100.
- ATE029: Lawrenz, F. P., Keiser, N., & LaVoie, B. (2003). Sustaining innovation in technological education. *Community College Review, 30*(4), 1-14.
- ATE030: Lawrenz, F. P., Keiser, N., & LaVoie, B. (2003). Evaluative site visits: A methodological review. *American Journal of Evaluation, 24*(3), 341-352.
- ATE052: Lawrenz, F. P., Keiser, N., & Lavoie, B. (2002). *A guide for planning and implementing site visits.* Kalamazoo, MI: Western Michigan University: The Evaluation Center.

The citing works were all publications or reports that addressed issues in project capacity building, workforce development, or evaluation. Figure S also shows the authorship of three citing works (yellow circles) was attributed to ATE projects, one attributed authorship to a CETP project (green circle), and one was authored by an NSF source (orange circle). Each of these authors was drawn to different ATE products, interestingly.

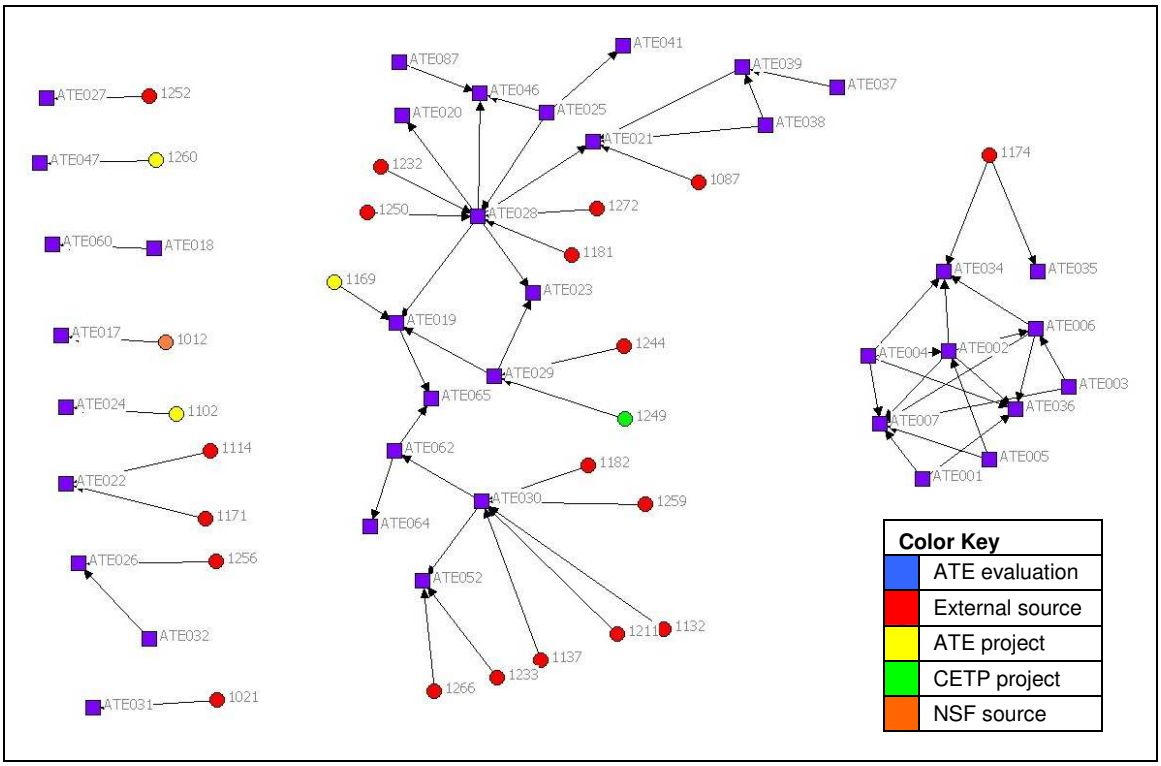

Figure 10. Citing authors of ATE products

ATE evaluation content area patterns. In Figure 11, the products and citations are all re-colored to represent the content areas that the publication addressed; a color key is provided with the figure. The ATE evaluation products are squares and other works are circles. As shown in the figure, the ATE evaluation products and citing works covered a wide variety of content areas, and there appears to be a link between the fields of the products and their citations. For example, products ATE030 [Lawrenz, F. P., Keiser, N.& LaVoie, B. (2003). Evaluative site visits: A methodological review] and ATE052 [Lawrenz, F. P., Keiser, N., & Lavoie, B. (2002). *A guide for planning and implementing site visits.*] were evaluation-related products, which were cited predominantly by evaluation-related citing works. Similarly, ATE028 [Zinser, R., & Lawrenz, F. P. (2004).

New roles to meet industry needs: A look at the Advanced Technological Education program.] addresses issues in workforce development and many of its citing works are also related to workforce development issues.

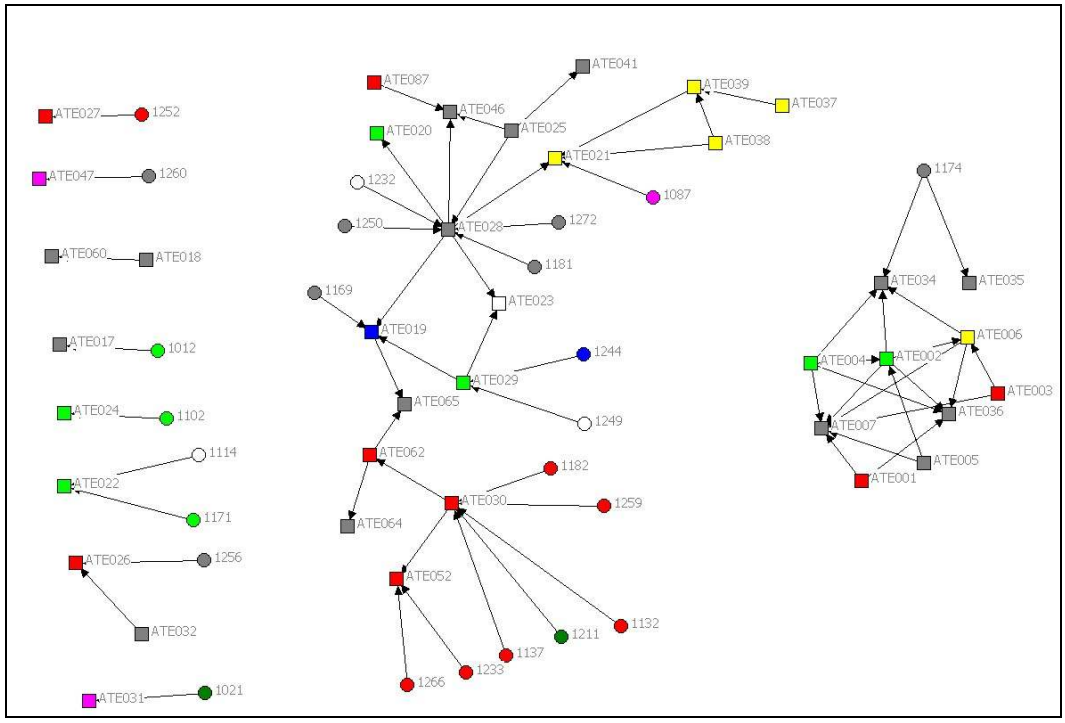

Figure 11. Fields of citing works and products

LSC core evaluation citation network. Figure 12 presents the citation network of LSC evaluation products (blue squares) and works authored by individuals other than LSC core evaluation staff or consultants (red circles).

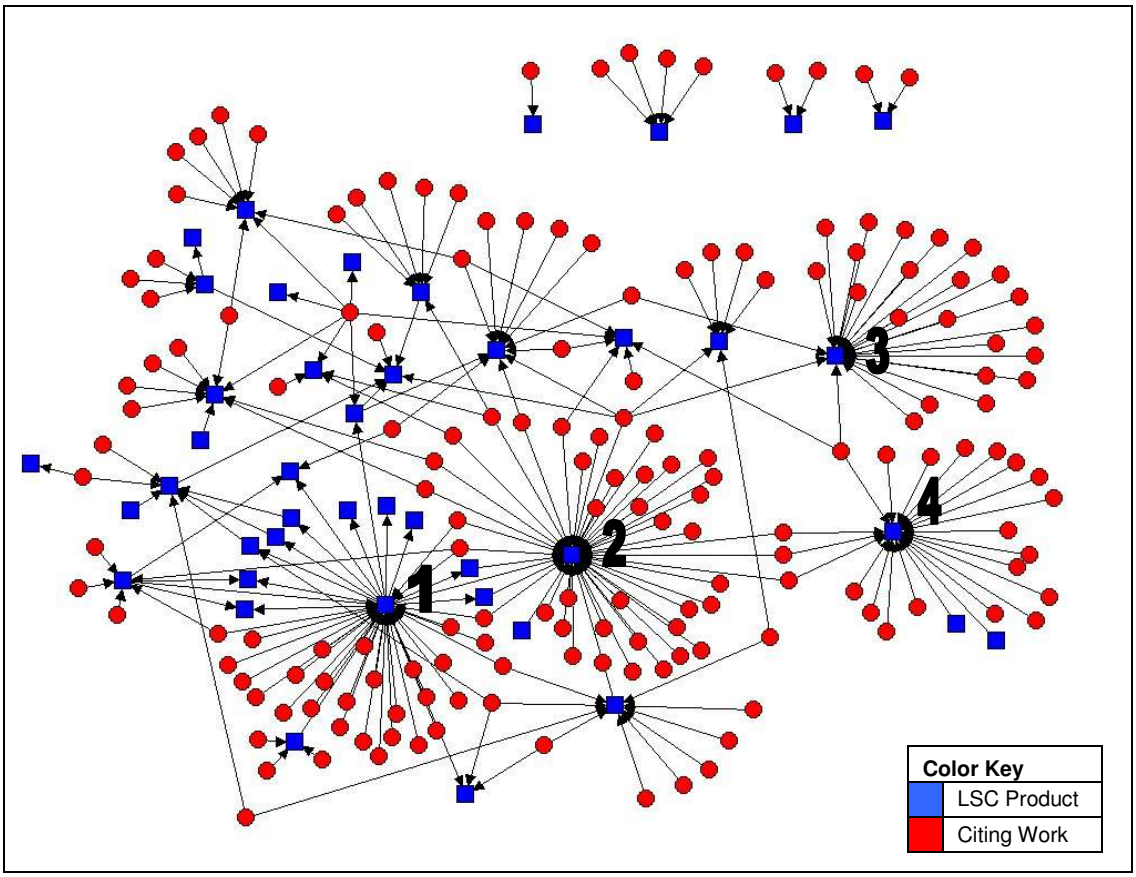

Figure 12. LSC evaluation products and citations

The LSC network map provides an interesting contrast to the ATE evaluation citation network. As shown above, there is widespread citing by non-LSC core authors, as well as evidence of internal citing between LSC evaluation products. Although there are a few LSC evaluation products that are cited by one or more citing works in isolation from the rest of the products, the majority of the products and citations are connected through networks of co-citation. There are relatively few instances of products that cite more than one or two LSC products, however. The majority of works cite one product only.

There are two central hubs in the network that show high levels of interconnectivity. The first, toward the left side of the figure and indicated by the number "1," is Banilower et al. (2006) "Lessons from a decade of mathematics and science reform: A capstone report for the local systemic change through teacher enhancement initiative." In addition to being cited 33 times, this product cites 14 other LSC evaluation products. The second hub, in the center of the figure and indicated by the number "2," is the "Local Systemic Change through Teacher Enhancement Classroom Observation Protocol," which is cited a total of 48 times by both internal and external sources. A close inspection of the citations to and from these two hub products shows a high degree of interconnectedness between these two products and other LSC products and citations. There are two additional highly-cited LSC evaluation products marked in the figure. The "Year-Three Cross Site Report" is marked with the number "3." The "Core Data Collection Manual" is marked with the number "4." While both of these products are highly-cited, they are less highly interconnected with other citations and products than the other two highly-cited works.

The LSC evaluation citation data also provide insights into relationships between the evaluation products and the authors and content areas of citing works. In Figure 13, the colors on the image have changed, although the LSC evaluation products remain blue squares and non-affiliated citing works are red. The remainder of the citing works have been colored to illustrate the affiliation of the authors to any of programs under consideration in the Beyond Evaluation Use project, to another NSF funded project, or to the NSF organization.

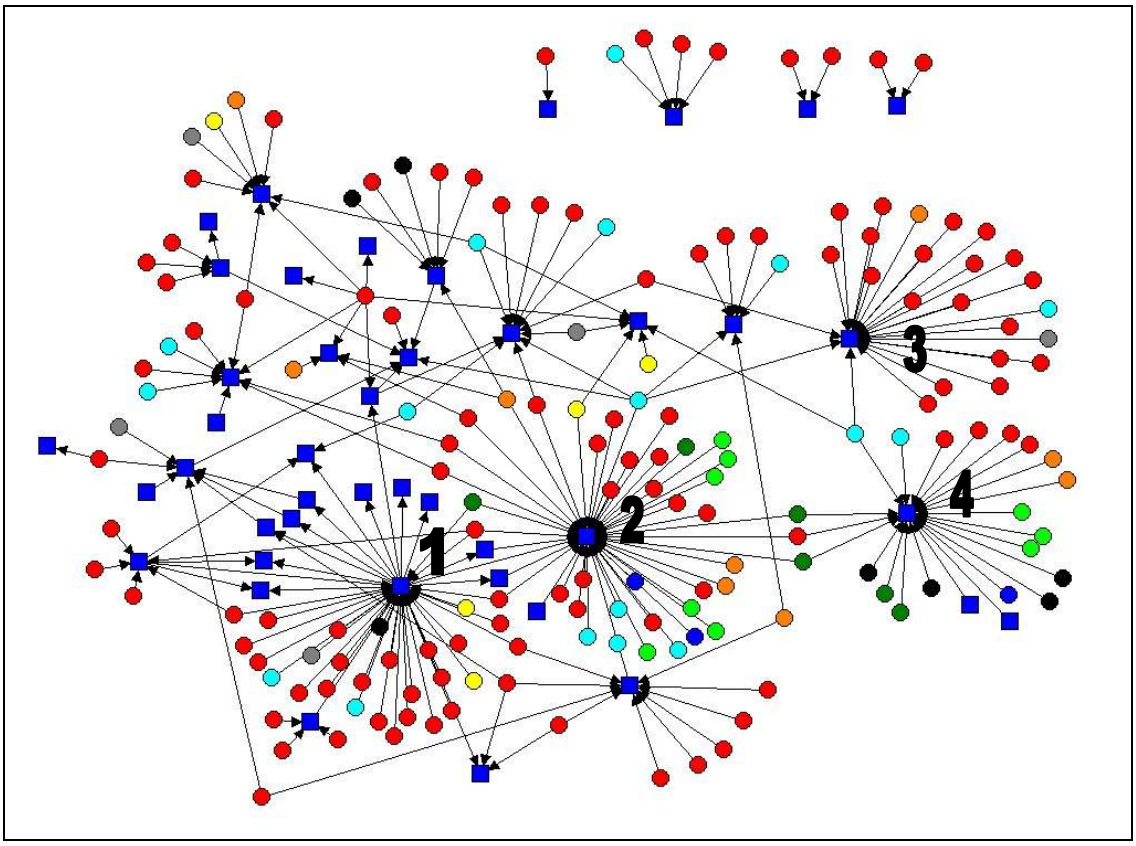

Figure 13. Authors of products and citing works

Examination of the author network reveals an interesting pattern. Among the four highly-cited products, the Lessons Learned article (#1) and the Year-Three Cross Site Report (#3) are predominately cited by non-affiliated sources. The two evaluation instruments have somewhat different citation patterns. The Classroom Observation Protocol (#2) is cited by both non-affiliated and sources connected to NSF programs and projects in some way, including several LSC projects, CETP projects, and other NSF projects. The Core Data Collection Manual (#4) is cited primarily by sources that are part of the NSF community.

LSC products' and citations' content areas. In Figure 14, the color coding scheme is changed again to describe the content areas within STEM education and evaluation that the LSC products and citations addressed.

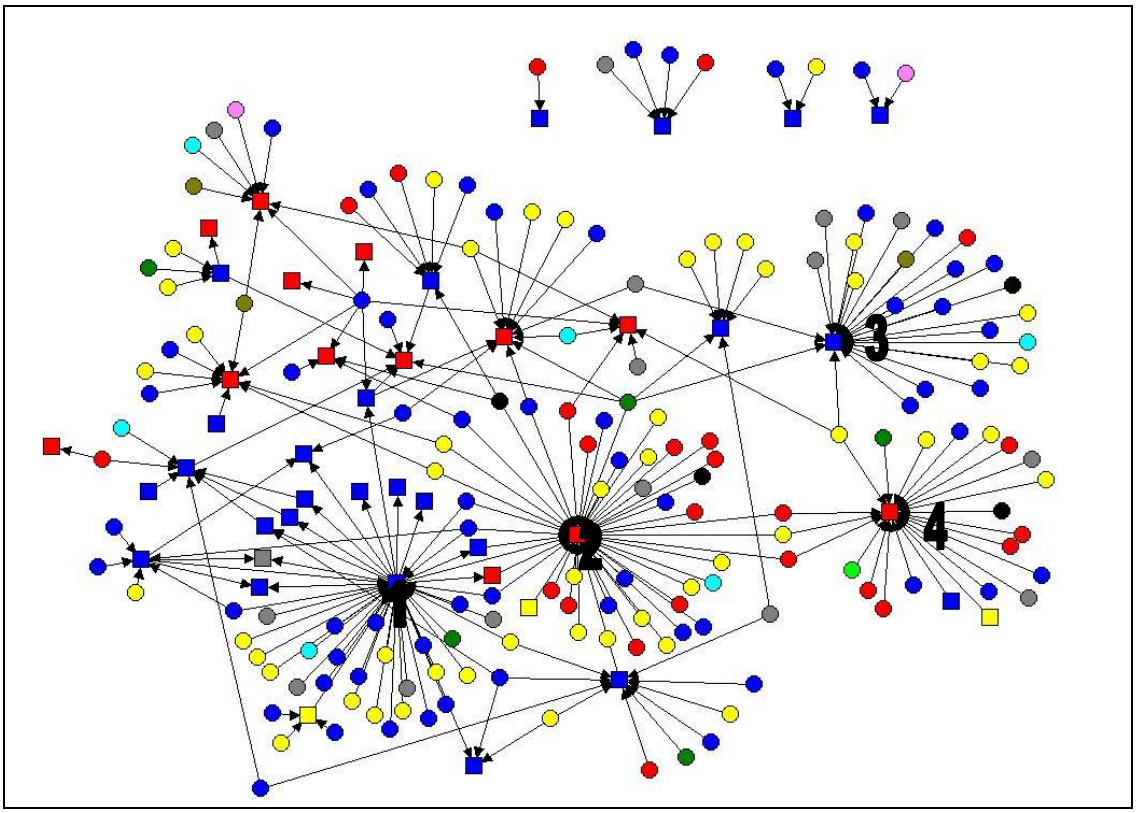

Figure 14. Content areas of products and citing works

Color Key			
	Professional Development		Teacher training
	Curriculum/Instruct/Assess.		Materials Development
	Evaluation		Project Capacity Building
	Partnerships/Systemic Reform		Workforce Development
	NSF		Other

As illustrated above, most of the evaluation products were cited by a variety of different content areas within STEM education. This is evident with three of the four highly cited products. The Classroom Observation Protocol (#2), Year-Three Cross Site Report (#3), and Core Data Collection Manual (#4) were all cited by a number of professional development-related works, but also reached a variety of other education-

related fields. The "Lessons Learned" article (#1), on the other hand, appears to be grounded in the professional development and curriculum, instruction, and assessment literature. Not only are the vast majority of the LSC products that it cited related to professional development, the majority of its citing works are from the professional development and curriculum fields, with a handful of citations from works addressing systemic reform.

CETP core evaluation citation network. Figure 15 presents the citation network of CETP core evaluation products (blue squares) and citing works (red circles). The figure highlights the impact of one of the CETP products (CETP012) – the CETP Classroom Observation Protocol. This one product accounts for 27 of the 42 total citations. The figure also shows that there are three citing works (i.e., 1205, 1139, and 1051) that cite both the classroom observation protocol and one other evaluation product.

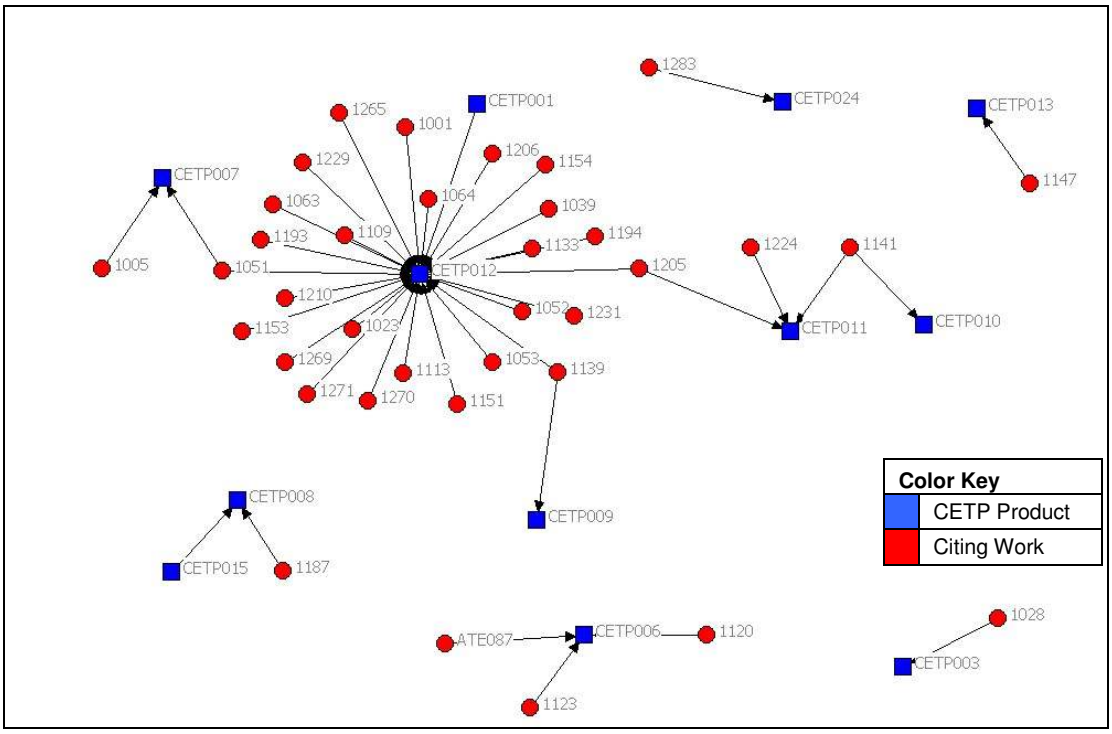

Figure 15. CETP core evaluation products and citations

CETP author patterns. In Figure 16, the color coding scheme is changed to illustrate the different authors of the external citing works. CETP core evaluation products are again represented as blue squares, however, other sources are colored according to the key. As illustrated in the figure, citations to the CETP Classroom Observation protocol were split between sources closely connected to the core evaluation (e.g., CETP projects or other NSF PIs or staff) and sources that were not affiliated with the core evaluation in any direct way. These patterns are similar to some seen in the LSC core evaluation network. The LSC and CETP core evaluations are both examples of evaluations that engaged a high proportion of project representatives in the development and/or training of evaluation instruments. It is not surprising that these instruments are then cited by projects and others within the NSF community. These patterns contrast

with the ATE evaluation citation network that has few citations from outside sources. The ATE evaluation was less participatory in nature.

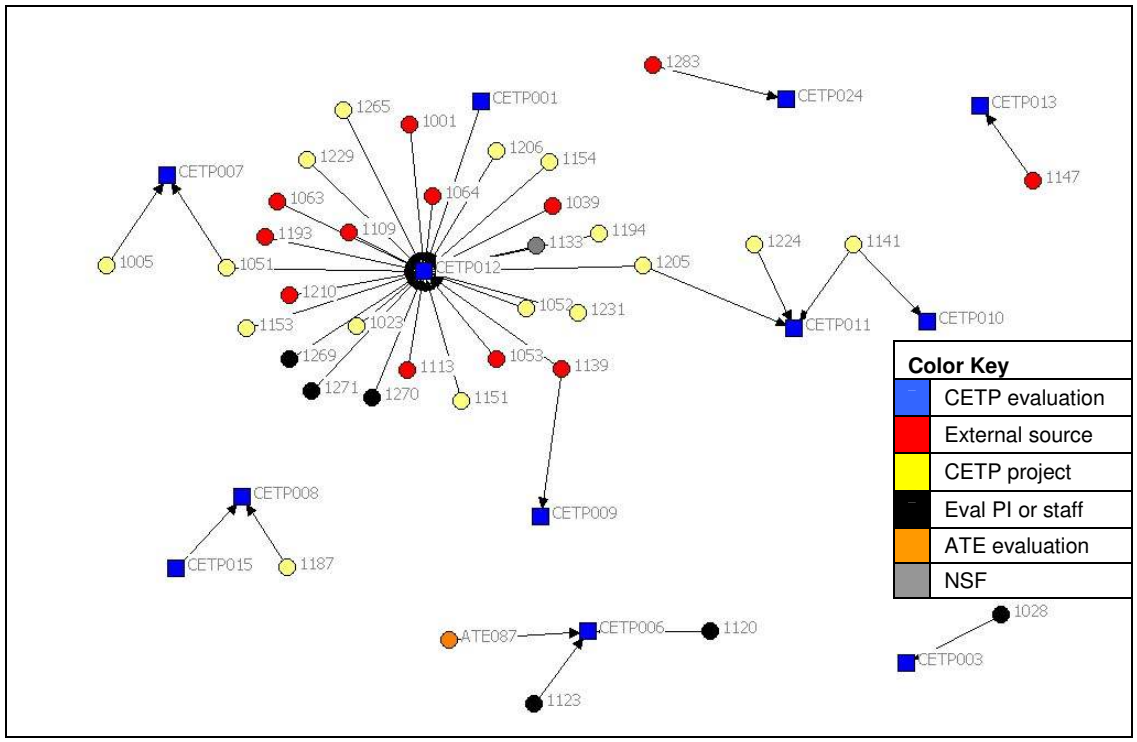

Figure 16. Citing authors of CETP core evaluation products

CETP content area patterns. In Figure 17, the products and citations are all re-colored to represent the content areas that the work addressed; a color key is provided with the figure. The CETP core evaluation products are squares and citing works are represented as circles. As shown in the figure, many of the cited evaluation products were evaluation-related in nature, including several instruments. Many of the citing works were also related to evaluation issues. The content areas of the works citing the CETP Classroom Observation protocol are somewhat varied, in that they address issues in all five of the content area categories represented among the CETP products. The works citing the remaining CETP products, however, are concentrated in areas of

evaluation and teacher pre-service. This suggests that while the Classroom Observation Protocol received interest in a number of areas, the rest of the CETP core evaluation products had somewhat limited reach in terms of the content areas that cited them. This is in contrast to the LSC evaluation, which reached a wider range of content areas within STEM education.

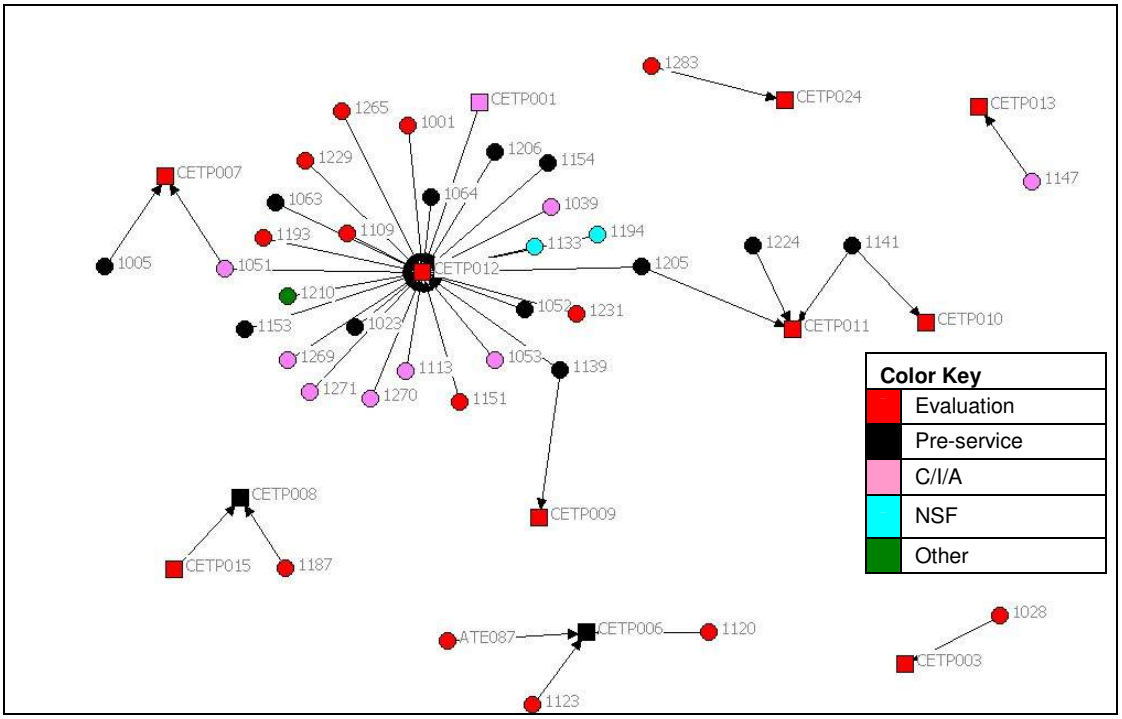

Figure 17. Content areas of citing works and products

The visual analysis of the networks of products and citations provides additional insights that extend the findings from the quantitative analysis. Visually inspecting the data illustrates differences among the four program evaluations in terms of degrees of self-citation, citation by affiliated projects or others in the NSF community, and reach into a variety of content areas within STEM education and evaluation. The network

analysis also highlights the importance of highly-cited product hubs and how these hubs are related to the overall citation structure.

Assessments of the Evaluation Primary Investigators

Together the quantitative and visual analysis of the citation data appears to provide useful distinctions in the patterns of citation between, among, and within the four program evaluations. The perceptions of the four evaluations' primary investigators regarding the representativeness and accuracy of the interpretations of the data were solicited to see if they found the citation analyses useful.

The primary investigator of the CETP core evaluation, Frances Lawrenz, felt that the citations were about what she expected, although she feels that the evaluation had a greater influence than that which the citations capture. In terms of the representativeness and accuracy of the citation data, she said, "I would say this is sort of what I would have expected. I think that some of my articles about CETP might have been more cited, but I guess it is right." She continued, "I think that the project and its products had more influence than can be seen here. There was lots of interpersonal contact and influence that isn't reflected here but I think this is representative." Lawrenz also commented that "most people seem interested in instruments, mostly observations," which is a conclusion highlighted in both the quantitative and visual analyses. She was discouraged, however, at the overall level of impact of the products, stating "I just think it is depressing that we all did so much work and there seems to be so little impact" (F. Lawrenz, personal communication, February 23, 2008).

Arlen Gullickson, PI of the ATE evaluation, found the analyses to be clear and helpful and in particular thought the graphics helped to "make sense of the references and show linkages." Gullickson reflected on the choice of the methodology and stressed that the choice of a particular strategy affects the types of use seen. As he explained, "Here use is viewed through the lens of writers so it gives no insight into the large distribution of information to ATE PIs who read and used materials… but likely never referenced it…. This is but one way of measuring use and others (number of printed copies, number of Web hits) provide additional perspectives" (A. Gullickson, personal communication, March 9, 2008).

Cathy Callow-Heusser, PI for the MSP-RETA project, stated that she was surprised at the low numbers of citations to the DIO cycle. She stated that she was able to locate an additional citation doing a quick search, and more importantly, from her perspective should have been cited more highly: "given the lip service it has been given among NSF program officers, its prominent position on several NSF web pages, and the number of NSF MSP folks who have contacted me about it, I would have suspected more references." She concluded that "given the feedback we've received from NSF program officers and MSP folks, I am not sure it captures influence. But I don't have hard evidence (other than e-mails and meeting conversations) or suggestions for other ways to measure influence." Callow-Heusser speculated that perhaps the citation databases are not adequately capturing the grey literature in which she believes the DIO cycle should have been cited more highly. She also suggested that perhaps there are differences in the ways individuals cite "how-to" works, such as the DIO cycle model and instruments such as those developed by the LSC and CETP evaluations. She thought that perhaps

instruments were cited more highly because people actually "use/apply" them (C. Callow-Heusser, personal communication, March 11, 2008).

Iris Weiss, PI for the LSC evaluation, shared some of the same concerns as Callow-Heusser in terms of the citations under-representing actual influence. As she wrote,

> I think the data underestimate the influence (but I have no way of knowing to what extent); for example, we received requests for the videos we had used in training evaluators from people who wanted to use the observation instruments to establish a common vision of effective classroom instruction/quality PD with teachers/teacher leaders/principals, etc., and those kinds of uses may not have wound up reported in the literature.

Although Weiss was concerned about underestimation, she did feel that the data were representative of her perception of the LSC's influence on the field. As she wrote, "the citation data seem about right to me." Like Gullickson, Weiss found the visual presentation interesting: "The network analysis is a very interesting way to present these data!" (I. Weiss, personal communication, March 17, 2008).

Assumption 8: Citation analysis is useful for understanding the influence of STEM program evaluations of different sizes.

All the preceding analyses were conducted on data gathered from four large-scale, multi-site STEM education program evaluations. As discussed in Chapter Three, the evaluations are diverse in several respects and represent a range of current practices within STEM education evaluations. First, the content areas of the four programs being

evaluated represent four key areas in STEM education (i.e., professional development, pre-service, workforce development, and partnerships). Second, the evaluations reflect a range of common NSF evaluation types including monitoring studies, program evaluations, and technical assistance projects providing tailored support to a variety of local projects. Third, the evaluations ranged widely in terms of the level with which local project PIs and evaluators were involved in the development of evaluation plans and instruments (from not at all involved to highly involved). Finally, the emphasis and means of disseminating evaluation findings ranged from dissemination being a mandated part of grant extensions for two projects to two projects that did not focus highly on formally disseminating findings to broad audiences.

Despite their differences, the four evaluations were similar in key ways which may limit the generalizability of the conclusions of this study. All four programs of the programs being evaluated were large-scale, multi-site national initiatives and three of the four evaluations were program-level evaluations of the local projects while one was an evaluation technical assistance project aimed at assisting the local projects. As stated in Chapter Three, while there are numerous examples of NSF-funded multi-site program evaluations, which these four evaluations represent well, there are also many single-site "project" evaluations and national or international "status" studies that these four evaluations do not represent. Because of the similarities in terms of scale of four evaluations in the sample, questions remain regarding the applicability of citation analysis to STEM evaluations of different sizes. To address this issue, rough citation counts were gathered for two NSF evaluations of different sizes: one is an example of smaller-scale, single-site "project" evaluations and the other is an example of an

international "status study" evaluation effort. By examining citation counts in these two contexts some tentative conclusions can be made regarding the usefulness of citation analysis to understanding the influence of program evaluations of different sizes.

The first example studied was the evaluation of the Wisconsin Academy Staff Development Initiatives Retention and Renewal Program (WASDI R2). The WASDI R2 evaluation was a formative and summative evaluation of a single-site teacher professional development program conducted for four-years (2003-2007) at a total cost of approximately $100,000. Approximately 30 evaluation reports were produced that provided formative and summative evaluation feedback to the program's primary investigators and NSF. One conference presentation was also made in conjunction with the evaluation. No evaluation instrument-related products were produced, nor were any journal articles published.

Citation data for the WASDI evaluation were collected by searching Web of Science, Google, and Google Scholar for references to WASDI evaluation products. Both Google Scholar and Google found one (the same) citation to the year two executive summary evaluation report: Blank, R.K., de las Alas, N., & Smith, C. (2007, February). *Analysis of the Quality of Professional Development Programs for Mathematics and Science Teachers: Findings from a Cross-State Study.* This study, commissioned by the Council of Chief State School Officers is a review and comparison of STEM professional development initiatives and their evaluations. The WASDI program and evaluation were included among 25 other similar professional development initiatives in the study. The Web of Science did not yield any citations.

The second example is the Trends in International Mathematics and Science Study (TIMSS). TIMSS is an international study of mathematics and science achievement conducted by the International Association for the Evaluation of Educational Achievement (IEA) four times since 1995. The 2003 study, the last completed study, collected standardized test data at the 4th and 8th grades, as well as survey data from teachers, students, and principles in 46 countries.

TIMSS was found to be the second most influential educational research study and by far the largest and most influential study in STEM education in a recent examination of influential educational studies conducted by the Editorial Projects in Education (EPE) Research Center (Swanson & Barlage, 2006). The EPE Research Center study used a two-stage survey methodology to produce a list of the top ten most influential studies conducted in education between 1996 and 2006 as determined by top educational policy experts. Then citation data from both scholarly sources and news sources were combined with the expert ratings to create a composite "influence index." Rather than trying to duplicate the EPE Research Center's efforts in collecting comprehensive citation data on TIMSS, their findings will be provided here as a point of comparison to the influence levels of the program evaluations and single-site evaluation.

The EPE Research Center study decided not to use Google Scholar as a way of gathering citation information, despite its acknowledged strength in capturing the grey literature. The researchers found that the Google Scholar was "insufficiently precise" (p. 6) to produce refined enough results to be used without investing high amounts of time and energy in manually collecting and cleaning data. Consequently, the search engines LexisNexis and EBSCO's Academic Search Premier were used to collect citation

information. LexisNexis was used to determine the number of references to studies in the news media while EBSCO was used to capture citations within peer-reviewed journals. The study found approximately 1,500 references to TIMSS within the news media and approximately 775 citations to TIMSS within peer-reviewed journals (p. 13).

A quick search on Google Scholar confirmed the high citation levels found by EBSCO. For example, searching Google Scholar for citations to the TIMSS report on middle school science achievement [Beaton, A.E, et al. (1996). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*] found over 650 citations alone, although these data were not cleaned and searched for duplicates. Regardless, this is only one of many reports, articles, and publications produced as part of the TIMSS study. This level of citations is clearly congruous with the EPE Research Center's finding that the TIMSS study was highly influential both by peer-review and citation data.

Assumption 9: The consequences of using citation analysis to measure STEM evaluation product influence are more beneficial than detrimental.

The final assumption to be examined regards the consequences of using citation analysis to measure STEM program evaluation impact. The three evaluation theorists were asked to comment on the possible positive and negative consequences of citation analysis to measure STEM evaluation influence and to judge whether the possible positive consequences outweigh any possible negative ones. Additionally, the evaluation

theorists were asked to comment on the overall validity and utility of using citation analysis for assessing STEM education evaluation influence.

Both Patton and Kirkhart stated that a positive consequence of citation analysis is that it provides a way of quantifying product influence. For example, as Kirkhart stated,

> The positive consequence is that it lays a quantitative foundation that describes the nature of information flowing from evaluation programs/projects to other intellectual undertakings through printed media. These data may provide accountability at the level of an individual researcher or a funded research program in demonstrating productivity in terms of output. The method largely maps dissemination from the end of an evaluation cycle forward. It focuses on products that then must be understood in the context of the intent of the funded research. It notices patterns of visibility over time and maps connections (K. Kirkhart, personal correspondence, March 30, 2008).

Alkin felt that citation analysis was helpful because it provides information about highly-used products and can inform scholars, policy makers, and funding agencies about productivity and influence levels of STEM evaluations (M. Alkin, personal correspondence, April 2, 2008). Additionally, Patton found the networking graphics "useful and powerful" while also finding the relative ease of access to citation data a strength of the method (M. Patton, personal correspondence, March 8, 2008).

Both Kirkhart and Patton were primarily concerned about using citation analysis to measure evaluation influence regarded its use as a stand-alone method. As Patton wrote, citation analysis

> doesn't strike me as a stand-alone method. The addition of case studies, as you're doing, is essential. If used as one method in conjunction with other indicators and sources of data, it's fine. If it is treated as a stand-alone or primary indicator, it is distorted.

Simiarly, Kirkhart commented that she "would advise caution in using this as a stand-alone and emphasize the importance of pairing this with other methodologies if one seeks to understand the scope of influence of a given evaluation or its products." Additionally, Kirkhart was concerned about using citation data counts, without an analysis of their content, which would lead to a limited understanding of influence. As she stated,

> The danger is that it represents a very narrow operationalization of influence (to the extent that actual influence is captured at all) and that the numbers alone cannot move us beyond a black box analysis of visibility. The contribution of the citations—whether used to support, refute, replicate or extend a prior argument—remain unknown, as are the impact and consequences of their use. I worry that the numbers stripped of context will underrepresent the influence to an extent that obscures the actual impact of the work.

Alkin's chief concern was that the results of citation analyses could be over-generalized leading to the possible negative consequence that program evaluations whose influence is hard to capture may be subjected to funding-cuts if decision makers do not use citation data carefully. In Alkin's mind, however, "the positive outweighs the negative."

In terms of the validity of using citation analysis as a method for understanding the influence of STEM education evaluation products, all three theorists felt that while

the method has some validity, caveats and cautions were warranted. Patton reemphasized his concerns about using citations as stand-alone indicators, finding them not useful in isolation from other measures of influence. Kirkhart concurred, stating that she thinks citation analysis

> is certainly a useful method for mapping the dissemination of products and the exchange of information among researchers—especially here, those working in the related areas. But I think that equating that with understanding influence is risky; it is but one piece of the larger influence puzzle.

She continues,

> Do I think that this method has potential for a partial mapping of the influence terrain? Absolutely! But taken as a criterion for *judging influence*, I have validity concerns; I think it underrepresents the construct, influence. So I would advise, 'Use with caution'."

Both Patton and Kikhart reacted strongly to the citation analysis report in which citation counts, without content, were reported and analyzed. Alkin also stressed that it is important to know whether the citation was made in a positive or negative context and, "even in instances where the citation was stated positively, we, of course, do not know how influential the idea was that was being cited." Both Patton and Kirkhart were also concerned about using citation analysis as a stand-alone method; however, both thought that the method had some usefulness in understanding a limited type of influence. Next, Chapter Five will discuss the strengths and limits of the methodology in further detail based on the findings of this validity study and connect the citation data to existing theories of evaluation influence.

CHAPTER FIVE

DISCUSSION AND IMPLICATIONS

This chapter discusses the study's findings and evaluates the evidence associated with each of the nine validity assumptions. Then, this chapter assesses the overall validity of using citation analysis for measuring STEM evaluation influence and examines limitations to the study. This chapter concludes with a discussion of the implications of the data collected for this study for extending existing theories of evaluation influence and outlines areas of future research.

Evaluation of the Validity Evidence

Assumption 1: Citation analysis is an established method for measuring the impact of STEM education evaluations or research efforts in related fields.

To assess this assumption a literature search was performed to locate citation analysis studies conducted within the fields of STEM, education, evaluation, and the union of these fields. Numerous studies using citation analysis methods for a variety of purposes were found within STEM fields. Using the ERIC database, 45 citation studies within education fields were found, although only one of these studies was related to STEM education, and none were related to a STEM education evaluation. A review of discussion within the field of evaluation regarding citation analysis methods was conducted as well. There appears to be some recent interest within the evaluation community in using citation analysis.

These findings suggest that citation analysis is used widely within STEM areas, less so in educational fields, and seldom within STEM education areas. The field of evaluation only recently began to pay attention to citation analysis as a methodology. No examples of the application of citation analysis for measuring program evaluation influence were found in the review of the literature, although the usage of citation analysis to measure the impact of individual researchers and research centers is widespread. The evaluation theorists all agreed that citations are a convention used with STEM education evaluation papers to give credit to influential ideas or methods. These findings suggest that while citation analysis is not an established methodology within the fields of education and evaluation, citations are used within the field, and the ubiquity of citation analysis studies within the STEM areas provides some support for the assertion that it is an established methodology. The usefulness of citation analysis methods for understanding STEM education evaluation influence is addressed in the following eight assumptions.

Assumption 2: The content of citations to STEM education evaluation products suggests they are used to give credit where credit is due or represent other indicators of influence.

This assumption was evaluated by conducting a review of the theoretical and empirical literature regarding the content of citations and conducting a content analysis of a random selection of 30 citations to the sample evaluation products. The literature review found that the citation analysis literature discussing the content of citations is somewhat divided regarding the assertion that citations are used to give credit where credit is due. Although the majority of opinions and empirical research on the subject

support the assertion that citations are indicators of influence, there are some dissenting voices that suggest citations are used in other ways and are not good indicators of influence. The review of existing content coding schemes found that no existing categorization scheme perfectly fits the application of assessing evaluation product influence. Several of the schemes were able to be combined and tailored to fit the evaluation context to provide a framework for the content analysis study.

The 30 randomly-selected citations were coded to one of eight content codes according to the context within which they were used in the original publications. The findings from the content analysis support the assertion that citations are used in ways that indicate that the STEM evaluation product had some influence on the citing author. The most frequent type of citation was a reference to an evaluation instrument or method used or considered to be used within another study. The second most common type of citation was a reference to the empirical findings of the evaluation. Together these two types of citations accounted for two-thirds of the total citation types. The remaining citations were divided among factual, informational, resource and other similar types of references. Only one of thirty-five citations was not consistent with the assumption that the citations captured in the study represent influence; this citation was inappropriately identified as a citation during the data collection process.

There are several implications from the citation content analysis. First, the finding that evaluation methods and instruments, as much if not more than results, are highly cited is not new, but remains important. Garfield in 1979 found that many methods papers tended to be cited highly and argued that this was an indicator of their "utility." As he wrote,

> People talk about citation counts being a measure of the "importance", or "impact" of scientific work, but those who are knowledgeable about the subject use these words in a very pragmatic sense: what they really are talking about is utility. A highly cited work is one that has been found to be useful by a relatively large number of people, or in a relatively large number of experiments. That is the reason why certain methods papers tend to be heavily cited. They describe methods that are frequently and widely used. (p. 363)

This study's data suggest that within the context of STEM education evaluations, many existing instruments are modified to fit what is seen as the unique context of a subsequent evaluation. This evidence of borrowing and modifying instruments raises questions about the validity of using instruments developed in one context in another. Nevertheless, these examples clearly demonstrate a high level of influence that the original program evaluations had on other evaluations and studies through the use of their instruments in some form.

Secondly, the remaining types of citations found through the content analysis provide evidence of the contributions of the four program evaluations to building knowledge within STEM education and related fields. There were many examples of citations to the program evaluations' findings, uses of the program evaluations to substantiate factual statements, and the inclusion of the program evaluations as "concept markers" representing a particular genre or type of study within the field. Additionally, the inclusion of program evaluation products within resource lists is another avenue through which the influence of these evaluations is spread. The implications of these findings for building a theory of evaluation influence will be discussed further later in

this chapter. In sum, the content analysis provides evidence to support Assumption Two as the vast majority of the citations to evaluation products were indicators of the use and/or influence of the products on the citing author.

Assumption 3: Citation databases exist that provide adequate coverage of the STEM education and evaluation fields.

The third assumption was evaluated first, by examining the coverage by Web of Science and Google Scholar of key journals within the fields of STEM education and evaluation and, second, by analyzing the actual types of citing works captured by the three databases. First, the review of the coverage of key journals in the fields found that Google Scholar comprehensively covers both the key research-oriented and practitioner-oriented journals in the fields. Web of Science, while a helpful database for collecting citations to papers within research-oriented, peer-reviewed journals within the STEM education and evaluation fields, is inadequate alone for assessing the broader influence on evaluation products on STEM education and evaluation practitioners as it does not cover a number of key practitioner-oriented or content-focused journals. Google was not able to be examined in this manner as it is not possible to assess its coverage.

Second, the types of citations to the sample evaluation products actually captured by the three databases found that Google and Google Scholar complement each other in terms of the types of sources in which they find citations. Google Scholar captures citations within the academic journal literature as well as references contained within evaluation reports and other types of grey literature. Google captures many of the more informal types of references contained within electronic resources and other types of

works while adding some references within reports and presentations. While Web of Science captures a handful of additional peer-reviewed journal citations, the number of unique citations it adds to the sample found in this study is relatively quite small.

The analyses of the coverage of the databases suggest that the three existing citation databases together gather a broad sample of the possible citations to the products. Using all three databases, citations in the academic literature appear to be adequately covered. It is not possible to precisely assess the extent to which the databases are adequately capturing influence within the grey literature, however, as it is impossible to construct a list of the grey literature against which to compare coverage. The data do suggest, however, that at least some citations within the grey literature and less formal sources are captured using Google and Google Scholar. The coverage evidence collected to evaluate this assumption, therefore, support the notion that the databases provide adequate coverage of academic literatures in the STEM education and evaluation fields, although these analyses do not provide empirical evidence to refute concerns that the databases may under-represent the grey literature. Additionally, it is important to note that the proportion of true existing citations the databases capture is not discernable from these data.

Assumption 4: The process of gathering STEM education evaluation product citation data can be conducted accurately.

The evidence is mixed concerning the accuracy with which the citation analysis process employed in this study is able to be conducted across repeated trials. This assumption was examined through a replication study in which the primary researcher

and one additional researcher attempted to replicate the results obtained from the original data collection process on a random selection of 25 evaluation products. The results of the replication study found that greater numbers of citations were obtained during the replication trials. However, a significant difference in the mean number of citations per products was found only between the data collected by the primary researcher during the original data collection period and the data found by the primary researcher during the replication study. The mean number of citations was statistically the same for the two researchers during their first data collection attempts, as well as the two researchers during the same point in time. The percent of agreement between the different data collection trials and across the three databases ranged from approximately 65% to 100% agreement. The lower agreement levels were related to finding more citations to the same products at the later time point of the replication study and to differences between the researchers conducting the studies.

Several factors come into play that could affect the accuracy of the data collected using citation analysis. First, the findings suggest that researcher experience is an important factor in the amount of data found using the databases. While the secondary researcher was trained in the primary researcher's search process, apparently there is a skill to constructing the best search terms, refining search results, and in capturing accurate data. While it is hard to assess the skill levels of each of the collectors, in terms of the amount of time it took for each researcher to conduct the replication study, there were large differences. It took the original researcher approximately 12 hours to collect citation data on the 25 products while it took the secondary researcher approximately 27 hours to do the same task.

Second, characteristics of the Google search engine itself may contribute to differences in citations found by different researchers. Google searches are "tailored" based on what Google knows about a particular user (W. Teitelman, personal communication, October 20, 2007). For example, when a user searches for "Jaguar", the results that are returned contain links to either cars or cats, but the order in which they are presented depends on the user. Moreover, these results change with increased use of the search engine, as Google "learns" a user's preferences based on which links he or she clicks on. Consequently, the results returned for a particular search may have been more accurate for the primary researcher during the replication study, as Google had already "learned" her search preferences during the original data collection period.

There are also two sources of error that affect the accuracy of the data. First, it is harder to find citations to certain types of evaluation products than to others. In particular, finding citations to instruments is difficult. Based on the "checking" completed as part of this study, it appears that some authors do not include references to instruments formally in their reference lists. An example is in the dissertation: Sciulli, J.A. (2004). *Teaching Science through Inquiry in K-5 Classrooms: Analysis of Change in Practice.* Doctoral dissertation: Duquesne University. The dissertation includes a letter from Iris Weiss, PI of the LSC program evaluation, stating that he can use her questionnaire if recognition is given. However, there is no reference to the instrument in the reference list – just a mention in the text itself and the inclusion of the letter. This more informal way of citing instruments makes it harder to capture them using the search tools, in particular, Web of Science and Google Scholar, which index reference lists.

More importantly, time is also a critical factor in collecting citation information: every day new works are being published with new reference lists, and at some regular (and unknown) interval citation search engines are being updated and expanded to include this additional content. Any study conducted in the future has the potential of finding a greater number of citations to the same works. While it is not possible to estimate exactly how much more information is available as time passes, it is reasonable to expect that the greater the time that has passed since citation data have been collected, the more the original sample of data collected under-represents the current "true" citation level. The most important conclusion from these assertions is, therefore, that any sample of citation data should be viewed as a cross-sectional snapshot of the true citations that existed at the time the data were collected.

In conclusion, while the consistency of the data appears to be fairly low using Google Scholar and Google, these search engines capture so much more relevant data than Web of Science that they are necessary to gather an adequate sample of citing works. The issues addressed above concerning reasons that different citations are found in repeated trials suggest that the data obtained from a citation analysis are affected by several variables. Citation analyses such as these, therefore, should be viewed cautiously as a time-bound, researcher-influenced sample of the true numbers of citations that exist to a product.

Assumption 5: Citation data can be transformed into meaningful indexes for comparing levels of STEM education evaluation product influence.

The fifth assumption was examined by conducting a review of the literature regarding citation indexes and an examination of the relative rankings of the program evaluations calculated from several of the most widely-used indexes. The literature review located several existing indexes that can be used to transform raw citation counts into measures of influence. Each of the indexes measures a particular aspect of influence or impact. Two of the most recently developed indexes, the *h* index and the *g* index, attempt to provide a metric appropriate for comparing the research impact of an individual or group. The *h* index controls for the length of a researcher's publication history, while the *g* index corrects for the *h* index's bias against a small number of highly influential works.

Next, a comparison of six indexes using the program evaluation citation information was conducted. The analyses found that while all the indexes showed the same evaluation as having had the greatest influence, the relative rankings of the other three evaluations varied based on the index. Overall, the *h* index appears to be most similar to the simple citation count in terms of the ranking and magnitude of differences between the influence levels of the four evaluations. The *g* index and citations per product indexes are similar in their rankings, but not in the magnitudes of difference between the evaluations. The highly-cited product index does not do a good job of distinguishing between the influence levels of the four evaluations. The agreement of these indexes with other measures of influence will be discussed in the examination of the next assumption.

Assumption 6: Citation indexes are related to other measures of STEM education evaluation influence.

This study's sixth assumption was evaluated by comparing the rankings of four citation indexes with a survey of individual perceptions within the STEM education evaluation field about the influence of the evaluations. Comparison of the citation indexes rankings to the data from the survey of project evaluators and PIs found similar patterns across the different measures. All six measures were consistent in their ranking of the most influential and least influential evaluations. The rankings of the middle two evaluations varied across the six measures. One measure found the two evaluations had the same levels of influence, and the remaining five measures were split three-two in judgments of the two evaluations' influence levels. Regardless of the particular ordering of the middle two evaluations, the measures of peer-judgment of the evaluations' influence levels appear to correspond to some degree with the rankings of the evaluations' influence levels according to the citation indexes.

The evaluation of this assumption should be viewed with some caution, however, as the peer-judgment measures that were used are less than ideal. First, the survey populations (project PIs and evaluators) and the population of interest in this study (e.g., the fields of STEM education and evaluation) are not the same. While the project PIs and evaluators are representatives of the field, they are closely connected to the evaluations and consequently may have different judgments as to their influence. Secondly, each project PI or evaluator only rated the influence of his or her particular program evaluation so differences in the relative rankings derived by comparing their mean scores are plausibly linked to factors other than actual influence perceptions. However, as no other

measures of evaluation influence have been developed to correlate with the citation rankings, the evidence collected for the evaluation of this assumption is as strong as possible at this time. The findings here, therefore, suggest a basic level of congruence between the citation indexes and other measures of influence.

Assumption 7: Citation analysis is useful for understanding differences in patterns of use and influence within and across STEM education program evaluations.

The seventh assumption was evaluated by conducting a multiple regression analysis regarding factors related to citation counts and a visual analysis of the network patterns of citing and cited works. Then, the four primary investigators who conducted the evaluations were asked to comment on the accuracy and representativeness of the data, as well as discuss if they found the data helpful for understanding their evaluation's influence.

The citation analysis findings from both the quantitative and visual analyses were found to be helpful in producing new understandings of the influence of the four evaluations. The differences in mean citations found among the product types, combined with the network graphics showing relationships between citing and cited works, authors, and content areas, raise interesting questions regarding the mechanisms and arenas in which the four evaluations were influential within the fields. The evaluation PIs cautioned, however, that while the patterns may be interesting and representative, from their perspective the citation data underestimate the absolute levels of influence of the evaluations. Implications of the findings from these analyses will be discussed in greater detail later in this chapter.

Assumption 8: Citation analysis is useful for understanding the influence of STEM program evaluations of different sizes.

The eighth assumption was evaluated by examining the raw citation counts of two evaluations of different sizes than the four primary evaluations studied. The assumption was tested on one example of a single site evaluation (WASDI-R2) and one example of a status study (TIMSS). It was found that the citation data vary according to expectations (e.g., only one citation to the single-site evaluation and thousands to the status study). That said, caution should be used when drawing conclusions as to the usefulness of citation analysis in small-scale and large-scale studies as only one example of each were used and only raw citation data were obtained.

The findings suggest that citation analysis is most applicable for evaluations that are well-funded, relatively large-scale initiatives. Single-site project evaluations are not intended to have a great deal of influence beyond the decision makers closely connected to the program, and few have the resources to devote to producing and disseminating works intended to have a great deal of influence. The fact that even one citation to a WASDI R2 evaluation product was found is surprising, given that the reports were not publically available. To obtain the report, the CCSSO researchers had to send a personal request to the program evaluation team. In terms of the TIMSS study, the massive number of citations would make it extremely difficult to do the data coding necessary to do the types of quantitative and visual analyses that help to make meaning out of citation data. While the citation numbers are useful in gauging the relative influence of a study,

such as done in the EPE Research Center study (Swanson & Barlage, 2006), this type of application is an expensive and time-consuming undertaking.

While this discussion suggests that citation analysis may be most applicable at the level of well-funded, multi-site program evaluations, such as the four multi-site program evaluations examined in this study, there is one additional limitation that should be mentioned. All of the six examples examined in this study were of evaluations funded by the National Science Foundation. While NSF is arguably the greatest funder of STEM evaluations in the country, there are other organizations such as the Department of Education and private foundations that also play a role in supporting evaluations of STEM programs. This study does not attempt to examine examples of program evaluations from these sources, but leaves the question for future replication studies to determine further boundaries of generalization.

Assumption 9: The consequences of using citation analysis to measure STEM evaluation product influence are more beneficial than detrimental.

The evaluation theorists were in agreement that citation analysis provides a way of quantifying evaluation product dissemination and use patterns and is useful as one possible way to map the terrain of evaluation influence. They cautioned, however, about using citation analysis as a stand-alone method for judging evaluation influence and stressed the importance of understanding the content of citations before drawing conclusions about whether citations are measuring impact. The theorists did not explicitly weigh the positive and negative consequences to make a judgment about the use of citation analysis to measure STEM education evaluation product influence.

Kirkhart's urge to use citation analysis but use it "with caution" is likely indicative of their overall recommendation about the approach.

Overall Evaluation of the Validity Evidence

While having limitations, which are discussed below, this study presents several pieces of evidence that support the interpretive argument guiding this validity study:

> *Citations are interpreted as indicators of the impact of STEM education evaluation products, and, as such, citation analysis has utility as a method for measuring the influence of program evaluations on the STEM education and evaluation fields.*

The data collected for this validation study suggest that citations are one possible indicator of STEM education evaluation product impact and that citation analysis methods do provide data to help understand, to a limited extent, the influence of the evaluations on the fields of STEM education and evaluation. The evidence presented in the evaluations of the assumptions suggests that citation analysis methods can be helpful, within certain contexts, for gaining knowledge about the patterns of influence of STEM program evaluations. There is high usage of citation analysis within STEM areas and in the application of studying research centers, and some evidence of citation analysis methods being applied in educational settings. The extension of the method to STEM education evaluations is a reasonable translation.

The content of citations supports the assertion that citations are measures of influence or impact. It appears that existing citation databases cover the academic literature adequately and do capture at least some proportion of the grey literature and

less formal sources. While the process of collecting citation information will necessarily result in somewhat different results across different timeframes and researchers, particularly using Google and Google Scholar, this type of "error" is unavoidable and understandable. The impact of researcher experience suggests the importance of a full understanding and that training in the methodology is important for collecting as accurate data as possible.

Once the evaluation product citation data are collected, there are ways of computing influence indexes that can be used to compare the relative results across program evaluations. In particular, the use of network graphics appears to be helpful in illustrating patterns and relationships between citing and cited works to assess areas of influence on the fields. As cautioned in Chapter Three, however, validity arguments cannot be "proven," but rather only supported by evidence to make a case for drawing appropriate inferences from the instrument or method. While the assumptions discussed above appear to be supported to a limited extent by the evidence collected in this study, there are several aspects of this study that are weak and warrant further investigation.

Limitations. The evidence used to judge two of the assumptions (Assumptions Six and Eight) is weak. While the analyses related to assumption six found congruence between citation indexes and peer-judgments, correlations with stronger measures of influence, when developed, will strengthen this assertion. Another weak part of the study is the examination of the usefulness of citation analysis for evaluating the influence of small-scale and very large-scale evaluation efforts (Assumption Eight). As discussed previously, the four NSF multi-site evaluation initiatives used as the sample in this study, while arguably representative of that type and scale of evaluation, are not representative

of all NSF evaluations, all STEM evaluations, and certainly not of evaluations in general. Therefore caution should be used when trying to generalize the findings of this study to other types of evaluations conducted in other contexts. Additionally, this study highlights the time-bound and researcher-bound nature of the data obtained while conducting a citation analysis. As cautioned above, citation data are only one snapshot in time of an evaluation's influence, and the citations found in any one sample, while arguably fairly representative, are an underestimate of true citation levels.

The greatest concern about using citation analysis as a method for measuring evaluation influence is that of construct-underrepresentation. While the citation data provide one picture of influence arising from an evaluation's products, as stressed by the evaluation theorists, citation analysis is not useful as a stand-alone method. Citations are only one among many possible measures of one limited type of influence arising from the dissemination of evaluation products. Citation data do not appear to be useful for exactly quantifying the actual level of influence of any one evaluation. Additionally, the examination of the content of citations is critical. Without understanding the content of the citations, judgments cannot be made about whether citations are actually measuring influence. Consequently, it is important to stress that citations are only one measure of one possible influence arising from an evaluation and are limited and should be interpreted as such. Citation analysis as a method, therefore, appears to be most useful in comparing influence types and levels across products produced by similar STEM evaluations at a particular point in time. Keeping these limitations in mind, some tentative implications for evaluation theory and practice that can be drawn from the data found for this study are discussed next.

Implications for Evaluation Theory and Practice

This paper proposed and evaluated the usefulness of one methodology for measuring one way in which evaluations have influence – through their products. Many evaluation theorists have called for greater research into evaluation practices to develop both prescriptive and descriptive theories (Henry & Mark, 2003b; Scriven, 2007). This paper presents a method that can provide useful information for obtaining greater understanding of the nature and mechanisms of evaluation influence. While it was not the express purpose of this study to build on existing evaluation influence theory, the data collected and analyzed herein provide support for and suggest modifications of existing influence models. These implications would benefit from future research and may provide useful considerations for practicing evaluators.

First, using the logic model framework suggested by Henry and Mark (2003a), the data collected in this study provide a more nuanced understanding of the pathway to influence that evaluation products can have. Figure 18 is the logic model presented in Chapter One:

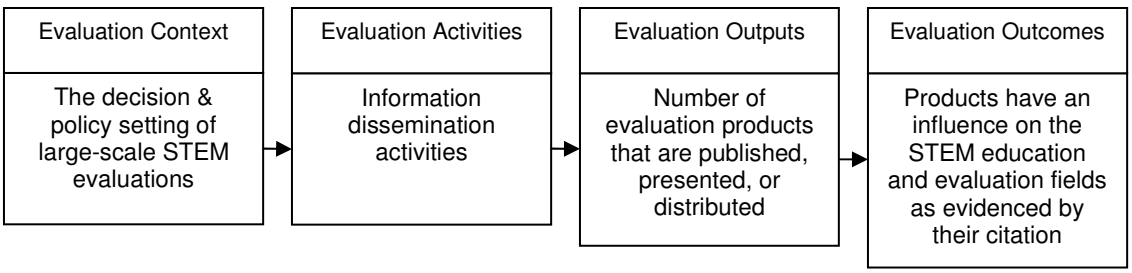

Figure 18. Logic Model of Evaluation Influence in Operation in this Study

The data from this study help to demystify what happens between the evaluation outputs and outcomes. The quantitative analysis of the citation data found that the type of evaluation product is important in predicting the amount of its influence, in terms of citations. The fact that evaluation instruments were cited more highly than other forms of products has potential implications for both theory and practice. This finding suggests that the development and dissemination of evaluation instruments is one way to increase the use and influence of evaluation efforts. It would be possible to test this finding in other contexts to see if it holds true or is unique to STEM education evaluation. In terms of implications for practice, evaluators should consider instrument development as an important activity, not only to meet the needs of one specific evaluation, but also as a possible service to the field. Making instruments available to others is one important step evaluators can take to ensure instruments' use beyond the scope of the particular evaluation for which they were developed. The content analysis findings show that instruments are often modified to fit new contexts, an issue that may affect the validity of their use and is important for the field to study as well.

The visual analysis highlighted additional patterns within the citation data. Comparing the network patterns of the four evaluations was helpful for contrasting levels of inter and intra-evaluation citation. A unique picture of influence within multi-site evaluation was illustrated through the connections. Citing works were often authored by individuals connected to the evaluation (both program evaluation team members and local project-level PIs and evaluators) in several examples. One example showed a much higher level of influence on the field in terms of citations from individuals not connected

to the evaluation in any discernable way. These patterns would be possible to study in different contexts and with evaluations of different sizes.

One additional possible implication for practice, if born out by further research, is that evaluations that are intended to have an influence on their content fields should consider incorporating participatory designs to engage representatives from the field in their work. This is one possible explanation of the relatively high citation levels for the CETP core evaluation compared to the ATE program evaluation. The CETP core evaluation, unlike the ATE evaluation, engaged local project evaluators and PIs in the development of evaluation plans and instruments. The use and influence of the CETP evaluation products, in terms of citation activity, by individuals connected to the evaluation is evident.

Secondly, the data from this study suggest two modifications to existing models of evaluation influence, namely the models proposed by Alkin and Taut (2003) and Kirkhart (2000). Alkin and Taut modified Kirkhart's integrated theory of influence model by adding a dimension of awareness to Kirkhart's existing three dimensions: source, time, and intention. The Alkin and Taut model identified three levels of awareness: aware/intended, aware/unintended, and unaware/unintended. Alkin and Taut also distinguished between evaluation uses, which are impacts that are aware/intended or aware/unintended and are immediate or shortly follow an evaluation. Evaluation influence they define to be those impacts that are unaware/unintended and/or arise after an evaluation's conclusion.

This study's data found examples of a fourth type of awareness in addition to Alkin and Taut's three types of awareness. This fourth type is unaware/intended

influences. For example, two of the evaluations were intended to disseminate results to the field, but were unaware as to the impact of their dissemination efforts. The current Alkin and Taut cube does not allow a space for this type of influence. Second, both the Alkin and Taut and Kirkhart frameworks dichotomize the source of evaluation influence as being either process- or results-related. This study has found that evaluation instruments are a third source of influence. While instruments are developed out of the evaluation process, they also have an existence and an influence on individuals not connected in any way to the evaluation. Furthermore, it is not the results that are found using the instruments that are a source of influence either. Therefore it would be inappropriate to categorize instruments into either the process or results dimensions in the current frameworks. Instead, a third category should be added to the source side of the cube.

These additions broaden the picture of dimensions of evaluation influence, suggesting that evaluation influence is best represented as a cube-like figure that is now 4x3x3 rather than Kirkhart's original 2x2x3 figure. Figure 19 presents a representation of this newly proposed cube of influence:

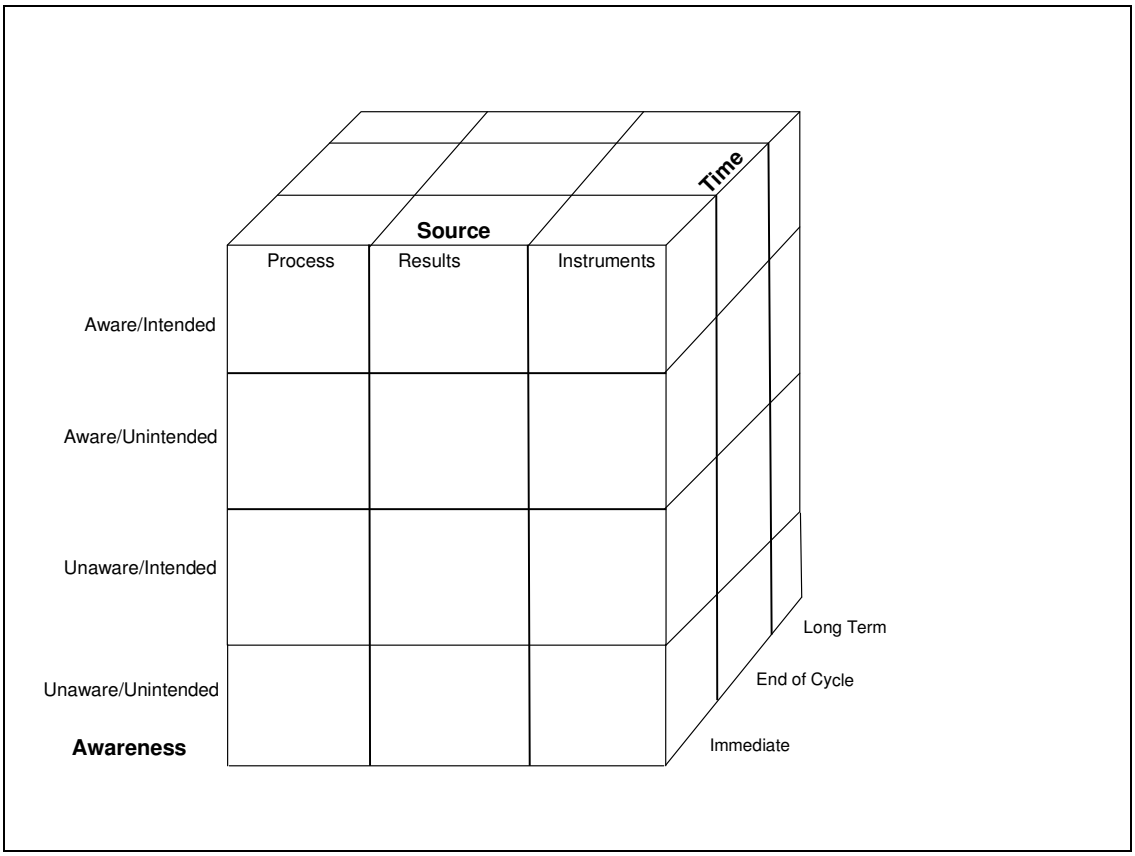

Figure 19. Newly proposed influence model (adapted from Alkin & Taut, 2003 and Kirkhart, 2000)

This new representation of evaluation influence challenges Alkin and Taut's assertion that influence is by its definition unintended in nature and therefore is not under the control of the evaluator. As they state,

> influences of evaluation are undoubtedly of importance, but they are unintended and cannot be addressed until after they have occurred. Practicing evaluators have to do their best in actively ensuring and promoting evaluation use, while at the same time noting evaluation influences that might occur but which are outside of their sphere of action. (2003, p. 10)

The data from this study suggest that influence, i.e., having an impact on individuals not connected directly to an evaluation or at some time later than the actual conduct of the evaluation, is an important intended outcome of large-scale multi-site evaluations. This is different from traditional conceptualizations of conceptual use of evaluation findings. Conceptual use is used to describe the impact of evaluation findings on decision makers – the "enlightenment" function described by Weiss (Weiss, 1980). In fact, NSF specifically extended funding in order to have evaluations influence the fields through the dissemination of evaluation methods and findings. Large-scale evaluations of this nature often have as their key stakeholders members of particular fields, communities, or even citizens of particular nations, rather than just decision makers at NSF or other funding agencies. These types of amorphous stakeholder groups, while important intended beneficiaries of the evaluations, make assessing the impact of an evaluation initiative difficult. Further investigation of other instances of intended evaluation influence may help to expand understanding of how evaluators can accomplish these goals.

Third and finally, this study's data provide support for Weiss, Murphy-Graham, and Birkland's (2005) assertions that the traditional three types of evaluation use, namely instrumental, political or symbolic, and conceptual, hold true for evaluation influences that may occur at some time and distance from the actual evaluation or with individuals not connected to the evaluation. The content analysis of the citations found examples of all three of these types of influence. For example, Weiss et al. argue that instrumental use, while having a "suggestion of immediacy" (p.13) is not tied to decisions made in the short run. An example of a type of instrumental influence found in the citation study is

the use of the LSC-developed Classroom Observation Protocols in other evaluations of professional development programs. This type of influence does not arise out of the findings or the process but rather the instrument (a third source of influence, as argued above). Additionally, the decision to use the instrument was not made by the decision makers involved in the original evaluation, but by a different group of decision makers connected to a subsequent evaluation.

Similar examples arose from the citation content analysis that fit Weiss et al.'s descriptions of conceptual influence and symbolic influence. Examples of conceptual influence include citations to evaluation findings by individuals with no affiliation to the evaluation. Symbolic influences were evident when authors used citations to evaluation products to support their own factual assertions, arguably to persuade readers to accept their propositions. This study also found examples of the new type of evaluation use found by Weiss et al. in their study of the influence of the D.A.R.E. evaluation, that of "imposed use." Imposed use is when the use of evaluations are imposed by a superordinate body. One of the citations analyzed by the content analysis showed that the National Science Foundation was mandating the use of one of the evaluation instruments developed for the LSC core evaluation – an example of imposed use of an evaluation instrument.

Future Research

This study suggests that citation analysis is, to some limited extent, a useful method for studying the influence of STEM education evaluations. Some of the analyses conducted in this study were weak, however, and some of the findings regarding the

validity of the method were mixed. Additional examination of the validity of citation analysis methods to draw inferences about influence in this and other contexts is therefore warranted and welcomed. The data collected for this study also suggest both support for, and challenges to, current theories of evaluation influence. Additional research on these topics and more attention to the implications of these findings for evaluation practice would benefit the field.

Conclusion

In conclusion, this study developed and evaluated a method for gathering data about evaluation influence. The method examined here, citation analysis, represents only one of many pathways to influence. Citation analysis appears to be best suited for conducting comparative research on the influence patterns of fairly large-scale evaluation efforts. Further work to validate methods for examining other pathways to influence and examinations of the application of citation analysis within contexts other than STEM education would be helpful. Evaluation funders, researchers, and practitioners may benefit from the development of theories and practices regarding the many ways in which evaluations can have influence.

Appendix A: Descriptive analysis of evaluation product types, fields, and content areas and category definitions

Evaluation Product Types

Each of the 245 evaluation products produced by the four evaluations was coded by type according to the following categories:

- Instrument/Tools: Evaluation instruments (observation protocols, surveys, etc.); instrument manuals, online evaluation tools (logic model tool, resource database);
- Presentations: Conference presentations, presentations to NSF or other audiences;
- Publications: Journal articles, books, book chapters, published monographs, dissertations, newsletters;
- Reports: Evaluation reports, briefing papers, fact sheets.

Evaluation reports accounted for almost 50% of the products produced by the four program evaluations in aggregate, however within each evaluation the distribution varied. Over half of the products the ATE evaluation produced were evaluation reports (n=52), while presentations (n=27) and publications (n=19) comprise the remaining half. ATE did not publish any evaluation instruments or tools. The CETP core evaluation produced almost equal numbers of instruments (n=5), reports (n=5), and publications (n=4) and a number of presentations as well (n=10). The LSC core evaluation produced 42 evaluation reports and 33 publications. The remaining products produced by LSC were instruments (n=13) and presentations (n=10). The MSP-RETA focused on presentations (n=18). The remaining quarter of MSP-RETA products were reports (n=4), instruments (n=2), and publications (n=1).

Table 22. Evaluation product types

Program Evaluation		Evaluation Product Type: Instrument/Tool	Presentation	Publication	Report	Total
ATE	N	0	27	19	52	98
	%	.0%	27.6%	19.4%	53.1%	100.0%
CETP	N	5	10	4	5	24
	%	20.8%	41.7%	16.7%	20.8%	100.0%
LSC	N	13	10	33	42	98
	%	13.3%	10.2%	33.7%	42.9%	100.0%
MSP-RETA	N	2	18	1	4	25
	%	8.0%	72.0%	4.0%	16.0%	100.0%
Multiple	N	0	0	1	0	1
	%	.0%	.0%	100.0%	.0%	100.0%
Total	N	20	65	58	103	246
	%	8.1%	26.4%	23.6%	41.9%	100.0%

Evaluation Product Fields

Products were coded as to their field and content area. The accuracy of the coding process was checked by recoding a random sample of 25 products. The percentage of intracoder agreement for the field codes was 88% and 100% for the content area codes. Some mistakes in the original coding were identified and corrected prior to running the analyses. Product fields were broadly defined according to the following categories:

- Education–general: education areas/topics outside of STEM education;
- Evaluation-general: evaluation areas/issues outside of STEM evaluation;
- STEM education/research: general topics or research related to STEM education;
- STEM evaluation: topics or findings connected to a specific STEM evaluation;

As shown in Table 23, as was expected, the vast majority of evaluation products were related to the field of STEM evaluation, meaning that they were evaluation reports or articles/presentations related to one of the four specific program evaluations. The ATE

evaluation produced a large proportion of general evaluation related products and the only general education product among the group. The CETP core evaluation, in addition to STEM evaluation products, produced a couple of general evaluation and STEM education/research related products. The LSC core evaluation had a number of STEM education/research type products in addition to STEM evaluation products. MSP-RETA had a few general evaluation products but no general or STEM education products.

Table 23. Fields of evaluation products

		Fields				
Program Evaluation		Education - general	Evaluation - general	STEM education/ research	STEM evaluation	Total
ATE	N	1	18	3	76	98
	%	1.0%	18.4%	3.1%	77.6%	100.0%
CETP	N	0	2	2	20	24
	%	.0%	8.3%	8.3%	83.3%	100.0%
LSC	N	0	1	14	83	98
	%	.0%	1.0%	14.3%	84.7%	100.0%
MSP-RETA	N	0	5	0	20	25
	%	.0%	20.0%	.0%	80.0%	100.0%
Total	N	1	26	19	199	245
	%	.4%	10.6%	7.8%	81.2%	100.0%

Product Content Areas

Products were coded as to their content areas within education or evaluation according to the following:

- Professional development: teacher training during service including the effects of PD on student achievement, etc.;
- Workforce development: high tech, STEM workforce training/development;
- Partnerships/systemic reform: university, school, community partnerships and systemic reform initiatives;

- Curriculum/instruction/assessment: related to classroom teaching: (i.e., instruction methods, curriculum, pedagogy, student assessment, standards-based instruction, math/science reform);

- Evaluation: evaluation-related topics such as multi-site evaluation theory, evaluation instrument protocols/manuals, other general evaluation theory or knowledge;

- Materials development: development of instructional materials;

- Project capacity building: issues related to program improvement, sustainability, advisory committee development, dissemination, replication;

- Other: other content areas.

The content areas of the evaluation products varied in expected ways according to the focus of each of the programs. ATE, where the primary goal of the program is to develop a high-tech workforce, had about a third of its products related to that topic (n=32). ATE also produced a high number of evaluation-related products (n=38). There are also a number of project capacity building, workforce development and professional development products. The CETP evaluation products were predominantly related to evaluation topics and instruments (n=19). A few products specifically related to issues in pre-service teacher training and curriculum/instruction/assessment were also produced. The LSC core evaluation predominantly produced products related to issues in professional development (n=72), also reflecting the emphasis of the LSC program. The MSP-RETA project produced a number of evaluation-related products (n=19) and products addressing issues in partnerships (n=6).

Table 24. Evaluation product content areas

Program evaluation		Content Areas: Teacher training/ Pre-service	Prof. Dev.	Work-force Dev.	Partner-ships/ systemic reform	Curr./ Instruct. / Assess.	Eval-uation	Mater-ials Dev.	Project Capacity Building	Total
ATE	N	0	5	32	1	2	38	6	14	98
	%	.0%	5.1%	32.7%	1.0%	2.0%	38.8%	6.1%	14.3%	100.0%
CETP	N	3	0	0	0	2	19	0	0	24
	%	12.5%	.0%	.0%	.0%	8.3%	79.2%	.0%	.0%	100.0%
LSC	N	1	72	0	3	5	17	0	0	98
	%	1.0%	73.5%	.0%	3.1%	5.1%	17.3%	.0%	.0%	100.0%
MSP-RETA	N	0	0	0	6	0	19	0	0	25
	%	.0%	.0%	.0%	24.0%	.0%	76.0%	.0%	.0%	100.0%
Total	N	4	77	32	10	9	93	6	14	245
	%	1.6%	31.4%	13.1%	4.1%	3.7%	38.0%	2.4%	5.7%	100.0%

References

Alkin, M. C. (1980). Naturalistic study of evaluation utilization. *New Directions for Program Evaluation, 5*, 19-28.

Alkin, M. C. (1985). *A guide for evaluation decision makers*. Beverly Hills: Sage Publications.

Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations: Does evaluation make a difference?*. Beverly Hills, CA: Sage Publications.

Alkin, M., & Taut, S. (2003). Unbundling evaluation use. *Studies in Educational Evaluation, 29*(1), 1-12.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2, part 2).

Ayers, T. D. (1987). Stakeholders as partners in evaluation: A stakeholder-collaborative approach. *Evaluation and Program Planning, 10*(3), 263-271.

Bamberger, M. (2004). Influential evaluations: Evaluations that improved performance and impacts of development programs. *Washington, DC: Operations Evaluation Department, the World Bank.*

Beyer, J. M., & Trice, H. M. (1982). The utilization process: A conceptual framework and synthesis of empirical findings. *Administrative Science Quarterly, 27*(4), 591-622.

Birkeland, S., Murphy-Graham, E., & Weiss, C. H. (2005). Good reasons for ignoring good evaluation: The case of the drug abuse resistance education (D.A.R.E.) program. *Evaluation and Program Planning, 28*(3), 247-256.

Borgman, C. L. (1990). Editor's introduction. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 10-27). Newbury Park: Sage Publications.

Bornmann, L., & Daniel, H. D. (2006). Selecting scientific excellence through committee peer review: A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics, 68*(3), 427-440.

Bornmann, L., & Daniel, H. D. (2007). What do we know about the h index. *Journal of the American Society for Information Science and Technology, 58*(9), 1381–1385.

Braskamp, L. A., Brown, R. D., & Newman, D. L. (1978, Summer). The credibility of a local educational program evaluation report: Author source and client audience characteristics. *American Educational Research Journal, 15*(3), 441-450.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (Fourth ed., pp. 1-16). Westport, CT: American Council on Education/Praeger.

Brooks, T. A. (1999). Core journals in the rapidly changing research front of "superconductivity." In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 235-247). Newbury Park: Sage Publications.

Brown, L. D., & Gardner, J. C. (1985). Applying citation analysis to evaluate the research contributions of accounting faculty and doctoral programs. *The Accounting Review, 60*(2), 262-277.

Brown, R. D., Braskamp, L. A., & Newman, D. L. (1978, May). Evaluator credibility as a function of report style: Do jargon and data make a difference? *Evaluation Quarterly, 2*(2), 331-341.

Brown, R. D., & Newman, D. L. (1982, Summer). An investigation to the effect of different data presentation formats and order of arguments in a simulated adversary evaluation. *Educational Evaluation and Policy Analysis, 4*(2), 197-203.

Brown, R. D., Newman, D. L., & Rivers, L. (1980, Sept.-Oct.). Perceived needs for evaluation and data usage as influencers on an evaluation's impact. *Educational Evaluation and Policy Analysis, 2*(5), 67-73.

Cawkell, T. (2000). Visualizing citation connections. In B. Cronin, & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 177-194). Medford, NJ: ASIS/Information Today.

Cheng, S. (2006). *A case study of evaluation use and influence in school settings.* Doctoral Dissertation: University of Minnesota.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions: An empirical examination. *American Journal of Evaluation, 28*(1), 8-25.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423-441.

Clark, K. E. (1954). The APA study of psychologists. *American Psychologist, 9*, 117-120.

Cole, J., & Cole, S. (1971). Measuring the quality of sociological research: Problems in the use of the science citation index. *American Sociologist, 6*(1), 23–29.

Cole, S. (1989). Citations and the evaluation of individual scientists. *Trends in Biochemical Science, 14*(1), 9-13.

Cole, S., & Cole, J. R. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review, 32*(3), 377-390.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52*, 291-302.

Coryn, C. L. S. (2006). The use and abuse of citations as indicators of research quality. *Journal of MultiDisciplinary Evaluation, 3*(4), 115-121.

Coryn, C. L. S., Hattie, J. A., Scriven, M., & Hartmann, D. J. (2007). Models and mechanisms for evaluating government-funded research: An international comparison. *American Journal of Evaluation, 28*(4), 437.

Cousins, J. B., Goh, S. C., Clark, S., & Lee, L. E. (2004). Integrating evaluative inquiry into the organizational culture: A review and synthesis of the knowledge base. *Canadian Journal of Program Evaluation, 19*(2), 99-141.

Cousins, J. B. (1996). Consequences of researcher involvement in participatory evaluation. *Studies in Educational Evaluation, 22*(1), 3-27.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331-364.

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 Publications.

Cronin, B., & Atkins, H. B. (2000). Introduction. In B. Cronin, & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 1-5). Medford, NJ: American Society for Information Science.

Culnan, M. J. (1986). The intellectual development of management information systems, 1972-1982: A co-citation analysis. *Management Science, 32*(2), 156-172.

Danin, S. (2000, 8/11/2000). Re: Classroom observation protocol. Message posted to Available from: http://www.edgateway.net/cs/ege/forum/cs_disc/573?x-showcontent=message_text.

Davenport, E., & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield. .Metford, NJ: Information Today Inc. ASIS Monograph Series,* 517-534.

Dickey, B. (1980). Utilization of evaluations of small-scale innovative educational projects. *Educational Evaluation and Policy Analysis, 2*(6), 65-77.

Egghe, L. (2000). New informetric aspects of the internet: Some reflections-many problems. *Journal of Information Science, 26*(5), 329.

Egghe, L. (2006). An improvement of the h-index: The g-index. *ISSI Newsletter, 2*(1), 8-9.

Elsevier B.V. (2008). *Scopus in detail: What does it cover?* Retrieved January 19, 2008, from http://www.info.scopus.com/detail/what/

Embretsen, S. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8), 449-455.

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

Gabel, D. L. (Ed.). (1994). *Handbook of research on science teaching and learning.* New York: Macmillan.

Garfield, E. (1979a). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359-375.

Garfield, E. (1979b). *Citation indexing: Its theory and application in science, technology, and humanities*. New York: John Wiley & Sons.

Garfield, E. (1998). The impact factor and using it correctly. *Der Unfallchirurg, 48*(2), 413.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science, 7*(1), 113-122.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goldberger, M. L., Maher, B. A., & Flattau, P. E. (1995). *Research-doctorate programs in the United States: Continuity and change.* National Academy Press.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8), 456-462.

Greene, J. C. (1987). Stakeholder participation in evaluation design: Is it worth the effort? *Evaluation and Program Planning, 10*(4), 379-394.

Greene, J. C. (1988). Stakeholder participation and utilization in program evaluation. *Evaluation Review, 12*(2), 91-116.

Greenseid, L. (2008, March). *Using citation analysis to measure evaluation influence.* Presentation to the Minnesota Evaluation Studies Institute, Bloomington, MN.

Greenseid, L. O., Johnson, K., & Lawrenz, F. (2008). *A citation analysis of the influence of the ATE, CETP, LSC, and MSP-RETA evaluations on the STEM education and*

evaluation fields (Report No. 6). Minneapolis, MN: Beyond Evaluation Use project, University of Minnesota.

Greenseid, L., & Toal, S. (2006, November 4). *Researching evaluation use and influence: Twenty years of empirical study*. Paper presented at the American Evaluation Association, Portland, Oregon.

Guba, E. G. (1972). The failure of education evaluation. In C. H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education* (pp. 250-265). Boston, MA: Allyn and Bacon.

Guba, E. G. (1978). *Toward a methodology of naturalistic inquiry in educational evaluation* (8th ed.). Los Angeles, CA: UCLA Center for the Study of Evaluation.

Gullickson, A. R., Wingate, L., Lawrenz, F., & Coryn, C. L. S. (2006). *The national science foundation's advanced technology education program: Final evaluation report*. Kalamazoo, MI: Western Michigan University, The Evaluation Center.

Hargens, L. L. (2000). Graphing micro-regions in the web of knowledge: A comparative reference-network analysis. In B. Cronin, & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 497-516). Medford, NJ: ASIS/Information Today.

Henry, G. T. (2000). Why not use? *New Directions for Evaluation,* (88), 85-98.

Henry, G. T. (2005). A conversation with Gary Henry. *The Evaluation Exchange, XI*(2), 10-11.

Henry, G. T., & Mark, M. M. (2003a). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*(3), 293-314.

Henry, G. T., & Mark, M. M. (2003b). Toward an agenda for research on evaluation. *New Directions for Evaluation,* (97), 69-80.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569-16572.

Hofstetter, C. H., & Alkin, M. C. (2003). Evaluation use revisited. In T. Kelleghan, & D. L. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 197-222). Great Britain: Kluwer Academic Publishers.

Holden, G., Rosenberg, G., & Barker, K. (2005). Bibliometrics: A potential decision making aid in hiring, reappointment, tenure and promotion decisions. *Social Work in Health Care, 41*(3-4), 67-92.

ISI/Thomson Scientific. (2007). *How do we identify highly cited researchers?* Retrieved August 20, 2007, from http://isihighlycited.com.floyd.lib.umn.edu/isi_copy/howweidentify.htm

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research, 50*, 438-460.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger Publishers.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*(2), 76-82.

Katzenmeyer, C., & Lawrenz, F. (2006). National Science Foundation perspectives on the nature of STEM program evaluation. In D. Huffman, & F. Lawrenz (Eds.), *Critical Issues in STEM Evaluation* (109th ed., pp. 7-18). San Francisco, CA: Jossey-Bass and the American Evaluation Association.

Kelly, C. D., & Jennions, M. D. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution, 21*(4), 167-170.

Kennedy, M., Apling, R., & Neumann, W. (1980). *The role of evaluation and testing programs in Title I programs.* Cambridge, MA: Huron Institute.

Kennedy, M. M. (1984). How evidence alters understanding and decisions. *Educational Evaluation and Policy Analysis, 7*(2), 207-226.

King, J. A., & Pechman, E. M. (1982). *The process of evaluation use in local school settings*. Final report of National Institute of Education Grant G-91-0900. New Orleans: New Orleans Public Schools.

King, J. A., Thompson, B., & Pechman, E. M. (1982). *Improving evaluation use in local school settings*. Final report of National Institute of Education Grant G-80-0082. New Orleans: New Orleans Public Schools.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation,* (88), 5-23.

Lawrenz, F., Gullickson, A., & Toal, S. (2007). Dissemination: Handmaiden to evaluation use. *American Journal of Evaluation, 28*(3), 275-289.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2005). Measures and mis-measures of scientific quality. *Arxiv Preprint physics/0512238.*

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2006). Measures for measures. *Nature, 444*(7122), 1003-1004.

Leviton, L. C., & Hughes, E. F. (1981). Research on the utilization of evaluations: A review and synthesis. *Evaluation Review, 5*, 525-548.

Leydesdorff, L., & Amsterdamska, O. (1990). Dimensions of citation analysis. *Science, Technology & Human Values, 15*(3), 305.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, November*, 437-448.

MacRoberts, M. H., & MacRoberts, B. R. (1986). Quantitative measures of communication in science: A study of the formal level. *Social Studies of Science, 16*(1), 151-172.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*(5), 342-349.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435-444.

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

McCain, K. W. (1999). Mapping authors in intellectual space: Population genetics in the 1980s. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 194-216). Newbury Park: Sage Publications.

Meho, L. I., & Sonnenwald, D. H. (2000). Citation ranking versus peer evaluation of senior faculty research performance: A case study of Kurdish scholarship. *Journal of the American Society for Information Science, 51*(2), 123-138.

Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *ISIS, 79*(299), 606-623.

Merwin, J. C. (1983). Dimensions of evaluation impact. *The Utilization of Evaluation: Proceedings of the Minnesota Evaluation Conference, May, 1983,* Minneapolis, MN.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (Third ed., pp. 13-103). New York: American Council on Education/Macmillan Publishing Company.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, March*, 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Moed, H. F. (2005). *Citation analysis in research evaluation*. The Netherlands: Springer.

Moed, H. F., & Bruin, R. E. (1999). International scientific cooperation and awareness: Bibliometric case study of agricultural research within the European Community. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 217-234). Newbury Park: Sage Publications.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86-92.

Nederhof, A. J., & van Raan, A. (1993). A bibliometric analysis of six economics research groups: A comparison with peer review. *Research Policy, 22*(4), 353-368.

Newman, D. L., Brown, R. D., & Braskamp, L. A. (1980). Communication theory and the utilization of evaluation. *New Directions for Program Evaluation, 5*, 29-35.

Newman, D. L., Brown, R. D., & Littman, M. (1979). Evaluator report and audience characteristics which influence the impact of evaluation reports: Does who say what to whom make a difference? *CEDR Quarterly, 12*(2), 14-18.

Newman, D. L., Brown, R. D., & Rivers, L. (1987). Factors influencing the decision-making process: An examination of the effect of contextual variables. *Studies in Educational Evaluation, 13*(2), 199-209.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, Calif.: Sage Publications.

Patton, M. Q., Grimes, P. S., Guthrie, K. M., Brennan, N. J., French, B. D., & Blyth, D. A. (1977). In search of impact: An analysis of the utilization of federal health evaluation research. In C. H. Weiss (Ed.), *Using social research in policy making* (pp. 141-163). Lexington, MA: Lexington Books.

Peritz, B. C. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics, 5*(5), 303-312.

Peters, H. P., & Van Raan, A. F. (1991). Structuring scientific activities by co-author analysis. *Scientometrics, 20*(1), 235-255.

Preskill, H., & Caracelli, V. (1997). Current and developing conceptions of use: Evaluation use TIG survey results. *Evaluation Practice, 18*(3), 209-225.

Reed, K. L. (1995). Citation analysis of faculty publication: Beyond science citation index and social science citation index. *Bulletin of the Medical Library Association, 83*(4), 503-508.

Rinia, E. J., van Leeuwen, T. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria evaluation of condensed matter physics in the Netherlands. *Research Policy, 27*(1), 95-107.

Rippey, R. M. (1973). The nature of transactional evaluation. In R. M. Rippey (Ed.), *Studies in transactional evaluation*. Berkeley, CA: McCutchan.

Rogers, E. M., & Cottrill, C. A. (1999). An author co-citation analysis of two research traditions: Technology transfer and the diffusion of innovations. In C. L. Borgman (Ed.), *Scholarly communications and bibliometrics* (pp. 157-165). Newbury Park: Sage Publications.

Rossi, P. H., & Freeman, H. E. (1982). *Evaluation: A systematic approach* (2nd ed.). Beverly Hills, CA: Sage Publications.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage Publications.

Rousseau, R. (2006). *A case study: Evolution of JASIS' hirsch index.* Available from: http://eprints.rclis.org/archive/00005430/.

Russell, Jr., R. L. (1999). *Testing for correlation between two journal ranking methods: A comparison of citation rankings and expert opinion rankings.* Unpublished Master's Thesis, Kent State University.

Schroeder, R. (2007). Pointing users toward citation searching: Using Google Scholar and Web of Science. *Libraries and the Academy, 7*(2), 243-248.

Scriven, M. (1993). Hard-won lessons in program evaluation. *New directions for program evaluation*, *58,* 1-107.

Scriven, M. (2007). Activist evaluation. *Journal of MultiDisciplinary Evaluation, 4*(7), 9/29/07.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ (Clinical Research Ed.), 314*(7079), 498-502.

Shadish, W. R., Tolliver, D., Gray, M., & Gupta, S. K. S. (1995). Author judgments about works they cite: Three studies from psychology journals. *Social Studies of Science, 25*(3), 477-498.

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *Evaluation Practice, 18*(3), 195-208.

Shulman, L. S., & Tamir, P. (1973). Research on teaching in the natural sciences. In R. M. W. Travers (Ed.), *Second handbook of research on teaching.* (pp. 1009-1148). Chicago: Rand McNally.

Small, H. (1982). Citation context analysis. *Progress in Communication Sciences, 3*, 287-310.

Small, H., & Greenlee, E. (1999). A co-citation study of AIDS research. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 166-193). Newbury Park: Sage Publications.

Smith, A. T., & Eysenck, M. (2002). *The correlation between RAE ratings and citation counts in psychology.* Available from http://cogprints.org/2749/

Spiegel-Rosing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science, 7*(1), 97-113.

Stokes, T. D., & Hartley, J. A. (1989). Coauthorship, social structure, and influence within specialties. *Social Studies of Science, 19*(1), 101-125.

Swales, J. (2001). Citation analysis and discourse analysis. *Applied Linguistics, 7*(1), 39-56.

Swanson, C. B., & Barlage, J. (2006). *Influence: A study of factors shaping education policy.* Editorial Projects in Education Research Center, Bethesda, MD.

Tamir, P. (1996). Science education research viewed through citation indices of major reviews. *Journal of Research in Science Education, 33*(7), 687-691.

Thackray, A., & Brock, D. C. (2000). Eugene garfield: History, scientific information, and chemical endeavor. In B. Cronin, & H. B. Atkins (Eds.), *The web of knowledge:*

A festschrift in honor of Eugene Garfield (pp. 11-23). Medford, NJ: American Society for Information Science.

The Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects and materials*. New York: McGraw-Hill Book Company.

Thomson Scientific. (2006). *Multidisciplinary resources: ISI web of knowledge.* Retrieved January 19, 2008, from http://scientific.thomson.com/media/scpdf/wok_multidisc_fs.pdf

Thomson Scientific. (nd). *Product specs - ISI web of knowledge.* Retrieved March 13, 2008, from http://isiwebofknowledge.com/currentuser_wokhome/cu_productspecs/

Trochim, W. M., Marcus, S. E., Masse, L. C., Moser, R. P., & Weld, P. C. (2008). The evaluation of large research initiatives: A participatory integrative mixed-methods approach. *American Journal of Evaluation, 29*(1), 8.

van Raan, A. F. J. (2001). Bibliometrics and internet: Some observations and expectations. *Scientometrics, 50*(1), 59-63.

van Raan, A. F. J. (2006). Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3), 491-502.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? *Journal of the American Society for Information Science and Technology, 54*(14), 1313-1322.

Walter, G., Bloch, S., Hunt, G., & Fisher, K. (2003). Counting on citations: A flawed way to measure quality. *The Medical Journal of Australia, 178*(6), 280-281.

Weiss, C. H. (1972a). Evaluating educational and social action programs: A treeful of owls. In C. H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education* (pp. 3-27). Boston, MA: Allyn and Bacon.

Weiss, C. H. (1972b). Utilization of evaluation: Toward comparative study. In C. H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education* (pp. 318-326). Boston, MA: Allyn and Bacon, Inc.

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Knowledge: Creation, Diffusion, Utilization, 1*(38), 381-404.

Weiss, C. H., Murphy-Graham, E., & Birkeland, S. (2005). An alternate route to policy influence: How evaluation affect DARE. *American Journal of Evaluation, 26*(1), 12-30.

White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology, 52*(2), 87-108.

White, H. D. (2004). Citation analysis and discourse analysis revisited. *Applied Linguistics, 25*(1), 89-116.

White, H. D. (1990). Author co-citation analysis: Overview and defense. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 84-106). Newbury Park: Sage Publications.

Williams, Harrison A. ,Jr. (1979). Foreword. In F. M. Zweig (Ed.), *Evaluation in legislation* (pp. 7-9). Beverly Hills: Sage Publications.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines* (2nd ed.). New York: Longman.

Wouters, P. (1999). *The citation culture.* Unpublished Doctoral Dissertation, Universiteit van Amsterdam.

Wouters, P. (2000). Garfield as alchemist. In B. Cronin, & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 5-71). Medford, NJ: American Society for Information Science.

Yang, K., & Meho, L. I. (2006). Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science. Paper presented at the *American Society of Information Science and Technology Conference, November 3-9, 2006,* Austin, TX. Available from http://eprints.rclis.org/archive/00008121/01/Yang_citation.pdf

Young, C. J., & Comtois, J. (1979). Increasing congressional utilization of evaluation. In F. M. Zweig (Ed.), *Evaluation in legislation* (pp. 57-79). Beverly Hills: Sage Publications.

Zsindely, S., & Schubert, A. (1999). Editors-in-chief of medical journals: Are they experts, authorities, both, or neither? In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 248-253). Newbury Park: Sage Publications.

Zubrowski, B. (2007). An observational and planning tool for professional development in science education. *Journal of Science Teacher Education, 18*(6), 861-884.