The Development and Validation of an Evaluation Use Scale for Multi-site Evaluations

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Kelli A. Johnson

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

Jean A. King, Adviser

July 2016

Acknowledgements

Countless thanks to Dr. Frances Lawrenz (PI) and Dr. Jean King (Co-PI) for the opportunity to work as part of the research team in the NSF-funded Beyond Evaluation Use study that made this research possible. I am honored and humbled to have been able to participate in this work. Special thanks to my longstanding colleagues on that research team – Stacie Toal, Ph.D., Lija Greenseid, Ph.D., Boris Volkov, Ph.D., Denise Roseland, Ph.D., and Gina Johnson, Ph.D. (listed in order of completion!) – for their hard work and dedication, as well as for their inspirational examples of focus and fortitude. Special thanks to Stacie Toal for blazing the validation trail and for her generous help and guidance along the way.

Thank you to my committee members Dr. Melissa Anderson, Dr. Jean King, Dr. Frances Lawrenz, and Dr. Karen Storm for the generous gift of time, energy, and attention, as well as for flexibility and patience in the homestretch. Thanks also to Dr. Frances Vavrus, Director of Graduate Studies, for approving my application for extra time to complete my degree requirements; and to the scores of individuals who signed and processed the other University forms required for degree completion.

A colossal thank you to Lynn Blewett for her enthusiastic professional support and academic assistance, and for getting this ball rolling by hiring me at the University in the first place. Thanks to my colleagues at the State Health Access Data Assistance Center (SHADAC) in the Division of Health Policy and Management (UMN School of Public Health) for providing unflagging support and camaraderie, for generously filling

in for me and picking up the slack, and for making each work day an enjoyable opportunity for learning and professional growth.

Thank you to the generous members of the editorial utility group, particularly Stuart Appelbaum and Nancy Evert for their time, energy, attention, focus, and talent (editorial *and* musical), as well as for proofreading, copyediting, and cheerleading with skill and gusto.

Profound thanks to my dear sisters (biological and spiritual) for believing in me: to Lorie Halvorson, for her unwavering support and quiet persistence in asking whether there had been any progress on the project whose name shall not be spoken; to Kari Johnson, for the indelible memory of her loving encouragement and irreplaceable sense of humor; to Sonnie Elliott, for meeting regularly to have coffee and "work on my dissertation"; and to Nancy Evert, for her tenacity in providing emotional support, editorial assistance, and gentle coaxing me to finally stick a fork in it.

Finally, boundless gratitude to my adviser, mentor, and friend, Jean King, for providing intellectual and professional inspiration, tireless encouragement, and unstinting patience and support. Without her generosity, equanimity, and pragmatic exuberance, this story would have had a different ending.

Abstract

Evaluation researchers and practitioners share a commitment to evaluation use, and the research community has focused on evaluation use because it is an essential component of the practice of evaluation. While evaluation use is among the most frequently studied topics in the field, a scale for measuring the use of evaluations in multi-site settings has yet to be validated. This study describes the development and validation of the Evaluation Use Scale for assessing program evaluation use and the factors associated with evaluation use in multi-site science, technology, mathematics, and engineering (STEM) education programs.

The data used in this study were collected as part of the NSF-funded Beyond Evaluation Use study (Lawrenz & King, 2009) and included the development and administration of the NSF Program Evaluation Follow-up Survey, a web-based survey of project leaders and evaluators in four multi-site STEM education programs. The study used Messick's unitary concept of validity as a framework for assembling empirical evidence and theoretical rationales to assess the adequacy and appropriateness of inferences and actions based on the Evaluation Use Scale.

The overall evidence in support of the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site evaluations is mixed and varies by aspect of validity. In four of the six aspects, the evidence is adequate to strong. However, the evidence in the remaining two aspects is sharply split and, therefore, inconclusive.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE
INTRODUCTION

One of the principal objectives of program evaluators is to inform program and policy decisions by producing useful findings, evidence, and information (House, 1980; Patton, 1997; Stufflebeam & Shinkfield, 2007). The field of program evaluation grew in the 1960s and 1970s fueled, in large part, by federally mandated evaluations of education and social service programs. At the same time, evaluators and researchers recognized that many evaluations were not being used. This realization motivated discussion, investigation, and soul searching among many early leaders in the field (Alkin, Daillak, & White, 1979) and the extensive and continuing study of evaluation use had begun.

Evaluation Use

For nearly half a century, evaluation use has been examined, categorized, defined, and re-defined, making it among the most frequently studied topics in the field (Christie, 2007; King & Pechman, 1984; Leviton & Hughes, 1981). Researchers and practitioners alike shared the commitment to evaluation use, and the research community focused on evaluation use because it is an essential component of the practice of evaluation (Kirkhart, 2000).

In the 1970s and 1980s, some evaluation theorists shifted to a more pragmatic approach. Weiss focused on the importance of producing results that would be useful to people and programs. She saw the importance of political context and highlighted the need to address the complicated organizational contexts in which use is anticipated, noting that though a project may look simple in theory, organizational obstacles invariably arise. Weiss (1979) described the complexity of the concept of research utilization and identified the importance of understanding what “utilization” really means before effective assessment of whether or to what it extent it can occur. In contrast to the pervasive understanding that utilization meant the direct application of research results to a specific decision, Weiss wrote, “The process is not one of linear order from research to decision, but a disorderly set of inter-connections and back-and-forthness that defies neat diagrams” (p. 428).

In an important theoretical and practical development, Michael Quinn Patton (1997) introduced the concept of utilization-focused evaluation in which the evaluator designs and focuses the evaluation, including the choice of methods, on the expressed needs of the specific individuals identified as *primary intended users*. Patton also coined the term *personal factor* encapsulating his finding that when asked to identify the most important factor affecting evaluation utilization, people often identify individual, named, intended users.

The *Program Evaluation Standards* published by the Joint Committee on Standards for Educational Evaluation (2011) include *utility* as one of the standards intended to provide guidance both to evaluators and to users of evaluation. An evaluation’s utility is based upon “the extent to which program stakeholders find evaluation processes and products valuable in meeting their needs” (p. 4).

Many prominent evaluation researchers and practitioners have discussed the importance of program evaluation use. Rossi (2004) noted that one of the distinguishing characteristics of an evaluation is how it is used, along with the importance of judging evaluations by their utility. He further asserted, “Evaluation is infrequently if ever thought of as ‘knowledge for knowledge’s sake’ but rather it is intended to be useful.” Patton (1997) joined Rossi, commenting that in developing or tailoring evaluations, priority should be given to evaluation questions that will yield information most likely to be used by program personnel and policy makers.

Evaluation Participation

In addition to use, another relevant area of evaluation research is participatory evaluation, which asserts that involving people throughout the evaluation process will result in a greater sense of ownership and, ultimately, more use of the information produced from the evaluation. Extensive research has demonstrated that stakeholder participation can enhance the use of both the evaluation process and evaluation findings (Cousins & Earl, 1992; Patton, 1997).

Burke (1998, p. 47) identified key moments and decision points in the participatory evaluation process. These include: (1) deciding to do the study, (2) assembling the evaluation team, (3) making a plan, (4) collecting data, (5) synthesizing, analyzing and verifying the data, (6) developing action plans for the future, and (7) controlling and using outcomes and reports. Lawrenz and Huffman (2003) applied the evaluation use and participatory evaluation frameworks to five large, multi-site evaluations to assess the relationship between high levels of involvement in program evaluations and the potential association of that involvement with higher levels of use of the evaluation results.

Statement of the Problem

Even as the need for continuing research to enhance and encourage the use of evaluation results among stakeholders has long been acknowledged, there is currently no validated measure of evaluation use for multi-site settings. This study describes the development and validation of a scale to measure evaluation use in four National Science Foundation (NSF) sponsored, multi-site Science, Technology, Engineering and Mathematics (STEM) education programs. The development and validation of the Evaluation Use Scale will potentially increase our ability to measure evaluation use in these settings, thereby enhancing our understanding of the levels of use and informing efforts to increase evaluation use. Such a scale will provide stronger evidence when it is used to measure the effect of an intervention or of some factor(s) influencing the use of evaluations.

Despite broad agreement in the field on the importance of using evaluation findings, no one has yet developed and validated an instrument or "use scale" by which to measure evaluation use. The quality of a measurement instrument is critical to the social science research process. DeVellis (2003) contends that regardless of how well-designed and executed a research study might be, "measurement can make or break a study" (p. 6). To that end, the focus of this study is the development and validation of an evaluation use scale for multi-site program evaluations.

Purpose of the Study

The purpose of the present study is twofold: first, to benefit the practice of evaluation by identifying factors critical to evaluation use that will, in turn, inform evaluation design to maximize the use of multi-site evaluations; and second, to improve research on evaluation use by providing a tested, effective tool for measuring evaluation use in multi-site evaluations. Should the validation study prove inconclusive and unable to support the validity of inferences about evaluation use based on the information collected using the scale, this study will help identify the areas where the scale is weak and where efforts to strengthen the validity evidence should be focused to effect adjustments to the scale.

Study Context

The present study deals specifically with NSF-funded programs across multiple sites in multiple states. Through the 1990s, evaluation use research primarily involved single-site programs. However, as federal agencies, including NSF, began to fund more large-scale, multi-site programs – often related to teacher professional development and STEM education improvement – the frequency of multi-site evaluations also increased (Lawrenz & Huffman, 2003; Straw & Herrell, 2002). Multi-site evaluations differ from single-site not only because they take place in more locations, but also because they may include cross-site evaluation activities or analyses.

Two grounding principles for multi-site evaluations in the social services were proposed by Leff and Mulkern (2002). According to these principles, evaluations should be 1) science-based (i.e., use scientific investigation techniques), and 2) participatory-based (i.e., include a broad range of stakeholders). The participatory-based principle is relevant to the present study, because part of the rationale for promoting participatory evaluation has been a tacit acceptance of the notion that more involvement in an evaluation will be associated with more use of the evaluation process and products; and that more use of the information generated in a program evaluation is better than less use.

However, it is especially challenging to generate higher levels of participation in evaluations across multiple sites. Lawrenz and Huffman (2003) identified three specific challenges faced by multi-site evaluators. These challenges stem from the unique characteristics of multi-site evaluations and are relevant here because the evaluation use scale that forms the center of this study was developed to measure evaluation use in exactly this type of multi-site situation. Further, they specified the challenges to increased participation and use faced by evaluators of multi-site programs who are dealing with: a) different layers of stakeholders across the sites, b) the wide diversity of stakeholders, and c) the fact that the personnel at different project sites within a given program may not know each other.

NSF requires program evaluations to be conducted for all of the programs funded through the Education and Human Resources Directorate. Moreover, NSF provides the funding to conduct these evaluations and strongly promotes the use of the results of these evaluations for program improvement purposes. In January 2005, NSF awarded a grant to the University of Minnesota to study the extent to which participation and involvement in four multi-site evaluations affected the use of the evaluation findings. The grant design for these four NSF-funded multi-level evaluations required a description of involvement by project personnel in the design, implementation and communication of these multi-site evaluations. In addition, the project addressed how this involvement affected the ultimate use of the evaluation findings, including: the use and influence of evaluations of four national, multi-site NSF programs; the relationship between the extent of involvement of evaluation stakeholders and the eventual use of the evaluations; and the long-term impact of the evaluations on project staff, on the STEM community, and on the evaluation field.

Validity Framework

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment.

Validation is the process of assembling empirical evidence and theoretical rationales to assess the adequacy and appropriateness of inferences and actions based on assessments (Messick, 1989a, 1989b, 1995a, 1995b, 1996a, 1996b). The validation process requires extended analyses of inferences and assumptions and involves developing a rationale for the proposed interpretation and a consideration of possible competing interpretations (Kane, 2001). Though considerable research on evaluation use has been published in the field of evaluation over the past two decades, no single instrument or approach has undergone a comprehensive validation study.

The present study employed a unitary validity framework based on the work of Samuel Messick (1989a, 1989b, 1995a, 1995b, 1996a, 1996b) to compile evidence of validity of the Evaluation Use Scale which was developed and implemented as part of a web-based survey of nearly 400 participants in NSF-funded multi-site program evaluations. Construct validity is central to the unitary view of validity (Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1989, 1995). The various types of validity (criterion validity, internal validity, and external validity) are facets that function together to support the validity of inferences made based on data generated using a particular instrument. Loevinger (1957) wrote that "since predictive, concurrent, and content validities are all essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view" (p. 636).

Messick (1989b) wrote "Construct validity comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables" (p. 7). Constructs are abstractions created to conceptualize latent variables, which are variables that are not directly observable, but can be inferred from other variables that can be directly measured.

Messick identified six aspects of construct validity that together serve as the framework for gathering and analyzing the evidence put forward in a validity argument (1989a, 1995a, 1995b, 1996b). These six aspects of construct validity are:

1. Content – relevance of the test content and representativeness of the domain;
2. Substantive – theoretical rationale for observed consistencies in responses;
3. Structural – alignment of statistical factors and test structure;
4. Generalizability – extent to which scores generalize across groups;
5. External – convergent and discriminant evidence; and
6. Consequential – value implications for score interpretations.

In conducting a validation study like this, the researcher assembles evidence in each of the aspects of validity by demonstrating the extent to which the instrument – in this case, the evaluation use scale from the web-based survey – meets the criteria for validity. These six aspects of construct validity afford a means of checking that the "theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases" (Messick, 1995b, p.747).

Research Questions and Method

A primary research question and eight sub-questions frame this validation study. The questions address the six aspects of construct validity and guide the assembly of evidence to assess the extent to which the Evaluation Use Scale meets the criteria for validity in the context of multi-site evaluations.

The primary research question is:

Is there sufficient evidence to support the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site evaluations?

The sub-questions supporting the primary research question are:

To what extent: . . .

Q1a. Is the construct of evaluation use theoretically sound?

Q1b. Is the Evaluation Use Scale internally consistent and reliable?

Q2a. Does the Evaluation Use Scale include all relevant activities and processes?

Q2b. Does the Evaluation Use Scale measure actual use?

Q3. Do the statistical factors of the Evaluation Use Scale align with the underlying theoretical and rational structures?

Q4. Does the Evaluation Use Scale consistently measure evaluation use in other multi-site settings?

Q5a. Does the Evaluation Use Scale correlate with other expected activities?

Q5b. Does the Evaluation Use Scale differentiate levels of use between groups of individuals that are rationally or theoretically different in terms of use?

Q6. Could possible interpretations or uses of the Evaluation Use Scale result in a negative impact?

Consideration of the primary research question and the supporting research questions formed the basis of the validation study, guiding the gathering of evidence and the assessment of the degree to which the evidence supports the adequacy and appropriateness of inferences based on scores from the Evaluation Use Scale. Scores reflect the underlying constructs more accurately in some cases than in others, but never perfectly; therefore, decisions about validity will not be clear-cut yes or no decisions, but rather will reflect a matter of degree based upon the evidence, e.g., strong, adequate, marginal, weak, unsatisfactory (Downing, 2004; Hubley & Zumbo, 2011; Messick, 1980; Toal, 2007).

CHAPTER TWO
LITERATURE REVIEW

This literature review addresses the concept of evaluation use as an important pillar of evaluation research and practice. Table 1 provides definitions of several key terms in the area of evaluation use. The chapter describes the various definitions and categories of evaluation use, traces the development and expansion of the concept, discusses the theoretical foundations, and outlines the various factors identified by researchers as being associated with evaluation use. In addition, this chapter includes a review of the validity literature that informed this validation study.

Evaluation Use

For over 40 years, researchers and practitioners in the evaluation field have maintained a strong focus on the importance of evaluation use. Scholars have defined evaluation use, tweaked the definitions, changed the words, debated the categorization of evaluation use components, presented at conferences, convened expert panels, and published multiple special issues of evaluation journals dedicated to the study of evaluation use.

One of the ongoing debates has been between the terms *use* and *utilization*. Early on, Weiss (1979) objected to the term *evaluation utilization* because it conjured images of the mechanical and unchanging. Other evaluation researchers joined Weiss in preferring the term *evaluation use*, rather than *evaluation utilization (*King, 1988). Daillak (1982), on the other hand, disagreed, explaining his position that *utilization* enhances the meaning of use by adding a “connotation of beneficial, profitable, or otherwise productive result” (p. 158).

Table 1

Key Evaluation Use Terms

Term	Definition
Evaluation Influence	Capacity of evaluation processes, products, or findings to indirectly produce a change in understanding or knowledge (King et al., 2006)
Evaluation Involvement	Active engagement in at least one phase of an evaluation, i.e., planning, implementing, or applying the results. (King et al., 2006)
Evaluation Use	The application of evaluation processes, or findings to directly produce an effect (King et al., 2006)
Instrumental Use	Action relates directly to a decision-making or problem-solving purpose (King & Pechman, 1984; Leviton & Hughes, 1981).
Conceptual Use	Over time, evaluation results influence a user’s thinking about a problem, but results in no documented use (King & Pechman, 1984; Leviton & Hughes, 1981)
Symbolic or Persuasive Use	Users apply the evaluation process or its results for personal ends (e.g., to garner political support or to discredit a policy; Drawing on evaluation evidence in attempting to convince others to support or defend a political position (King & Pechman, 1984; Leviton & Hughes, 1981)

The discussions over definitions and terms has stretched over several decades, leading evaluation scholars to broaden the definitions and expand the explanation of evaluation use to include basic knowledge and understanding of the evaluation plan and findings, possible justification for decisions already made, and the impetus for action (program improvement and/or decision making) based on the evaluation findings (Patton, 1997). For purposes of this paper, the term *use* will be used exclusively, rather than including the arguably interchangeable term, *utilization*.

Evaluation Use History

In the 1960s and 1970s, federally funded evaluation grew with the U.S. Elementary and Secondary Education Act (ESEA) and the Great Society programs of the Office of Economic Opportunity (Patton, 1997; Shadish, Cook, & Leviton, 1991; Stufflebeam & Shinkfield, 2007). At that time, social science researchers and evaluators anticipated that their work would be useful to public officials and policymakers, but they soon became concerned that their work was not being used.

Weiss (1979) described policy makers "displaying spurts of well publicized concern about the usefulness of the social science research that government funds support" (p.426). Alkin et al. (1979) wrote, "In the graveyard of ignored or disregarded evaluations rest not only those technically inferior studies which earned their consignment to oblivion; there are also many studies seemingly of high quality which somehow failed to move their audiences to action" (p. 13).

Concern about how to address this problem intensified with the knowledge that society's justification of the substantial government investment in evaluation came with the expectation of immediate benefits. Thus, if the evaluations did not prove useful, the funding was at risk of being redirected to other purposes (Shadish et al., 1991).

These concerns about the use evaluation findings resulted in continuing calls for more study of the issue. Researchers responded to these concerns by conducting empirical investigations of the use of evaluation information, and these involved establishing clear definitions of evaluation use. In the mid-1970s, researchers at the UCLA Center of the Study of Evaluation undertook a program of systematic research on evaluation use that was motivated to some degree by the "grandiose claims made by some of the early evaluation advocates… [that] evaluation would have a significant and highly visible impact on the program being evaluated" (Burry, Alkin, & Ruskus, 1985, p.133).

At the same time, researchers at the Center for Social Research at the University of Minnesota conducted a study of evaluation use and the factors related to it in several national health programs evaluations. The National Institute of Education at the U.S. Department of Health, Education, and Welfare funded the research which included in-depth interviews of the leaders of mental health program leaders to examine the extent to which evaluations were used and to identify factors the interviewees' described as being associated with the use or lack of use of evaluative information was used or not (Patton et al., 1977).

Initially, discussions of and research on evaluation use drew on a narrow characterization of use, focused primarily on the direct use of evaluation results in decision-making (Leviton & Hughes, 1981). Scholars labeled this narrow form *instrumental use* and asserted that it involved taking direct actions or making direct decisions about changing programs based on evaluation results (Alkin, 2005).

However, as evaluation use studies began to produce findings, scholars concluded that the field needed a broader, more-encompassing definition of evaluation use to properly recognize the evidence that evaluation use was occurring. In addition, several studies began to outline factors related to evaluation use, as well as to identify specific contexts that fostered the use or non-use of evaluations (Burry et al., 1985).

Research conducted during this period resulted in the development of a greater understanding of the nature of evaluation use. Despite concerns to the contrary, practitioners and scholars began to realize that people were, in fact, using the information produced through program evaluation. However, the uses of program evaluation information were not always consistent with the definitions and expectations that had been established by program evaluation researchers and scholars (Patton et al., 1977; Shadish et al., 1991). These studies contributed to expanding the accepted definition of evaluation use beyond direct, instrumental use to include changes in thinking or understanding.

Weiss and Bucuvalas (1980) studied the use of social science research in decision-making, focused specifically on the decision-makers' frames of reference. They found few instances of immediate application of information to decision-making and many more instances of information driving subtle shifts in attitudes or thinking rather than resulting in direct use.

In what they called one of the most important findings of this research, Weiss and Bucuvalas (1980) learned that the concept of use is "… an exceedingly ambiguous concept" that can be interpreted by decision-makers in wide range of ways. Specifically, they learned that decision-makers might interpret the notion of use to include the use of information in a conceptual manner to “clarify their own thinking, re-order priorities, [or] make sense of what they have been doing” (p. 305). The label c*onceptual use* indicated instances where knowledge generated by the study of a particular issue influenced a policy maker's thinking, but the information was not put to a direct, specific, documentable, or instrumental use (Alkin, 2005).

Weiss continued to study how policy makers and government officials reported how they did or did not use research findings in making policy decisions. Based on interviews with 155 officials in public mental health agencies (at the federal, state, and local levels) about whether and how they used or did not use social science research results in making decisions on the job, Weiss (1980) found that a majority of the officials reported having used social science research in their work but most could not cite specific studies the affected specific decisions.

Weiss concluded that rather than consciously applying the findings of a particular research study to an identified problem, the process for public officials and policy makers is more diffuse. Research information encountered through daily work, continuing education, and other means of keeping abreast of developments in the field entered into their awareness and affected their thinking and actions, but it was difficult to isolate exactly how or where that entry had occurred. Weiss (1980) labeled this indirect process whereby individuals absorb information and assimilate concepts as “knowledge creep” (p. 397).

In addition, Weiss also articulated her insight into the nature of decision-making in the policy arena. Contrary to the implied sense of direct action in executing decisions related to public policy, Weiss stated that policy actions are not decided in a clear cut fashion. Instead, policies may “come into being without systematic consideration.” Instead, policy decisions can take shape gradually, moving through formative stages, and little by little build up. In this way, Weiss asserts that “without conscious deliberation, the policy *accretes*.” (1980, p. 382).

Weiss called such conceptual, non-direct evaluation use “enlightenment” and described it as, “The percolation of new information, ideas, and perspectives into the arenas in which decisions are made…Over time, the ideas from evaluation seep into people’s consciousness and alter the way that issues are framed and alternatives designed (Weiss, 1999, p. 471).

Another type of evaluation use, namely persuasive or symbolic use, occurs when an evaluation is used to advocate for a particular position, or when individuals use evaluation results for political self-interest (Patton, 1997). Evaluation researchers have spent less time studying and writing about this type of use than the other types of use, and there is more variation in the definitions about this kind of use in the literature.

Finally, Weiss identified another type of use as *imposed use*. Imposed use occurs when decision makers use evaluation information to make program-related decisions under pressure from another entity, often a government mandate or the compliance requirements of other funders (Weiss, 2005).

In addition to the use of evaluation findings, the evaluation process itself may have impact, instrumental or conceptual (Alkin, 2005). First labeled *process use* by Patton (1997) it refers to the way people use what they learn from participating in the evaluation process and to behavioral and cognitive changes in people as a result of their involvement in evaluation. It also includes programmatic or organizational changes in procedures and culture that result from the learning that occurs during the evaluation process (Patton 1997; Preskill & Caracelli, 1997; Shaw & Campbell, 2014; Shulha & Cousins, 1997).

Evaluation Use Frameworks

Beginning in the early 1980s, evaluation researchers began to summarize the research of the previous decade through literature reviews identifying lists of factors related to evaluation use (Alkin, 1985; Leviton & Hughes, 1981).

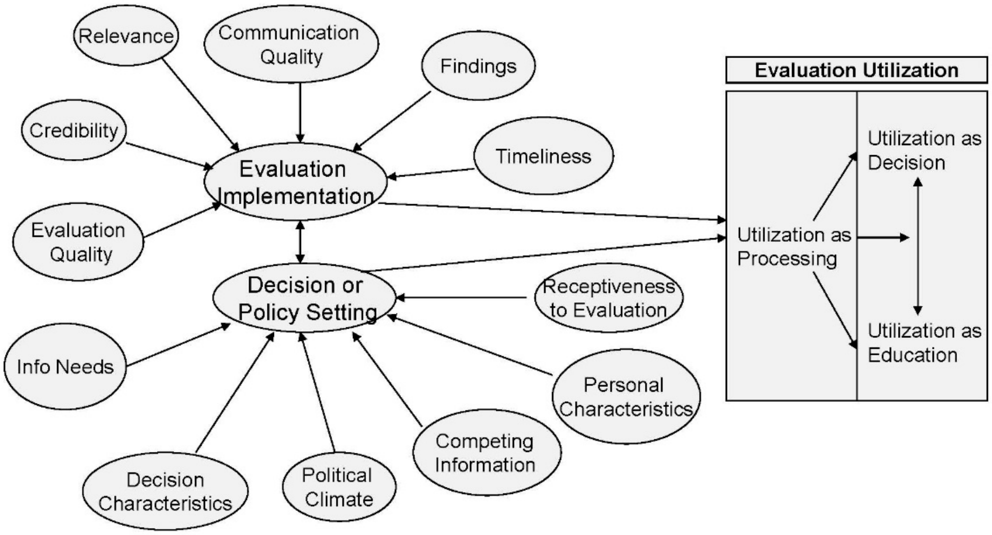
Weiss and Bucuvalas (1980) identified two main factors associated with the use of social science research findings: (1) research quality, which includes technical quality, objectivity, internal consistency, data support for recommendations, and other related factors; and (2) action orientation, which includes, among others, the presence of explicit recommendations, direct implications for a course of action, and applicability with a specific program (pp. 303-04).

Leviton and Hughes (1981) identified five factors that were likely to result in evaluation clients thinking or acting differently than they would have without the evaluation information. These five factors include: (1) relevance, (2) communication, (3) information processing, (4) credibility, and (5) user involvement/advocacy (p. 534).

Alkin (1985) published a *Guide for Evaluation Decision Makers*, geared toward increasing the use of evaluation findings among non-profit and public sector decision makers. The guide identified three factors related to the use of evaluation information – (1) human factors (e.g., characteristics of the evaluator and the evaluation users, such as political sensitivity, credibility, experience level, and commitment to the use of evaluation results); (2) context factors (e.g., project characteristics, evaluation constraints, contract and project details); and (3) evaluation factors (e.g., methods, reporting) (p. 27).

Figure 1.

Evaluation Utilization Framework – Cousins and Leithwood (1986)



From “Current empirical research on evaluation utilization. Review of Educational Research, 56 (3), p. 348. Copyright 1986 by the American Educational Research Association. Reprinted with permission.

Broadening Use to Influence

For the past 15 years, the evaluation literature has called for an expansion of the concept of evaluation use to a broader concept of evaluation influence. Over time, the concept of evaluation has evolved to encompass additional attributes beyond the original direct or instrumental use components (Cousins & Leithwood, 1986; Shulha & Cousins, 1997). Kirkhart (2000) argues that these developments notwithstanding, “An inclusive understanding of the influence of evaluation has been hampered by the scope and language of past approaches.” She goes on to suggest that evaluation scholars “step back and reconceptualize the terrain of evaluation's influence by mapping influence along three dimensions – source, intention, and time" (p. 5).

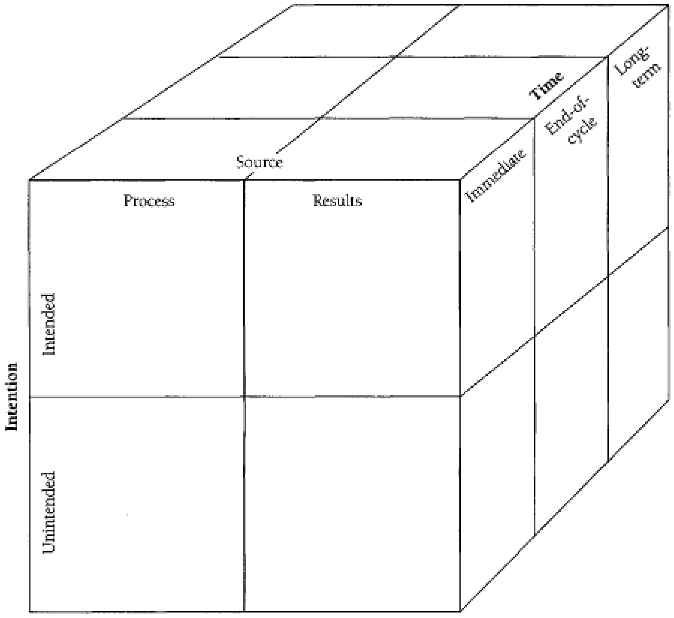
Alkin and Taut (2003) describe ongoing dissonance in the literature relative to evaluation, noting that "…the evaluation use concept has come under attack from several different directions" (p. 1). These dissenters include scholars and practitioners who, like Kirkhart, prefer the concept of evaluation influence to evaluation use; as well as those who do not distinguish between evaluation use and knowledge use (Cousins & Simon, 1996).

Kirkhart (2000) defined influence as "the capacity or power of persons or things to produce effects on others by intangible or indirect means" (p. 7). According to her framework using the three dimensions of source, intention, and time, evaluation use is viewed as direct, intended, and tangible, but influence is a broader concept that includes all possible impacts of an evaluation.

Alkin and Taut (2003) challenge Kirkhart's characterization of influence asserting instead that use and influence should describe different types of evaluation impacts. The authors suggest the addition of the concept of awareness to the influence model such that when a potential evaluation user can specify a particular impact, that would be called awareness; but when a user cannot specify the impact, that would be considered influence. In addition, they maintain that because influence is unintended, "it is outside the domain of the evaluator to affect such possible influences" (p. 9).

Figure 2.

Integrated Theory of Influence – Kirkhart (2000)

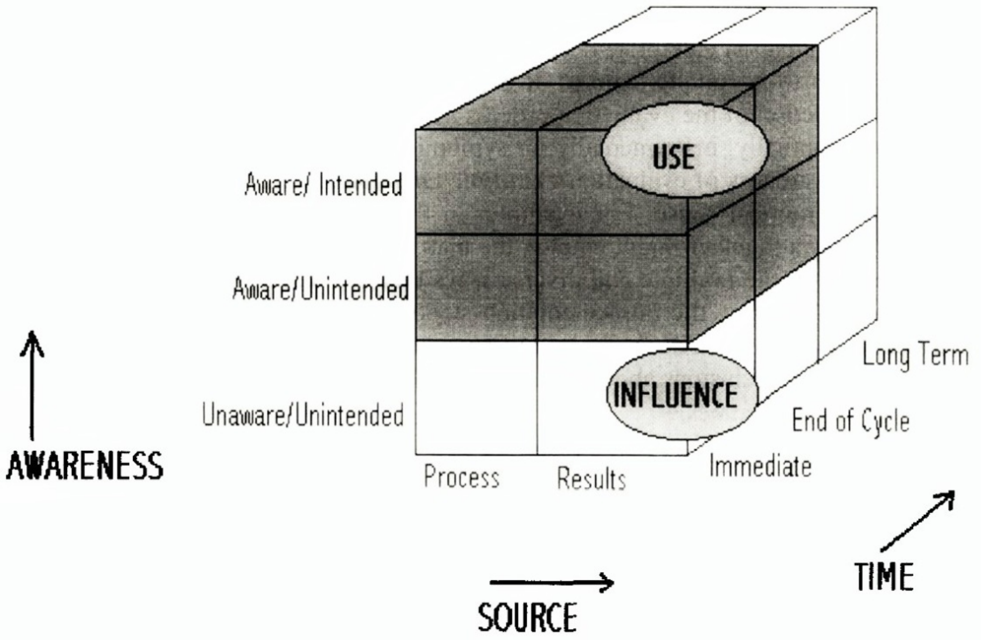


From “Reconceptualizing Evaluation Use: An Integrated Theory of Influence,” by K.E. Kirkhart, 2000, *New Directions for Evaluation*, no. 88, p. 8. Copyright 2000 by John Wiley and Sons. Reprinted with permission.

Mark and Henry (Henry & Mark, 2003; Mark & Henry, 2004) join Kirkhart in advocating for broadened perspective on the use and influence of program evaluations. Unlike Alkin and Taut, however, Mark and Henry argue that the goal of evaluation is broad social betterment, not merely use. Henry and Mark (2003) saw the value of Kirkhart’s concept of evaluation influence in their focus on social betterment as the goal of evaluation. They view evaluation as a social intervention that achieves goals by effecting changes at the individual, interpersonal, and collective levels. Moreover, they suggest that influence results from a web of pathways and mechanisms of influence.

Figure 3.

Evaluation Impact within the Program Context (adapted from Kirkhart)



From “Unbundling Evaluation Use” by M.C. Alkin and S.M. Taut, 2003, Studies in Educational Evaluation, 29, p. 9. 2003. Copyright 2003 by Elsevier Science Ltd. Reprinted with permission.

Evaluation Influence

Kirkhart (2000) published an integrated theory of evaluation influence to build toward a broader understanding and potential new insight into how to understand and enhance evaluation use. Kirkhart argued that in order to examine the question of how evaluation affects and changes people and systems, scholars should take a more broad-based approach. She suggested the use of the more expansive term *influence* as it would not inappropriately constrict our perceptions and would help build a deeper and more broadly inclusive understanding of the impact of evaluations.

Studying Evaluation Influence.

While the body of literature on evaluation influence is growing, it is not nearly as abundant as research on evaluation use. Some researchers address evaluation influence as the focus of their studies (Mark & Henry, 2004; Shaukat, 2010), while many others simply use the term influence as an add-on to the term evaluation use.

The growing attention to evaluation influence has been an important development in research on evaluation over the past 15 years. After decades of wrestling with and redefining the concepts associated with evaluation use, expanding the discussion is helpful understanding the effects evaluation can have in the broadest sense which, in turn, enables evaluation researchers to better describe and understand what occurs during and following an evaluation. However, just like the term “use” the term “influence” is a commonly used word with multiple meanings and interpretations.

Herbert (2014) conducted a literature review to identify studies of evaluation influence. Overall, he found that studies that attempted to incorporate elements of Mark and Henry’s framework found varying degrees of success in following the interaction and interrelationship between different mechanisms. Five studies included in the review were more successful in tracing influence, with each beginning their research fairly soon after the evaluation began. To date, however, the literature still lacks published research putting forward ways of measuring of evaluation influence. Nunneley, King, Johnson, and Pejsa (2015) concluded that the addition of the concept of evaluation influence to the ongoing evaluation research agenda may be helpful overall as it has broadened the practitioners’ and researchers’ thinking about the effects of evaluation.

Evaluation use remains an important component of program evaluation. To better understand and contribute to its ongoing development, it is necessary to be able to specify the concept and know how to measure it. Conducting a validation study of a scale to measure evaluation use in multi-site settings will enhance the reliability of the scale and make it more useful to the field by facilitating the measurement of evaluation use in multi-site settings. The next section will review the literature on validity as it applies to the Evaluation Use Scale.

Validity

This section reviews the validity literature as it applies to the Evaluation Use Scale. Validity is 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" (Messick 1989a, p.6; 1995a, p. 5; 1995b, p. 741). This paper presents theoretical and empirical evidence to support the claim that scores on the Evaluation Use Scale adequately represent the construct of evaluation use in multi-site settings.

Conceptualizations of Validity Past and Present

In the early 20th century, the theoretical view of validity was that validity is the degree to which a test measures what it supposed to measure (Garrett, 1937, cited in Sireci, 2009). Although this characterization of validity is still seen in textbooks, scholars have recognized is as insufficient for decades. Messick (1980) wrote, "Different kinds of inferences from test scores require different kinds of evidence, not different kinds of validity" (p. 1014).

In the 1940s, a three-part conceptualization of validity comprising content, criteria and predictive effectiveness served as the organizing framework for validity discussions. The three–sometimes known as the holy trinity–were seen as distinct types of validity linked to different testing objectives, e.g., content validity for achievement tests, construct for personality tests, and predictive for selection testing (Goodwin & Leach, 2003; Shepard, 1993; Sireci, 2009).

In the early 1950s, the American Psychological Association's Committee on Test Standards recommended four distinct categories of validity: predictive validity, status validity, content validity, and congruent validity. These four categories later became four types of validity and congruent validity was re-named construct validity.

Cronbach and Meehl (1955) asserted that construct validity is involved any time a test is used to measure a latent construct, or an attribute or quality that is not operationally defined. Construct validity inspired the work of validity theorists and the contemporary notion that all validity is construct validity was set in motion.

In 1957, Loevinger wrote, "Since predictive, concurrent, and content validities are all essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view" (p. 636). Thus, the work of Cronbach and Meehl, as well as Loevinger, created the foundation for Messick's unitary concept of validity – that all validity is construct validity.

Unitary Concept of Validity

The present validation study employed a model grounded in the unitary concept of validity and based on a framework of the six aspects of construct validity established by Samuel Messick. This model guided the process of gathering and reporting empirical evidence and theoretical rationales related to each of six distinguishable aspects of construct validity to inform a judgment about the quality of the inferences based on data from the Evaluation Use Scale. (Messick, 1989a, 1989b, 1995a, 1995b, 1996a, 1996b). This scale, developed and implemented as part the NSF Program Evaluation Follow-up Survey, was administered online to nearly 400 participants in multi-site NSF-funded program evaluations, as described in the next chapter.

Construct validity is the unitary concept that encompasses the other forms of validity; and it is the “evidential basis for score interpretation’ (Messick, 199b, p. 743). The central premise of the unitary view of validity casts all types of validity as falling under the umbrella of construct validity (Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1989, 1995a, 1995b). Messick (1995a, 1995b) identifies six distinguishable aspects of construct validity as a means to address the central issues of validity as a unified concept. These aspects are interdependent strands of validity evidence; they are not separate and substitutable validity types. (Messick, 1995b).

Messick’s six aspects of construct validity stand as the framework for approaching a validation study under the unitary concept of validity. This framework guided the collection of multiple forms of evidence to establish the extent to which inferences based on the Evaluation Use Scale are adequate and appropriate for measuring evaluation use in multi-site evaluations (Messick, 1995b, p. 745). The components of the framework are (Messick 1995a, 1995b):

1. The content aspect of construct validity requires assessing content relevance and representativeness, specifying the boundaries of, establishing the theoretical

soundness of the construct, and assessing the consistency and reliability level of the scale.

2. The substantive aspect of construct validity refers to theoretical rationales for the observed consistencies in test responses and empirical evidence that the theoretical processes are actually engaged in by the respondents.
3. The structural aspect of construct validity assesses how well the scales statistical factors align with the underlying theoretical and rational structures.
4. The generalizability aspect of construct validity assesses the consistency with which the scale could measure the construct in other settings.
5. The external aspect of construct validity examines the extent to which the scale is able to differentiate between theoretically or rationally different groups and the extent to which it correlates as expected with other variables.
6. The consequential aspect of construct validity looking at whether test items are fair, unbiased, and useful, as well as considering possible ways the use or interpretations could potentially have negative consequences.

Argument-Based Approach to Validation

Kane (1992) proposed an argument-based approach to validation. This approach calls for building an argument in defense of the use of a test for a particular purpose. Kane’s approach has gained popularity and is recognized as a leading approach to addressing validation in many settings. Kane (2004) suggested that the current “unified view of validity, based on the construct model, provides an elegant theoretical framework for validity, but it does not provide an effective methodology for validation (p. 137).

The *2014 Standards of Educational and Psychological Testing* recognize the argument-based approach to validation, as did the 1999 edition, stating,

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses….The validity argument may indicate the need for refining the definition of the construct, may suggest revisions in the test…and may indicate areas needing further study, (2014, p. 21; 1999, p. 17).

Kane does not label the argument-based approach as a type of validity, and thus avoids getting caught up in the nomenclature debates (Kane, 1992). The argument-based approach to validation is consistent with the unitary concept of validity in that the six aspects of validity and the supporting evidence fit within the structure of an interpretative argument at the center of the argument-based approach to validation.

Consensus Position

Over the past several decades, the unified concept of validity – that all validity is construct validity – has gained wide acceptance (Lissitz & Samuelson, 2007; Sireci, 2007; Lissitz, 2009). And, though debate among measurement scholars and theorists continues, there have been some consensus shifts in the field related to measurement validity over the past several decades. Sireci (2007) wrote, “The unitary conceptualization of validity that has dominated the validity theory literature for decades … is theoretically sound and widely endorsed” (p. 478).

However, even as consensus developed, validity continued to be debated, re-defined, re-interpreted, and relabeled; and both the conceptualization and the nomenclature of validity have evolved in the measurement literature. (Lissitz & Samuelsen, 2007; Lissitz, 2009). The unitary concept of validity remains the consensus position in the validity literature today; and, as such, it is at the center of the debate and serves as the reality toward which challenges and criticism are directed (Borsboom, Mellenbergh, & van Heerden, 2004).

Critics of the unitary concept of validity include scholars who find construct validity confusing, difficult to describe, and impractical to apply in a real world situation. (Lissitz, 2009; Sireci, 2007). Lissitz (2009) observed that construct validity "flunks the grandmother test" – in reference to Einstein's reported comment: "You do not really understand something unless you can describe it to your grandmother." (p. 2)

Another area of dispute involves terminology. Measurement scholars and practitioners largely agree that validity is not an inherent property of a test, but instead refers to the appropriateness of inferences derived from a test (Mislevy, 2009). For decades, the *Standards for Educational and Psychological Testing* have embraced the understanding that validity is not a property of a test or instrument. The 1985 Standards stated, "Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores" (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1985, p. 9).

However, validity theorists and researchers take issue with this notion as incorrect, confusing, or both. Lissitz & Samuelson (2007) wrote, "We also find

Messick's assertion that validity cannot reside in the test to be essentially incorrect and confusing" (p. 442). Borsboom et al. (2009) charge that,

> Validity, as *normally understood* – that is, as it is understood by almost everybody except construct validity theorists—does patently *not* refer to a property of test score interpretations, but to a property of tests (namely, that these test measure what they should measure). (p. 138)

Terminology

Despite the lack of uniform agreement on terminology and nomenclature, it is important to clarify some frequently used terms in the validity literature for purposes of the present study. In addition to the common understanding of educational tests, the term "test" also refers to "any means of observing and documenting behaviors or attributes, including scales, instruments, and observations" (Messick, 1995b, p. 741). The term "score" or "test score" is not used only in the sense of an accumulation of points; but, rather, in its broadest sense to mean, any coding or summarization of observed consistencies or performance regularities on a test, questionnaire, observation procedure, or other assessment devices such as work samples, portfolios, and realistic problem simulations (Messick, 1995b, p. 741).

The approach taken in this study is one in support of the prevailing position that it is not the test or instrument that is validated. "Validity is not a property of a test or instrument as such, but rather of the meaning of the test scores…as well as any implications for action that this meaning entails" (Messick, 1995a, p. 5). Sireci (2007) highlights, "Any conceptualization of validity theory must acknowledge that what is to be validated is not a test itself but the use of the test for a particular purpose," (Sireci, 2007, p. 477).

Validation is the "process of gathering and reporting evidence to evaluate the use of a test for a particular purpose" (Sireci & Sukin, 2013, p. 61) and evidence refers to "data, or facts, and the rationale or arguments that cement those facts into a justification of a test-score inferences" (Messick, 1980, p. 1014).

The analysis described in the coming chapters uses the Messick's unitary validity framework, but does not take the step of developing an interpretive argument. Table 2 displays the research questions associated with each aspect of the unitary validation framework (Toal, 2009).

Chapter Two reviews the literature related to program evaluation use, influence, and validity. Chapter Three describes the methods used in developing and validating the Evaluation Use Scale for use in multi-site evaluation settings. Chapter Four presents the results of the analyses. Finally, Chapter Five weighs the evidence collected as part of the validation study and assesses the strength of the evidence for the validity of inferences about the use of evaluation in multi-site settings based on data collected using the Evaluation Use Scale.

Table 2.

Unitary Validation Framework for Evaluation Use Scale

Aspects of Construct Validity	Research Questions
Primary RQ	Is there sufficient evidence to support the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site evaluations?
Content	1a.To what extent is the construct of evaluation use theoretically sound? 1b. To what extent is the scale internally consistent and reliable?
Substantive	2a. To what extent does the scale include all relevant activities and processes? 2b. To what extent does the scale measure actual use?
Structural	3. To what extent do the scale's statistical factors align with the underlying theoretical and rational structures?
Generalizability	4. To what extent does the scale consistently measure use in other multi-site settings?
External	5a. To what extent does the scale correlate with expected activities? 5b. To what extent does the scale differentiate between rationally or theoretically different groups?
Consequential	6. To what extent could possible interpretations or uses of the scale result in a negative impact?

CHAPTER THREE
METHOD

This study used data collected by the Evaluation Use Grant (EUG) team as part of an NSF-funded evaluation research project. This small research team of faculty and graduate research assistants (including me) developed and fielded the survey described below. The EUG team employed a mixed methods design including a web-based survey completed by evaluators and project leaders in four NSF-sponsored, multi-site STEM education programs and the associated evaluation efforts. The EUG team also conducted in-depth interviews with a subset of survey respondents who had agreed to be considered for participation in follow-up interviews. The validation study was conducted by the author using the data collected by the EUG team. This chapter outlines the methods employed in the development and validation of the Evaluation Use Scale and describes the approach taken in gathering the evidence needed to consider the primary question in this validation study:

Is there sufficient evidence to support the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site evaluations?

The chapter begins with a brief description of the research setting and context of the study. It identifies the primary data sources, namely the “NSF Program Evaluation Follow-Up Survey” (developed and conducted by the EUG research team) as well as the project-level interviews that followed the survey. The chapter also describes the study participants, the data collection procedures used in the study, the development of the Evaluation Use Scale, and the analyses used in the validation study. All statistical analyses were conducted using SPSS, version 23.

Participants

Respondents to the NSF Program Evaluation Follow-up Survey were project leaders and evaluators from four multi-site NSF programs: (1) Advanced Technological Education (ATE), (2) Collaboratives for Excellence in Teacher Preparation (CETP), (3) Local Systemic Change Through Teacher Enhancement Program (LSC), and (4) Math & Science Partnership—Research, Evaluation, and Technical Assistance (MSP-RETA) program. All four of these NSF programs focused on STEM education. However, some important program differences between the MSP-RETA program and the other three programs bear discussion.

Table 3 identifies a few important terms with specialized usage in this context and provides definitions tailored for the EUG study.

Three of the four programs – ATE, CETP and LSC – included a program-level (national) evaluation and a project-level (local) evaluation in which they were encouraged (and, in one case, required) to participate. For purposes of this study, project participation refers to local project sites being involved in decision-making for the larger, overarching program-level evaluation. The ATE, *C*ETP, and LSC programs varied in their levels of project participation in decision making, demonstrating the additional complexity common in multi-site evaluations. The ATE program evaluation afforded more limited opportunities for project-level participation; the LSC program evaluation mandated project-level participation; and the CETP program offered opportunities, but did not require project-level participation in program evaluation decision-making.

By contrast, the MSP-RETA program did not conform to the national- and local-level evaluation structure. Instead, the MSP-RETA project at Utah State University provided evaluation technical assistance to other MSP projects. Also known as the Consortium for Building Evaluation Capacity, the Utah State MSP-RETA project worked with the MSP projects across the country to assist in identifying their project-level evaluation needs and, further, to help them develop and implement evaluations tailored to those needs in the MSP focus areas of partnership, teacher quality, challenging course offerings, evidence-based outcomes, and institutional change.

The MSP-RETA program also differed in the size of its representation in the EUG survey. The number of overall respondents to the survey from the MSP-RETA project was a small (n=56) representing only 15% of the all survey respondents (n=369), but of that number, only three identified themselves as evaluators. In addition, the MSP-RETA respondents had more missing data in their survey responses than respondents from the other three programs. This is likely due, in large part, to the different structure of that project and the activities of its participants, which rendered several of the survey items less applicable to them and therefore more difficult or answer. Because of these systematic differences between MSP-RETA and the other three programs (ATE, CETP, and LSC), the data from the MSP-RETA survey respondents and interviewees is used in a descriptive manner in this study, but the quantitative analyses excluded data from the MSP-RETA project.

Table 3.

Key Terms Related to the EUG Study

Term	*Definition*
Program	A program is a large national funding effort that comprises multiple, smaller funding efforts at multiple sites (e.g., the four NSF multi-site STEM education programs at the center of the present study).
Project	A project is one of the smaller funding efforts funded under a program or one of the sites in a multi-site program.
Project Leaders	Project leaders include respondents to the NSF Program Evaluation Follow-up Survey (EUG Survey) who are Principal Investigators, Co-Principal Investigators, or other non-evaluator staff leading efforts in one of the projects.

Procedures

The EUG team used project and meeting rosters obtained from evaluation personnel at the program level. Using those data, the EUG team compiled a list of names and email addresses for the entire population of project leaders and evaluators for each of the programs: ATE, CETP, LSC, and MSP-RETA. Where the email contact information provided was out-of-date or missing, a member of the EUG team searched the Internet for current email addresses. Ultimately, the population for this study totaled 935 project leaders and evaluators.

Two distinguishing characteristics of the design of the EUG study are important for understanding the dynamics of the survey and the interpretation of the findings. First, the study addressed large, multi-site evaluations where the *participants* were defined not as individual people but as individual *projects*, each implementing a different approach to solving a national problem at different sites across the country. Therefore, the data were collected from individuals (project leaders and evaluators) who served as representatives of their respective projects. The survey results are based on projects as the unit of analysis and are discussed using terms like *project-level use*, *project participation*, etc.

Second, this research focused on the use of evaluation by secondary, somewhat unintended, users. NSF was the primary user and client in each of the four program evaluations. However, the focus of the EUG study was not on NSF, but rather on the local projects and on the fields of STEM education and evaluation, regardless of whether the national program evaluators were instructed to consider project and field use in the design of the program evaluation.

Survey Process

The EUG team developed the NSF Program Evaluation Follow-up Survey (EUG Survey), a web-based survey of project leaders and evaluators in the four multi-site NSF-funded STEM education programs. EUG team members included two professors, both experts in evaluation, one post-doctoral associate with expertise in evaluation, and three evaluation doctoral students. The team followed an iterative survey development process, meeting weekly to discuss and further revise the survey over a period of approximately one year.

The EUG team sent emails inviting participation in the survey to all of the 935 project leaders and evaluators in the sample population. The emails were tailored to each program. The team staggered the email distribution by program to avoid technical problems that might have resulted from a single larger-scale mass email to respondents from all four programs (e.g., server overload, large number of non-deliverable emails).

The email described the purpose of the study, invited the recipient to complete the survey, and provided a unique user account name and password and a link to a University of Minnesota website for the recipient to access the web-based survey. The email also offered a small incentive for early participation and provided contact information for Drs. Lawrenz and King, as well as for the Research Subjects' Advocate Line at the University of Minnesota. The team sent email notices two weeks after the original email and again at four weeks after the original email to those who had not yet completed the survey.

Based on questions and feedback from the first group of respondents, the team modified the language of the email invitation to clarify and stress that this study was interested in the program-level (national) evaluation, rather than the project-level (local) evaluation; and, as such, the survey questions were aimed project-level activities. Also, based on feedback from this first group, the reminder emails were substantially shortened. The text of the email invitation and the follow-up reminder emails is in the appendix.

Data were collected from August 10, 2006 through January 17, 2007. Of the 935 emails sent, 215 (or 23%) were returned as undeliverable. Members of the EUG team followed up on each of the undeliverable emails using Internet searches and telephone research to find updated email contact information for just over half of the un-deliverable emails (109 of 215). Of the 109 new email addresses, 90 worked and 19 failed a second time. Thus, the final sampling frame included 810 project leaders and evaluators from the four NSF programs.

The overall response rate for the survey was 46% (372 of 810). Respondents came from the District of Columbia and Puerto Rico and all but three states (Arkansas, Kansas, and Wyoming). The respondent population was 54% male (437/810), 43% female (348/810), and 3% (25/810) did not respond to the gender question. Data provided by three respondents were deemed unusable due either to technical problems or to a clear misunderstanding of the survey's focus, as demonstrated by responses to the open-ended comments in Section VI of the survey. Ultimately, the team analyzed data from 369 respondents.

Table 4 shows the distribution of respondents by role (project leaders vs. evaluators) across the four programs (ATE, CETP, LSC, and MSP-RETA). Of the 369 total respondents, 306 were project leaders and 63 people were evaluators.

The survey design addressed the four major sources of error in survey design: sampling error, coverage error, measurement error, and non-response error (Dillman, 2000, p.11; Fowler, 2009, p. 16).

Table 4.

EUG Survey Respondents by Role and Program

Role	ATE	CETP	LSC	MSP-RETA	Total Respondents
Project Leaders	175	35	43	53	306
Evaluators	11	19	30	3	63
Total	186	54	73	56	369

First, the research design attempted to avoid sampling error (the result of surveying only some, not all, of the survey population) by surveying the entire population of project leaders and evaluators, rather than drawing a sample of that population.

Second, the EUG team sought to reduce coverage error (the result of all elements of the population not having an equal chance of being included) by choosing electronic survey distribution – reasoning that postal addresses were likely to be less accessible and up-to-date than email addresses. In addition, the team devoted substantial time and resources to finding replacements for all undeliverable email addresses in the original list.

Third, the survey development procedures aimed to minimize measurement error (the result of poor question wording or questions being presented in a way that results in inaccurate or uninterpretable answers) through careful and iterative item construction, external expert reviews, and the use of think-aloud interviews.

Finally, the EUG team assessed non-response error (which results when the people who responded are different from those who did not respond in some way that is relevant to the study) by conducting a non-response study.

Survey Non-response Analysis

To check for non-response bias, the EUG team contacted a subset of individuals in each of the four programs who had not responded to the web-based survey and asked them to answer three questions: (1) How involved were you in the evaluation? (2) How much impact did the evaluation have on you? (3) What kept you from responding to the initial survey request? A member of the EUG team drew a random sample of non-respondents in each of the programs using SPSS. The selected non-respondents received a personal, individually addressed email requesting information/input on the three questions in the non-respondent survey. These potential respondents received up to two reminder emails and a telephone call, if the email communication did not result in a response.

Table 5 displays a comparison of respondents and non-respondents in terms of overall involvement and use by program. The respondents' scores are the means of the involvement items and the use items in the original survey. The non-respondents' mean scores are the means of answers to the questions they responded to via email as part of the non-respondent study. Although these measurements are not identical, the two sets of means mirror one another, both using a 4-point response scale. In the non-response study, the label "use/impact" distilled multiple items from the larger survey into a single score encompassing questions about use of the process, instruments, and results of the evaluation.

For ATE and CETP, the mean scores for both involvement and use were higher for respondents than for non-respondents. For both LSC and MSP, the involvement mean was higher for non-respondents than for respondents, and the use mean was lower for non-respondents than for those who responded to the initial survey.

Table 5.

Summary of Non-Response Study Comparisons

	Involvement Mean			Use/Impact Mean		
NSF Program	Respondents	Non-Respondents	p-value	Respondents	Non-Respondents	p-value
ATE	2.39	2.36	.91	2.50	2.06	.03
CETP	2.38	1.82	.07	2.52	1.76	.01
LSC	2.71	3.14	.15	2.85	2.69	.70
MSP	1.76	1.95	.49	2.02	1.90	.48

Of the 89 total responses to the non-respondent survey across all four programs, almost 25% reported that they did not feel the survey applied to them, and nearly 24% did not remember having received the initial survey. In addition, approximately 18% cited a lack of time, and 5% reported having had technical problems. Over 20% offered a range of specific reasons in the “other” category, the basic themes included: a) no longer working with the project, b) confused this survey with another survey from the still-ongoing ATE program; c) uncertainty within their project about who would respond to the survey, and d) personal reasons, e.g., a death in the family. Approximately 8% did not answer this question.

The non-respondent study showed no significant differences in the reported overall level of involvement for each of the four programs. However, it did show a significant difference between respondents and non-respondents from the ATE and CETP evaluation projects regarding the impact of the evaluations. In both cases the respondents reported higher levels of impact than non-respondents, which may indicate the possibility of upward bias in the estimates of overall involvement and use/impact for survey respondents, or it may highlight the possibility of an overlap in interpretation of the terms use, influence, and impact. The items used in the non-respondent study are included in the appendix.

Project Interviews

The EUG team also conducted individual telephone interviews with project leaders and evaluators from each of the four NSF programs. These follow-up interviews served to gather more in-depth information about the scope and type of involvement in and use of the program level evaluations. Drawing from a pool of survey respondents who indicated that they were willing to be contacted for a follow-up interview, the EUG team purposefully selected two samples from the group of project leaders and evaluators.

One sample contained respondents with varying reported levels of involvement and use based on the calculated mean of each individual’s responses on the involvement and use items in the survey, i.e., high involvement/high use, high involvement/low use, etc. The second sample of interviewees included those drawn exclusively from survey respondents who reported high levels of use.

The team conducted 29 interviews or approximately six to eight interviews per program. The interviews lasted between 30-50 minutes; and all were taped and transcribed. The interviews covered the main topics of involvement in and use of program evaluations, as well as motivation for involvement, perceived influence on the evaluation, and ways in which evaluations were used. The interviews followed a standardized, semi-structured protocol with flexibility in probing for additional details and description as necessary. The interview protocol and consent form are included in the appendix.

Scale Development

To measure the two principal constructs – involvement in evaluation and evaluation use – the survey includes seven sections: (1) respondents' involvement in the program evaluation; (2) respondents' influence on the evaluation; (3) the influence of the evaluation on the respondents' knowledge, skills, and beliefs; (4) the use of knowledge, products, and processes from the evaluation; (5) the use of evaluation findings, (6) open-ended questions, and (7) demographics.

In each section, the questions follow a logical structure based on three stages of an evaluation: (1) planning (activities before the evaluation formally begins, discussing the focus of the evaluation, suggesting individuals to participate in the evaluation planning team, and developing the evaluation plan), (2) implementation (activities during the evaluation, developing data collection instruments and processes, analyzing data, reviewing collected data for accuracy and/or completeness, interpreting data), and

(3) communication (developing future project plans, writing and reviewing evaluation reports, and presenting evaluation findings) (Burke, 1998).

Planning the content of a scale requires clear thinking about the construct(s) to be measured, and “theory is a great aid to clarity” (DeVellis, 2003, p. 60). The EUG team engaged in extensive and ongoing literature review in developing the theoretical and rational foundations for the development of the Evaluation Use Scale and the Evaluation Involvement Scale.

A review of the existing literature, an initial brainstorming session that yielded more than 200 possible items, and months of iterative testing and revising yielded a final survey instrument with 69 items: 65 fixed response items (including an opt-in for respondents willing to be interviewed), and 4 open-ended questions. To promote ease of completion, the EUG team incorporated input from experts in web-based survey layout and graphic design, and the instructions and section headers in the survey were tailored to each of the four programs.

Evaluation Use Scale

The extensive evaluation use literature, as well as the emerging evaluation influence literature (Kirkhart, 2000; Mark & Henry, 2004) grounded the development of the Evaluation Use Scale. The EUG team’s in-depth empirical literature review on evaluation use (Johnson et al., 2009) informed the process; and the team incorporated extensive input from experienced evaluators and evaluation researchers to inform the development of the scale items aimed at measuring the construct of evaluation use. The scale items asked survey respondents about the extent to which the evaluation increased knowledge, understanding, skills, and beliefs about evaluation, as well as about the use of knowledge and skills in future evaluations.

Table 6 presents the 20 items in the Evaluation Use Scale. Response options for each item included: *No; Yes, a little; Yes, some; and Yes, extensively.*

Table 6.

Evaluation Use Scale Items (20 items)

The evaluation increased my knowledge/understanding of …

- How to plan an evaluation (e.g., discussing the focus of the evaluation, identifying evaluation planning team members, developing the evaluation plan).
- How to implement an evaluation (e.g., developing data collection instruments and processes, collecting, analyzing, reviewing, and interpreting data).
- How to communicate evaluation findings (e.g., developing future plans for your project, writing and reviewing evaluation reports, and presenting evaluation findings).
- Science, Technology, Engineering, and Math (STEM) education evaluation.
- My project.

The evaluation improved my skills in . . .

- In planning an evaluation (e.g., discussing the focus of the evaluation, identifying evaluation planning team members, developing the evaluation plan).
- In implementing an evaluation (e.g., developing data collection instruments and processes, collecting, analyzing, reviewing, and interpreting data).
- In communicating evaluation findings (e.g., writing and reviewing evaluation reports, presenting evaluation findings)
- As a STEM education evaluator.
- For working on my project.

The evaluation increased my belief in the importance of . . .

- Planning an evaluation (e.g., discussing the focus of the evaluation, identifying planning team members, developing the evaluation plan).
- Implementing an evaluation (e.g., developing data collection instruments and processes, collecting, analyzing, reviewing, and interpreting data).
- Communicating evaluation findings (e.g., developing future plans for your project, writing and reviewing evaluation reports, and presenting evaluation findings).
- STEM education.
- STEM education evaluation.

I used …

- What I learned from planning this program evaluation in another evaluation.
- The evaluation plan from this program evaluation as a model in another evaluation.
- What I learned from implementing this program evaluation in another evaluation.
- Data collection instruments from this program evaluation in another evaluation.
- What I learned from communicating evaluation findings in another evaluation.

Involvement Scale

Like the Evaluation Use Scale, an extensive review of the literature informed the development of the survey items aimed at measuring the construct of involvement (the evaluation involvement scale). The traditional notions of participatory evaluation – broadening the decision-making base and/or reallocating power in the course of conducting an evaluation (Cousins & Whitmore, 1998) – do not apply to large-scale, multi-site evaluations where stakeholders are not able to meet and contribute to decision making due to the many layers of stakeholders and the number of widely dispersed sites that are frequently unfamiliar with each other.

Both Cousins and Whitmore's (1998) systematic collaborative inquiry and Burke's (1998) key decision points in evaluation support the concept of project involvement as defined by Lawrenz and Huffman (2003) and underpin the development of the scale to measure the extent of evaluation participation in large-scale, multi-site evaluations.

Burke's (1998) key decision points include: deciding to do the evaluation; assembling the team; making the evaluation plan; collecting the data; synthesis, analysis, and data verification; developing future plans; and dissemination and use of outcomes and reports. Based on this framework, the EUG team assigned these decision points to one of three evaluation stages: planning, implementation, or communication of results.

Table 7 shows the items in the Evaluation Involvement Scale, the response options for which included: *No; Yes, a little; Yes, some; Yes, extensively; and I don't think this activity took place.*

Table 7.

Evaluation Involvement Scale Items (11 items)

- I was involved in the discussions that focused the evaluation.
- I was involved in identifying evaluation planning team members.
- I was involved in developing the evaluation plan.
- I was involved in developing data collection instruments.
- I was involved in developing data collection processes.
- I was involved in collecting data.
- I was involved in reviewing collected data for accuracy and/or completeness.
- I was involved in analyzing data.
- I was involved in interpreting collected data.
- I was involved in writing evaluation reports.
- I was involved in presenting evaluation findings (e.g., to staff, stakeholders, external audience).

Think-Aloud Interviews

Near the end of the survey development process, a researcher from the EUG team conducted individual "think-aloud" (or concurrent cognitive) interviews with three doctoral students studying evaluation at the University of Minnesota. In a think-aloud interview, individuals complete a questionnaire while telling the interviewer what they are thinking as they do so. The interviewer takes notes on how the survey items are being interpreted and whether the intent of each question is understood. At the end, the interviewer may also ask a series of questions prepared in advance (Dillman, 2007).

As the volunteer respondents verbalized their thought processes, the EUG team interviewer took notes on their reactions, comments, points of confusion, etc. Then, the interviewer asked a series of questions prepared in advance to elicit information to inform the research team about areas for potential improvement in the questionnaire items. One of these questions was related specifically to validation of the use scale – "Did you have questions related to the concept of use?" The specific questions and associated responses from the three people interviewed are available in the appendix.

Expert Review

The EUG team worked on survey development for almost a year, extensively reviewing and discussing the best way to structure the questionnaire. The EUG team's two internal evaluation experts, Dr. Frances Lawrenz and Dr. Jean King, provided input throughout the process. Both are established evaluation experts, and Dr. Lawrenz also has four decades of experience as an NSF evaluator with knowledge of each of the four program-level evaluations being studied. The expertise of these two leaders of the EUG team ensured the relevance and utility of the items included in the survey.

External experts also reviewed the questionnaire before the survey instrument was finalized. The EUG researchers sent the draft instrument to the national program evaluators for the ATE, LSC, and MSP-RETA programs (Arlen Gullickson, Iris Weiss, and Catherine Callow-Heusser) as well as to a EUG advisory board member and well-known expert evaluator, Michael Quinn Patton. Each of these experts reviewed the instrument and sent written feedback, suggestions, and questions to the EUG team. An additional outside expert, Marvin C. Alkin, Ph.D. (who also served as the meta-evaluator for the NSF grant) provided a final external review and suggested minimal changes to the instrument before the wording was finalized. The appendix provides brief biographical information for each of the external expert reviewers who contributed to the design of the survey content.

Validity Analysis

The validation of the Evaluation Use Scale followed Messick's unitary concept of validity and included identifying and gathering evidence of validity across the six aspects of construct validity. The six aspects of Messick's framework function together constituting complementary forms of validity evidence, not representing separate different types of validity (1989a, 1989b, 1995a, 1995b, 1996a, 1996b). In addition, this analysis also includes gathering evidence for two test design components – theoretical soundness and internal consistency.

Table 8 presents the framework for the validation study including: (1) the research questions associated with each of Messick's aspects of validity, (2) the validity evidence sources; and (3) the analysis methods employed for each research question. This framework guided the design and evidence gathering required by the validation study of the Evaluation Use Scale. The findings are detailed in the next chapter. The final chapter examines the relative strength or weakness of the evidence and assesses the degree to which it supports the adequacy and appropriateness of inferences about evaluation use in multi-site settings.

Table 8.

Evaluation Use Scale Validation Analysis Framework

Aspect of Validity/ Research Questions	Source of Validity Evidence	Analysis
Content Aspect 1a. To what extent is the construct of evaluation use theoretically sound? 1b. Is the Evaluation Use Scale internally consistent and reliable?	Literature Base Think-Alouds Expert Judgment Survey Data Project Interviews	Review published literature Describe think-aloud findings Calculate item variance and scale alpha Compare interview responses and scale mean scores
Substantive Aspect 2a. To what extent does the scale include all relevant activities and processes? 2b. To what extent does the scale measure actual use?	Literature base Expert judgment Project Interviews Scale Items	Describe expert views Compare interview examples of use and survey responses Compare evaluation use activities documented in the literature with scale items.
Structural Aspect 3. To what extent do the scale's statistical factors align with the underlying theoretical and rational structures?	Survey Data	Exploratory factor analysis (EFA) Calculate reliability (alpha) of new factors
Generalizability Aspect 4. To what extent does the scale consistently measure use in other multi-site settings?	Survey Data	Compare reliability (coefficient alpha) and scale means
External Aspect 5a. To what extent does the scale correlate with expected activities? 5b. To what extent does the scale differentiate between rationally or theoretically different groups?	Involvement and Use Items Survey Data	Conduct One-way ANOVA, mean comparison of use by role Correlation between use and involvement
Consequential Aspect 6. To what extent could possible interpretations or uses of the scale result in a negative impact?	Scale Items Range of Sources	Correlation between use and involvement Explore possibilities

IRB Approval

This study is exempt from full IRB review under two federal guidelines: 45 CFR Part 46.01 (b) Category 2 Surveys/Interviews and Category 4. Data for the present study (#1510E79121) were collected for the “Beyond Evaluation Use: Determining the Effects of Project Participation on the Influence of NSF Evaluations” grant research conducted at the University of Minnesota.

CHAPTER FOUR
FINDINGS

This chapter presents the results for each research question, framed according to the six aspects of construct validity. These aspects function as general validity criteria and structure the approach to this validation study (Messick, 1995a). The extent to which these findings are responsive to the primary research question – *To what extent does the evidence gathered in this study support the Evaluation Use Scale as measure of evaluation use in multi-site evaluations?* – will be discussed in the final chapter.

Content Aspect of Construct Validity

The content aspect of construct validity addresses the relevance and representativeness of the content of a test or other measurement. Gathering validity evidence under this aspect involves specifying the construct's boundaries and establishing its theoretical soundness. Two supporting research questions guided the exploration of validity evidence in the content aspect of construct validity. First, the question of the theoretical soundness of the construct was addressed:

Question 1a: To what extent is the construct of evaluation use theoretically sound?

Establishing the soundness of the construct involves reviewing the theoretical literature and incorporating the review and judgment of content experts in the field (Question 1a). Then, when the existence of the construct is established, examining the extent to which the scale reliably measures the construct provides a description of the quality of the measure (Question 1b). This analysis of reliability involves conducting statistical tests of the scale’s internal consistency and reliability as detailed in the next section of this chapter.

Evaluation use literature

Researchers and evaluators have studied the use of evaluations for decades (Christie, 2007; Cousins & Leithwood, 1986; Fitzpatrick et al., 2004; Johnson et al., 2009), and theorists assert that evaluation use is fundamental to a theoretical understanding of evaluation (Shadish et al., 1991). Demonstrating that it is indeed a recognized construct, evaluation use is in the top five most frequently addressed topics in the Research on Evaluation articles published in the *American Journal of Evaluation* since 1998 (Vallin, Phillipoff, Pierce, & Brandon, 2015). In addition, the growing body of published peer-reviewed literature on evaluation influence supports the claim that evaluation influence is also a recognizable construct in the evaluation field today (Kirkhart, 2000; Henry & Mark, 2003; Mark & Henry, 2004) and that it is likely to be discussed, examined, and deconstructed by researchers in the coming years.

Expert judgment

Two established and well-regarded evaluation experts, Dr. Jean King and Dr. Frances Lawrenz, provided expert review and input related to the construction of the instrument, lending credibility to the process and support for the quality of the content, and contributing to the relevance and utility of the Evaluation Use Scale. Each iteration of instrument development by the EUG team incorporated expert feedback on what items should be amended, augmented, or deleted. Final decisions on content inclusion or exclusion were made by Drs. Lawrenz and King.

Additional experts from each of the involved NSF programs reviewed and commented on the survey instrument before it was finalized. The expert review and judgment is an important component of establishing content-related validity evidence.

Think-aloud interviews

In the process of constructing the survey instrument used in this study, individuals familiar with the field of evaluation participated in an individual “think-aloud” process with a member of the EUG team. This exercise involved volunteers with some knowledge of the evaluation field completing the survey instrument while verbalizing their thoughts, questions, concerns, and uncertainties related to the instrument, its content, the flow of the survey, and any other input relevant to the survey instrument.

The volunteers recruited to participate in the think-aloud interviews were current graduate students in evaluation students or those with experience in evaluation practice. The EUG team identified this profile as the desired target population for the think-aloud interviews to ensure that their feedback was grounded in a base of knowledge about the field, rather than from a novice perspective. Because the interpretations of individuals familiar with program evaluation concepts was important, the design of the think-aloud process prioritized individuals with education and/or experience in program evaluation over those with STEM education experience.

The findings from the think-aloud interviews revealed that all three participants recognized and understood the terms used in the survey and understood that there are different ways in which a person can use an evaluation. None of the participants raised any concerns or questions about the concept of evaluation use. The lack of confusion or questions among the think-aloud interviewees who were involved in the evaluation field suggests that the construct of evaluation use is recognized and understood in the evaluation field. A summary of the think-aloud process, the prompts offered by the researchers, and the responses and feedback provided by the interviewee are provided in the appendix.

Project interviews

Post-survey follow-up interviews with survey respondents who agreed to participate, yielded examples of responses or comments from the interviewees describing examples of how they had used evaluation findings (which can also include evaluation processes or products). The interviewees readily talked about evaluation use and demonstrated their understanding of the concept; they did not hesitate or seek clarification related to the term evaluation use.

The project leaders and evaluators who participated in the interviews shared feedback and examples of evaluation use in their projects. One interviewee offered a specific example saying: “We used the instruments to develop teacher training within our project… [W]e ended up using the instruments as a way of helping teachers understand what a good lesson looked like.” Another interviewee made a broader statement about the overall influence of the evaluation on his/her project: “I think it had significant impact on how we went about doing evaluation of the project.”

Question 1b: To what extent is the Evaluation Use Scale internally consistent and reliable?

Quantitative analysis of survey data collected with the Evaluation Use Scale assessed the internal consistency and reliability of the scale. This analysis provides information to assess the internal reliability and consistency of the Evaluation Use Scale by illustrating how well the items that reflect the same construct yield similar results. Table 9 displays the item variances, distributions, and inter-scale correlations, and Table 10 displays the inter-item correlations of the Evaluation Use Scale.

Table 9 is the Item Correlation Matrix which presents the inter-item correlations illustrating the extent to which scores on one item relate to scores on all other items in a scale. All items in the Evaluation Use Scale are designed to measure evaluation use, so all items in the scale are included in the calculation of the correlation between each pair of items in the scale. The average inter-item correlation represents the mean of all the correlations. This analysis provides data to inform an assessment of the extent to which the items in the Evaluation Use Scale reflect the same construct. Strong inter-item correlations among scale items suggest that the scale items represent a single construct and comprise a reliable scale.

Table 10 displays the item statistics for the Evaluation Use Scale, including item means, standard deviations, and corrected item-total correlations. The Corrected Item-Total Correlations represent the correlation between each individual item and the total score from the scale (minus itself). A high degree of correlation between each item and the scale overall is characteristic of strong scale reliability.

Table 9.

Evaluation Use Scale Item Correlation Matrix

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.00																			
2	0.82	1.00																		
3	0.81	0.80	1.00																	
4	0.69	0.70	0.72	1.00																
5	0.62	0.68	0.65	0.65	1.00															
6	0.75	0.73	0.74	0.67	0.68	1.00														
7	0.73	0.82	0.74	0.66	0.65	0.79	1.00													
8	0.72	0.72	0.85	0.72	0.65	0.79	0.79	1.00												
9	0.66	0.70	0.70	0.80	0.68	0.76	0.77	0.80	1.00											
10	0.60	0.66	0.63	0.66	0.79	0.70	0.67	0.71	0.75	1.00										
11	0.58	0.49	0.53	0.51	0.40	0.61	0.50	0.55	0.50	0.49	1.00									
12	0.54	0.47	0.49	0.46	0.36	0.54	0.47	0.53	0.45	0.48	0.92	1.00								
13	0.62	0.51	0.61	0.56	0.46	0.60	0.53	0.64	0.52	0.55	0.89	0.88	1.00							
14	0.45	0.39	0.51	0.60	0.52	0.48	0.41	0.52	0.51	0.53	0.51	0.51	0.56	1.00						
15	0.49	0.47	0.53	0.63	0.50	0.58	0.46	0.54	0.53	0.49	0.71	0.68	0.71	0.80	1.00					
16	0.66	0.60	0.59	0.51	0.50	0.66	0.60	0.63	0.60	0.50	0.49	0.48	0.46	0.38	0.43	1.00				
17	0.57	0.53	0.53	0.44	0.43	0.53	0.59	0.59	0.56	0.45	0.44	0.45	0.43	0.47	0.41	0.74	1.00			
18	0.55	0.65	0.62	0.53	0.50	0.61	0.66	0.68	0.63	0.56	0.39	0.44	0.44	0.36	0.39	0.75	0.74	1.00		
19	0.38	0.45	0.37	0.37	0.35	0.42	0.51	0.50	0.52	0.37	0.22	0.25	0.23	0.23	0.23	0.59	0.66	0.68	1.00	
20	0.56	0.49	0.62	0.51	0.41	0.58	0.56	0.70	0.59	0.46	0.49	0.50	0.55	0.46	0.49	0.73	0.71	0.72	0.53	1.00

Table 10.

Descriptive Statistics for Evaluation Use Scale Items

Survey Item #	Evaluation Use Scale Item	Mean	Standard Deviation	Corrected Item-Total Correlation
25	Knowledge/understanding of evaluation planning	2.79	1.03	0.80
26	Knowledge/understanding of evaluation implementation	2.86	0.99	0.79
27	Knowledge/understanding of communicating evaluation findings	2.66	1.04	0.82
29	Knowledge/understanding of STEM evaluation	2.84	0.97	0.78
30	Knowledge/understanding of my project	2.99	0.97	0.71
31	Skills in planning an evaluation	2.71	1.06	0.83
32	Skills in implementing an evaluation	2.78	0.95	0.81
33	Skills in communicating evaluation findings	2.64	1.08	0.87
35	Skills as a STEM education evaluator	2.75	1.00	0.82
36	Skills for working on my project	2.80	1.00	0.75
38	Belief in the importance of planning an evaluation	2.80	1.12	0.70
39	Belief in the importance of implementing an evaluation	2.82	1.11	0.68
40	Belief in the importance of communicating evaluation findings	2.84	1.11	0.74
41	Belief in the importance of STEM education	2.38	1.16	0.63

42	Belief in the importance of STEM evaluation	2.62	1.16	0.69
45	Used what I learned from planning the evaluation in another evaluation	2.82	1.02	0.74
46	Used evaluation plan as a model in another evaluation	2.29	1.13	0.70
47	Used implementation knowledge in another evaluation	2.68	1.00	0.74
48	Used data collection instruments in another evaluation	2.43	1.09	0.53
49	Used communication info in another evaluation	2.42	1.12	0.73

Note: Response options were: 1 = No; 2 = Yes, a little; 3 = Yes, some; 4 = Yes, extensively.

Table 11 addresses the reliability of the Evaluation Use Scale in terms of coefficient alpha, a measure of internal consistency. Coefficient alpha describes the scale items in terms of how closely related they are as a group. It represents the proportion of total item variance due to both the true score and error.

A high coefficient alpha indicates internally consistent scale items that measure the effect of one latent (unobserved) variable. Coefficient alpha will increase as the number of items measuring the same trait increases (DeVellis, 2003). A high coefficient alpha could also potentially provide some support for removing items if that is called for based on other analyses.

Table 11.

Evaluation Use Scale Reliability

Coefficient Alpha	Coefficient Alpha Based on Standardized Items	N of Items
.963	.964	20

Substantive Aspect of Construct Validity

Adding to the expert judgment of test content considered in the first aspect of construct validity, the substantive aspect of construct validity identifies “the need for empirical evidence of response consistencies or performance regularities reflective of domain processes” (Messick, 1995a, p. 6). A valid scale should include an effective sampling of tasks, and it should also provide evidence that the processes included are demonstrated by respondents (Messick, 1995a, 1995b).

In the case of the Evaluation Use Scale, the validity evidence should demonstrate that the dimensions of evaluation use included in the scale capture actual evaluation use. Two supporting research questions guided the exploration of this aspect of validity. First, the question of the adequacy of coverage of the construct by the items included in the scale was addressed with the following research question:

Question 2a. To what extent does the Evaluation Use Scale include all relevant activities and processes of evaluation use?

This aspect of construct validity examines the extent to which a test or an instrument includes all relevant processes and tasks to effectively measure the construct of interest. The sources of evidence to demonstrate adequate coverage of the content domain include: expert judgment, qualitative data, and content from the published literature. (Brualdi, 1999; Messick, 1989b).

Expert judgment

Again, the EUG team included two experienced and well-regarded evaluation experts, Dr. Frances Lawrenz and Dr. Jean King, provided expert review and input related to the construction of the instrument, including the identification of the relevant evaluation use-related dimensions for inclusion in the Evaluation Use Scale. Both of these individuals have an extensive and far-reaching grasp of the evaluation research literature, in addition to their own experience in teaching, conducting research on evaluation, as well as in practicing evaluation in a range of settings, including NSF-funded evaluation programs. As such, they are well-qualified to provide insightful and accurate assessments of the extent to which the Evaluation Use Scale was inclusive of all relevant activities and processes involved with evaluation use.

Project interviews

Data were gathered from a number of telephone interviews with program participants to understand the experiences of the projects participating in these large national programs. As part of the examination of the extent to which the Evaluation Use

Scale includes all relevant evaluation use content, analysis of the project interview transcripts yielded evidence about the extent to which the items in the Evaluation Use Scale matched the actual evaluation use activities and experiences described by the interviewee.

Data analysis included comparing the relevant project interview content to the evaluation use topics represented in the Evaluation Use Scale and assessing the extent to which the topics or domains represented in the Evaluation Use Scale were also described or touched upon by interviewees. The overlap or lack of overlap between the two serves as evidence that can be brought to bear in addressing questions about the extent to which the Evaluation Use Scale is inclusive and representative of all the relevant domains.

A review of the qualitative data collected during telephone interviews with survey participants revealed some highlights of participants' experiences relative to evaluation use based on their descriptions of their activities consistent with the domains of evaluation use included in the Evaluation Use Scale. Overall, the interviewees' examples of evaluation use activities matched items or types of evaluation use included in the Evaluation Use Scale. Mismatches occurred in one of two ways: 1) items in the scale not mentioned by interviewees as examples of evaluation use, or 2) items not in the scale being mentioned by interviewees as examples of evaluation use.

Selected interview quotations matched to scale items

Below are some illustrative examples of project interviewee quotations that illustrate items included in the Evaluation Use Scale:

- Increased knowledge and skills for evaluation planning: "*…there's no question, we improved our evaluation techniques and instruments over time compared to when we first started."*
- Increased knowledge and skills for evaluation implementation: "*They educated us about some components of evaluation that we were relatively unfamiliar with, things like logic models, for example."*
- Increased knowledge and skills for communicating about evaluation: "*It turned out to be useful to us to have those papers to share with other audiences."*
- Belief in importance of planning, implementing evaluations: "*I think people are more aware of the importance of doing evaluation."*
- Increased belief in the importance of evaluation: "*The opportunity to meet other projects, to exchange ideas with other projects… will have a lasting impact on those that are out in the field."*
- Use in future evaluations: "*I became more sensitive to the complexity and difficulties in doing program-wide evaluations in terms of handling the uniqueness and the individual characteristics of the projects."*
- Used data collection instruments in another evaluation: "*We know that the NSF likes the surveys that were created. And so, in our current work, we decided to use the HRI surveys for middle school and high school in mathematics."*

Question 2b. To what extent does the Evaluation Use Scale measure actual evaluation use?

The descriptive statistics in Table 12 show the distribution of survey responses relative to each item in the Evaluation Use Scale. Survey respondents were asked to respond with the level of use of evaluation in terms of improving their knowledge and skills about evaluation, their belief in the importance of different aspects of evaluation, and their future use of evaluation products.

The response distribution data illustrate that two-thirds of respondents or more reported at least a little evaluation use for every item in the Evaluation Use Scale. This finding provides evidence for an assessment of the extent to which the items in the scale represent activities in which the respondents are actually engaged. Respondents indicated that an increased belief in the importance of STEM education was the most frequently reported as not having happened at all. Similarly, an increased belief in the importance of planning an evaluation was the most frequently reported as having happened extensively. Overall, a strong majority (not less than 67% of respondents) reported that that activity happened at least a little for each item in the scale.

Table 12.

Evaluation Use Scale Item Distributions

#	Item	No	A Little	Some	Extensively	N
Increased knowledge/understanding of…						
25	Evaluation planning	63	67	112	68	310
		20.3%	21.6%	36.1%	21.9%	100.0%
26	Evaluation implementation	53	74	113	69	309
		17.2%	23.9%	36.6%	22.3%	100.0%
27	Communicating evaluation findings	70	73	109	57	309
		22.7%	23.6%	35.3%	18.4%	100.0%
29	STEM education evaluation	57	67	122	60	306
		18.6%	21.9%	39.9%	19.6%	100.0%
30	My project	53	51	116	87	307
		17.3%	16.6%	37.8%	28.3%	100.0%
Increased skills …						
31	In planning an evaluation	63	78	102	64	307
		20.5%	25.4%	33.2%	20.8%	100.0%
32	In implementing an evaluation	54	83	109	61	307
		17.6%	27.0%	35.5%	19.9%	100.0%
33	In communicating evaluation findings	72	71	104	58	305
		23.6%	23.3%	34.1%	19.0%	100.0%
35	As a STEM education evaluator	67	77	107	53	304
		22.0%	25.3%	35.2%	17.4%	100.0%
36	For working on my project	57	67	115	68	307
		18.6%	21.8%	37.5%	22.1%	100.0%
Increased belief in the importance of …						
38	Planning an evaluation	56	63	93	98	310
		18.1%	20.3%	30.0%	31.6%	100.0%
39	Implementing an evaluation	58	52	104	93	307
		18.9%	16.9%	33.9%	30.3%	100.0%
40	Communicating evaluation findings	64	58	98	90	310
		20.6%	18.7%	31.6%	29.0%	100.0%
41	STEM education	102	60	92	54	308
		33.1%	19.5%	29.9%	17.5%	100.0%
42	STEM education evaluation	82	62	95	69	308
		26.6%	20.1%	30.8%	22.4%	100.0%

#	Item	No	A Little	Some	Extensively	N
Used … in another evaluation						
45	What I learned from planning the evaluation	22	32	54	45	153
		14.4%	20.9%	35.3%	29.4%	100.0%
46	Evaluation plan as a model	51	30	46	26	153
		33.3%	19.6%	30.1%	17.0%	100.0%
47	Implementation knowledge	22	41	54	35	152
		14.5%	27.0%	35.5%	23.0%	100.0%
48	Data collection instruments	44	27	56	26	153
		28.8%	17.6%	36.6%	17.0%	100.0%
49	Communicating evaluation findings	42	37	42	32	153
		27.5%	24.2%	27.5%	20.9%	100.0%

Note: Response options were: 1 = No; 2 = Yes, a little; 3 = Yes, some; 4 = Yes, extensively.

Structural Aspect of Construct Validity

Question 3: To what extent do the Evaluation Use Scale’s statistical factors align with the underlying theoretical and rational structures?

Factor analysis serves a number of purposes, including:

- Guiding the determination of the number of latent variables underlying a set of test items;
- Explaining variation among a set of variables by reducing the number of variables to a smaller number of factors; and
- Defining the content or meaning of the factors that account for variation among the larger set of items (DeVellis, 2003).

Consequently, factor analysis is an effective means of connecting the statistical factors of a scale to the underlying theoretical and rational structures; and factor analysis also yields data about the scale and its component members that can contribute to the evidence needed to complete a validation study.

Exploratory Factor Analysis (EFA) of the 23 evaluation use-related items in the EUG survey used Principal Axis Factoring (PAF) and Varimax rotation and was conducted in SPSS version 23. An alternative EFA approach is to conduct Principal Components Analysis (PCA). PCA is useful when the primary goal is item reduction. This method was not selected, however, as items reduction was not the primary purpose.

Table 13.

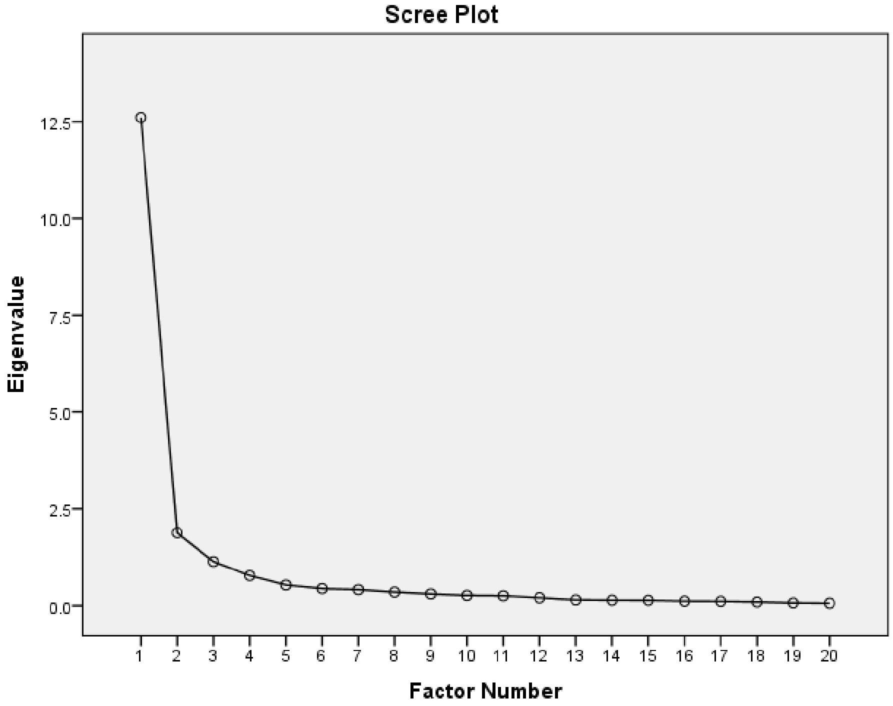
Eigenvalues for Evaluation Use Scale EFA

Evaluation Use Factor	Eigenvalues	% of Variance	Cumulative %
#1 Increased Knowledge, Skills, and Understanding	12.608	63.042	63.042
#2 Increased Belief in Importance of Evaluation Components	1.877	9.386	72.428
#3 Increased Future Use	1.132	5.661	78.089

The PAF analysis supported an assessment of the internal consistency of the Evaluation Use Scale. It also enabled the identification of the underlying factor structure of the Evaluation Use Sale and a demonstration of support for the contention that the scale is measuring a single trait.

The Eigenvalues greater than 1.0 supported the decision to retain three factors with values above 1.0 (i.e., 12.61, 1.88, and 1.13) as shown in Table 13. The cumulative percentage column indicates that these three factors explain 78.1% of the variance. The scree plot (Figure 4) also indicates that three factors should be retained, based on the bends in the curve.

Figure 4.
Scree Plot



Items that clearly loaded onto one of the factors were defined as those that showed at least a 0.40 factor loading and where there was at least a 0.20 factor loading difference across factors. These items were retained as part of the Evaluation Use Scale.

Items that did not meet these criteria were deemed not to have loaded clearly onto one of the factors. These items were, therefore, eliminated from the scale. The details of the factor loading are provided in the Rotated Factor Matrix in the appendix.

Items Removed from Evaluation Use Scale

This application of these selection criteria supported the removal of three of the 23 items, resulting in a 20-item Evaluation Use Scale. These three items did not fully load onto the three factors that emerged from the factor analysis. The following three evaluation use-related items on the EUG survey did not become part of the Evaluation Use Scale. They were removed from the scale subsequent to factor analysis (principal axis factoring) where they did not clearly load onto one of the three factors that emerged. The three removed items are:

1) Evaluation increased my knowledge/understanding of STEM education (item # 28 in the EUG survey questionnaire)
2) Evaluation improved my skills as a STEM educator (item #34 in the EUG survey questionnaire)
3) Evaluation increased my belief in the importance of my project (item #43 in the EUG survey questionnaire)

To assess the extent to which these deleted items differ from or conform to the retained items in the Evaluation Scale, Table 14 and Table 15 show the item distributions and descriptive statistics for the three removed items compared to the 20 retained items.

Table 14 illustrates the similarities between the item distributions for the three removed items and the 20 retained items, and Table 12 provides the same for the retained items. The response distributions for the removed items mirror the others in showing the largest proportion of responses consistently in the “some” category and generally follow a similar pattern across the board.

Table 14.

Response Distributions for Items Removed from Evaluation Use Scale

Survey Item #	Item	No	A Little	Some	Extensively	N
28	Increased my knowledge/under-standing of STEM education	84 27.5%	73 23.9%	100 32.7%	49 16.0%	306 100.0%
34	Improved my skills as STEM educator	103 33.6%	77 25.1%	93 30.3%	34 11.1%	307 100.0%
43	Increased my belief in the importance of my project	78 25.3%	47 15.3%	109 35.4%	74 24.0%	308 100.0%

Note: Response options were: 1 = No; 2 = Yes, a little; 3 = Yes, some; 4 = Yes, extensively.

An examination of the descriptive statistics (Table 15) for the three survey items not included in the Evaluation Use Scale reveals no strong divergence from the survey items that were retained in the scale (see Table 10). The range of means for the 20 items retained in the scale was 2.29 to 2.99. The means for the deleted items fell within or slightly below the range of means for the retained scale items. The standard deviations for the scale items ranged from 0.95 to 1.16; and the standard deviations for the removed items fell within that same range.

Table 15.

Descriptive Statistics for Items Removed from Evaluation Use Scale

Survey Item #	Evaluation Use Scale Item	Mean	Standard Deviation	Corrected Item-Total Correlation
28	Increased my knowledge/ understanding of STEM education	2.37	1.052	0.76
34	Improved my skills as a STEM educator	2.19	1.024	0.67
43	Increased my belief in the importance of my project	2.58	1.111	0.64

Note: Response options were: 1 = No; 2 = Yes, a little; 3 = Yes, some; 4 = Yes, extensively.

The items removed from the scale do not differ dramatically from the retained items in statistical terms. However, the removed items are substantively different from the retained items in one important respect: none of the removed items directly addresses the topic of evaluation nor contains the word "evaluation." The majority of the retained items (17 of 20) explicitly address evaluation knowledge, skills, beliefs, etc., whereas the removed items do not mention evaluation. Instead, the removed items touch on increased knowledge of STEM education, improved skills as a STEM educator, and increased belief in the importance of the project – but they do not mention evaluation.

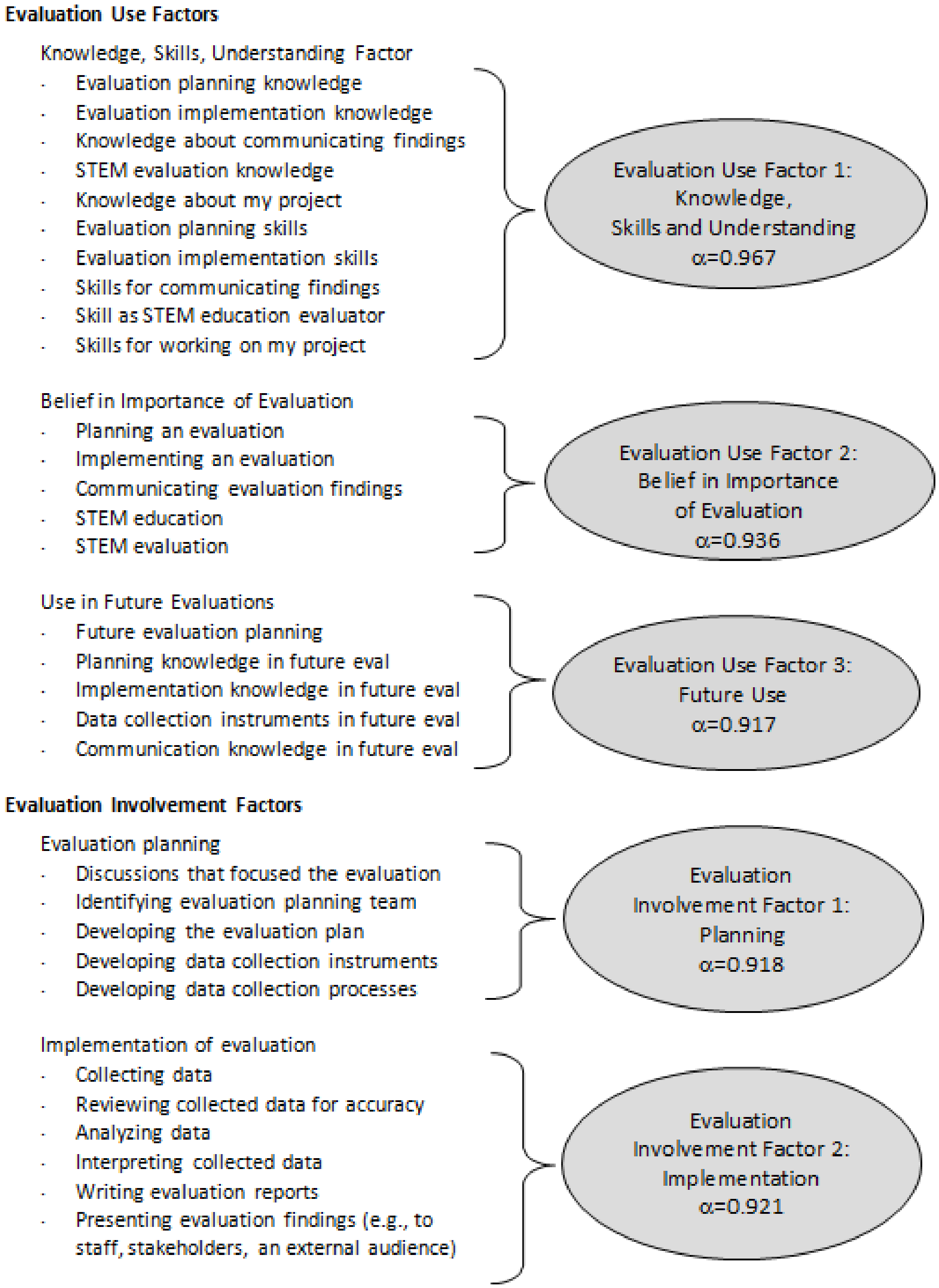
In considering how survey responses to these items may have varied such that the items did not load onto the three identified factors in the Evaluation use Scale. The removed items may have evoked different responses because they moved outside of the field of evaluation and asked respondents about topics where they may have felt more ownership or expertise. For example, these questions focused on the profession, the field, and the specific projects of the respondents.

As such, the respondents may have felt more ownership, expertise, and ability to speak to these issues. They may have concluded that their knowledge, skills, and beliefs in their specialty areas were a) not in need of improvement, so could not have been positively affected by the evaluation, or b) not likely to be helped by the evaluation asked about in the survey, if they were in need of improvement at all. These scenarios, plus others, could have resulted in the removed items measuring something other than evaluation use, and thus not loading onto the factors in the evaluation use scale.

Figure 5 illustrates the factor structure of the Evaluation Use Scale, as well as the factor structure of the Evaluation Involvement Scale (Toal, 2007). The figure displays the results of the factor analysis conducted as part of the Evaluation Use Scale validation study, along with the factor analysis of the Evaluation Involvement Scale conducted by Toal (2007).

Figure 5.

Factor Structure of Evaluation Use and Involvement Scales



Evaluation Use Factors

Knowledge, Skills, Understanding Factor

- Evaluation planning knowledge
- Evaluation implementation knowledge
- Knowledge about communicating findings
- STEM evaluation knowledge
- Knowledge about my project
- Evaluation planning skills
- Evaluation implementation skills
- Skills for communicating findings
- Skill as STEM education evaluator
- Skills for working on my project

Evaluation Use Factor 1: Knowledge, Skills and Understanding α=0.967

Belief in Importance of Evaluation

- Planning an evaluation
- Implementing an evaluation
- Communicating evaluation findings
- STEM education
- STEM evaluation

Evaluation Use Factor 2: Belief in Importance of Evaluation α=0.936

Use in Future Evaluations

- Future evaluation planning
- Planning knowledge in future eval
- Implementation knowledge in future eval
- Data collection instruments in future eval
- Communication knowledge in future eval

Evaluation Use Factor 3: Future Use α=0.917

Evaluation Involvement Factors

Evaluation planning

- Discussions that focused the evaluation
- Identifying evaluation planning team
- Developing the evaluation plan
- Developing data collection instruments
- Developing data collection processes

Evaluation Involvement Factor 1: Planning α=0.918

Implementation of evaluation

- Collecting data
- Reviewing collected data for accuracy
- Analyzing data
- Interpreting collected data
- Writing evaluation reports
- Presenting evaluation findings (e.g., to staff, stakeholders, an external audience)

Evaluation Involvement Factor 2: Implementation α=0.921

Source: Evaluation Involvement Factors from Toal, 2007

Generalizability Aspect of Construct Validity

Question 4: To what extent does the Evaluation Use Scale measure evaluation use in other multi-site settings?

The validity of a scale is increased by the extent to which it is reliable in different settings, often referred to as generalizability. The Evaluation Use Scale was administered to four, distinct programs with multi-site evaluations as part of the EUG study. This analysis compares the reliability of the Evaluation Use Scale in measuring evaluation use across four NSF-funded national programs: ATE, CETP, LSC, and MSP-RETA.

These four national programs shared several common characteristics, namely, a focus on STEM education evaluation, a program structure with multiple sites across the country, and a strong evaluation component. While sharing these high-level goals, the programs differed in terms of program objectives and operations. And, as described in the discussion of participants in Chapter 3, the MSP-R ETA program, in particular, diverged the most in terms of its structure and purpose.

Unlike ATE, CETP, and LSC programs, because of differences in program structure and goals, the MSP-RETA did not conduct a single, distinct program evaluation. Instead, the Utah State MSP-RETA project was a capacity building project designed to provide technical assistance and evaluation expertise to the MSP projects around the country. Based on these structural differences, we might expect to see weaker evidence of the consistency and reliability of the Evaluation Use Scale in measuring evaluation use in other multi-site settings for MSP-RETA respondents. But that was not the case.

Table 16 displays a comparison of the reliability coefficients for each of the three factors in the Evaluation Use Scale across the four programs involved in the study. The scale means are lower for the MSP-RETA, and this lower level of reported evaluation use among MSP-RETA respondents may be a function of differences in the structure of the program, i.e., the MSP-RETA program did not conduct a distinct program evaluation so there was less use to report. But the alpha levels are consistently high for all four programs across the three factors in the Evaluation Use Scale. These results suggest that the Evaluation Use Scale performed consistently and functioned similarly in measuring evaluation use in different settings.

Table 16.

Evaluation Use Scale Reliability Coefficients Across Different Settings

	Influence on Knowledge, Skills, Understanding			Influence on Belief in Importance of Evaluation			Use in Future Evaluation		
Project	Alpha	Scale Mean	N*	Alpha	Scale Mean	N*	Alpha	Scale Mean	N*
ATE	0.969	24.46	175	0.955	12.97	183	0.952	11.70	67
CETP	0.972	25.65	52	0.913	12.45	53	0.898	12.45	31
LSC	0.946	29.62	69	0.903	13.49	69	0.872	13.81	54
MSP	0.982	21.06	50	0.974	10.06	51	0.955	9.79	19

*Listwise deletion based on all variables in the procedure.

The EUG survey was also administered to another group of MSP project personnel in a subsequent data collection effort. A review of the summary data report from the subsequent fielding of the EUG Survey to another group of MSP projects shows that the means and standard deviations for the items in the Evaluation Use Scale fall in similar ranges across the two administrations of the survey (Johnson & Toal, 2009).

In the first fielding, the MSP-RETA respondents' (N=56) mean responses on the Evaluation Use Scale items ranged from 1.7 to 2.7, and the standard deviations ranged from 0.6 to 1.3. In the second fielding to a different group of MSP participants (N=41), the mean responses for items in the Evaluation Use Scale ranged from 1.8 to 2.8 and the standard deviations ranged from 0.9 to 1.2. These ranges are quite similar between the two administrations of the survey indicating roughly comparable performance by the Evaluation Use Scale items when administered in another setting.

Similarly, a different group of ATE projects received the EUG Survey at a later date. The mean responses in both administrations of the surveys covered roughly similar ranges -- 2.1 to 2.7 for the first group (N=175) and 2.0 to 2.9 for the second group of ATE projects (N=155); and the overall mean scores for the evaluation use items in the surveys were quite similar at 2.5 for the first and 2.6 for the second administration of the survey (Johnson & Toal, 2009).

These comparisons are rough estimates based on summary statistics presented in reports on the second fielding of the EUG Survey to different MSP and ATE projects. In both cases, the Evaluation Use Scale rendered comparable survey means and standard deviations as described above. While the score ranges appear to be somewhat consistent between the two administrations of the survey within the MSP and ATE programs, no statistical testing was conducted given that data access was limited to summary data tables only. More in-depth analysis, statistical comparisons, and reliability coefficient calculations would be necessary to improve the strength and quality of conclusions.

External Aspect of Construct Validity

Question 5a: To what extent does the Evaluation Use Scale correlate with expected activities (convergent validity)?

A scale has external validity to the extent that its relationship with other variables is rationally or empirically expected (Messick, 1995). It is important to consider both predictive and discriminant correlational patterns (Smith & McCarthy, 1995). A sound theoretical foundation helps identify where convergence and divergence are expected. A good theory articulates not only what the construct is, but what it is not (Clark &Watson, 1995). As a result, if unexpected correlations arise during the validation process, it is typically not a surprising new discovery; instead those relationships suggest that the measure is perhaps not capturing the construct under study.

The EUG focused on exploring the relationship between evaluation use and involvement, and a full exploration of this relationship is beyond the scope of this study. However, there was substantial theoretical and empirical support for a relationship between involvement and use, so it was also reasonable to expect the Evaluation Use and Involvement Scales to show a positive correlation between them.

Table 17 shows the results of a bi-variate correlation among the three factors of the Evaluation Use Scale with the Evaluation Involvement Scale, showing a significant, positive relationship.

Table 17.

Correlations between Evaluation Use and Involvement Scale Factors

Factors	Influence on Knowledge, Skills, Understanding	Influence on Beliefs about Evaluation	Use in Future Evaluation	Involvement Mean
Influence on Knowledge, Skills, Understanding	1.00	.75*	.73*	.55*
Influence in Beliefs about Evaluation	.75*	1.00	.55*	.41*
Use in Future Evaluation	.73*	.55*	1.00	.46*
Involvement Mean	.55*	.41*	.46*	1.00

* Correlation is significant at the .01 level (2-tailed).

Question 5b: To what extent does the Evaluation Use Scale differentiate between rationally or theoretically different groups (discriminant validity)?

An expectation of divergence between survey respondents who are evaluators versus those who are project leaders was supported by the knowledge that the duties and experience of these two groups were likely to be quite different. For example, project leaders often have a wide range of responsibilities outside of program evaluation (e.g., project management, budgeting, implementation, and reporting) whereas evaluators are primarily responsible for a more limited scope, namely the evaluation.

Consequently, it is reasonable to conclude that strong evidence in support of the discriminant aspect of construct validity would include statistically higher levels of evaluation use among evaluators than among project leaders or principal investigators.

Table 18.

Evaluators vs. Non-evaluators on Evaluation Use Scale Factors

	Influence on Knowledge, Skills, Understanding Factor		Influence on Beliefs about Evaluation Factor		Use in Future Evaluation Factor	
	Evaluator	Non-Evaluator	Evaluator	Non-Evaluator	Evaluator	Non-Evaluator
Mean	2.75	2.52	2.62	2.60	2.56	2.50
Std.Dev.	0.91	0.89	0.98	0.98	0.94	0.91
N	84	226	84	226	58	95

Note: Response options for scale items: 1 = No; 2 = Yes, a little; 3 = Yes, some; 4 = Yes, extensively.

However, Table 18 compares evaluators to non-evaluators (i.e., survey respondents who were PIs/project leaders compared to those who identified themselves as evaluators) and shows the close similarities in mean scores for both groups across all use factors. Results from an independent samples t-test suggested no significant differences between evaluators and project leaders in terms of the levels of use reported:

- Influence on knowledge, beliefs, and skills (F=2.07, p=.049)
- Influence on beliefs about the importance of evaluation (F=.086, p=.053)
- Use in future evaluations (F=.326, p=.70)

Consequential Aspect of Validity

Question 6: To what extent could possible interpretations or uses of the Evaluation Use Scale result in a negative impact?

Messick’s validation framework includes a comprehensive consideration of the possible consequences of the interpretation and use of a particular test; and a validation study should compile evidence of positive impacts, along with evidence that negative consequences are minimized. Validity is compromised when the use of a test creates bias, which can result from: (1) irrelevant content that interferes with the demonstration of competence, or (2) missing content which would have provided an opportunity to demonstrate competence (Messick, 1989a; 1998).

Messick (1998) also identified reliability, validity, and fairness as standards to be met in order to support a proposed test use, specifying that

> Fairness requires evidence that score meaning does not differ consequentially across individuals, groups, or settings… [And] because value implications as a basis for action are integral to score meaning, construct validation has to include an appraisal of these value implications in terms of the actual and potential consequences of test use (pp. 3-4).

Intended Consequences for Use of the Evaluation Use Scale

The process of gathering evidence of the consequential aspect of construct validity and considering the potential positive and negative consequences includes addressing the questions: What are the intended consequences for use of the Evaluation Use Scale?

What is the purpose of the Evaluation Use Scale? And, how are the scores to be used? And what does the testing practice claim to do? (Shepard, 1993, p. 429).

The Evaluation Use Scale was developed as part of an academic research study aimed at expanding the understanding of evaluation practices that promote evaluation use. To inform this study, researchers developed the Evaluation Use Scale as part of a survey to collect data from project leaders and evaluators in four, NSF-funded, multi-site evaluation programs.

The specific purpose of the Evaluation Use Scale was to collect data about how individuals and groups used evaluation processes, products, and results in multi-site evaluation settings. Within the context of the research study, the scores from the Evaluation Use Scale (along with the scores from the Evaluation Involvement Scale, another scale developed under this research project) were used to quantify the intensity of evaluation use in several areas, including: increased evaluation knowledge and skills, enhanced belief in the importance of evaluation, and respondents' used of what they learned in other evaluations.

The measures were then used to analyze how evaluation use is related to involvement in program evaluation, as well as to assess what evaluation practices are most directly related to enhancing evaluation use and influence. The measures informed the researchers' judgments and decisions about how and how much individuals and groups used the evaluation processes, products, and results.

Possible Adverse Consequences

The tasks and activities captured in the Evaluation Use Scale were typical of the majority culture, so it is possible that other ways of using evaluation findings were not included in the scale. This omission in the scale development phase could result in the scale underestimating the level of use for members of the non-majority culture.

Chapter Five weighs the relative strength of the evidence presented for each of the six aspects of validity, demonstrates the extent to which there is sufficient evidence to support the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site settings, and discusses the implications and limitations of this study.

CHAPTER FIVE
DISCUSSION AND IMPLICATIONS

Evaluation of Validity Evidence Framework

This chapter presents the validity evidence to address the primary question in this validation study: *Is there sufficient evidence to support the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site evaluations?*

The validity evidence is presented in the order of the research questions and organized using the framework of Messick's six aspects of construct validity. This chapter reports the assessment of the relative strength or weakness of the validity evidence in supporting the adequacy and appropriateness of inferences about the level of evaluation use in multi-site settings based on data collected using the Evaluation Use Scale. The chapter concludes with a discussion of implications, limitations, and suggestions for future research.

Content Aspect of Construct Validity

Q1a. To what extent is the construct of evaluation use theoretically sound?

The literature review, think-aloud interviews, and project interviews together form strong evidence that evaluation use is a sound theoretical construct and is widely acknowledged as such.

Literature Review: In both editions of the book, *Evaluation Roots* (2004, 2011), Alkin devotes an entire branch of his theory tree to evaluation use theory. He identifies nine evaluation theorists who have built the theoretical foundations of the concept of evaluation use. Notably, Alkin excludes several prominent evaluation experts and academicians known for writing about evaluation use, explaining that “all those who have written about evaluation cannot be considered to have developed a unique evaluation theory” (Alkin, 2004, p. 5).

In addition to this strong theory base for the construct of evaluation use provided by Alkin, evaluation use is also extensively studied and written about in the peer-reviewed literature, demonstrating the widespread knowledge about and recognition of evaluation use as a familiar concept in the field. A review of the evaluation literature revealed over 100 studies of evaluation use published since 1971 (Cousins & Leithwood, 1986; Johnson et al., 2009). And evaluation use was also among the top ten research topics published in the *American Journal of Evaluation* in the years 1998-2014 (Brandon, 2015).

Think-Alouds: The feedback from the think-alouds process provides evidence that the individual participants in the study both recognized and understood the concept of evaluation use and were able to discuss it knowledgeably. No participant in the think-aloud exercises conveyed confusion or questioned the meaning of evaluation use.

Project Interviews: Similarly, none of the survey respondents provided feedback indicating a lack of familiarity with the concept of evaluation use. And, finally, none of the survey respondents who consented to the follow-up interviews displayed any uncertainty about the topic of evaluation use.

Overall, the information from these three sources uniformly indicates that the evidence for the theoretical soundness of the construct of evaluation use is strong. The evidence collected as part of the think-aloud interviews adequately supports the contention that evaluation use is a sound and well-recognized construct. While no doubts were raised about the conclusion that evaluation use is a viable and theoretically supported construct, the limited number of think-aloud interviews conducted leaves room for what could have been a more robust investigation. Thus, the think-aloud evidence is rated as having provided evidence of only an adequate level of support and strength.

Q1b: To what extent is the Evaluation Use Scale internally consistent and reliable?

The descriptive item statistics provided largely positive indications about the strength of the Evaluation Use Scale. First, the item means ranged between 2.29 and 2.99. Mean scores close to the center of the range are desirable, indicating that they are detecting values across the range of possible responses, rather than "piling up" at one extreme or the other (DeVellis, 2003).

Second, the standard deviations for each item (ranging 0.95 – 1.16) indicated that items varied sufficiently, thereby supporting the contention that the items in the Evaluation Use Scale effectively discriminate between levels of use, and therefore, that the sample was diverse enough to capture the range of use levels.

Third, the corrected item-total correlations (ranging 0.53 – 0.87) indicated strong, positive correlations between each item and the whole set of items (with the item itself removed from the calculation of the whole scale) (Pett, Lackey, & Sullivan, 2003). In a reliable scale, the correlations between each item and the whole scale would be high, which they are in this case. The scale item *"used data collection instruments in* *another evaluation"* had the lowest corrected item-total coefficient, indicating that this item had the highest potential for exclusion had that been indicated by other study findings.

Finally, the internal reliability coefficient (often called Cronbach's alpha) for the Evaluation Use Scale was high (alpha = 0.963), meaning that the internal consistency of the scale is high, and indicating that a high proportion of the variance in the scale scores was attributable to true differences between respondents, rather than to error (DeVellis, 2003).

Each of these statistical analyses yielded strong measures of the reliability and internal consistency, supporting the assertion that the Evaluation Use Scale is likely to be consistent in reflecting the construct of evaluation use in its measurements. These results provide strong evidence in support of the assertion that the Evaluation Use Scale is internally consistent and reliable.

Substantive Aspect of Construct Validity

Q2a. To what extent does the Evaluation Use Scale include all relevant activities and processes of evaluation use?

Passages or statements by interviewees reflected the evaluation use domains revealing considerable alignment between the content of the Evaluation Use Scale and the evaluation use activities identified by interviewees. The interview respondents provided examples of evaluation use activities that linked to each item in the scale, thus providing strong evidence of the inclusiveness of the Evaluation Use Scale. This alignment is strong evidence in support of the *representativeness* of activities and processes engaged in by evaluators and project leaders.

Additionally, based on the comparison to items in the Evaluation Use Scale, the evidence suggests that all scale items matched up to ways that respondents used evaluations. However, the interviewees also provided examples of uses of evaluation that were not included in the Evaluation Use Scale, indicating that at least one process of evaluation use was not included in the scale. Therefore the evidence on the inclusiveness of the Evaluation Use Scale was marginal.

The scale did not, for example, contain an item related to use of the evaluation as motivation for better and more timely evaluation documentation, as one interviewee described:

> [The evaluation] helped to motivate us to assemble our earlier evaluation results and get them into a form where we could share them. Without them, we might have been slower in getting those things pulled together.

There were other scale items that might involve the activities mentioned by this interviewee. For instance, “improved my skills as a STEM evaluator” could include this motivation for better and earlier organizing. But no item in the Evaluation Use Scale specifically mentioned motivation. Interestingly, motivation was included in the Evaluation Involvement Scale, where respondents were asked about their motivation for being involved in the evaluation (whether they were asked, encouraged, or required to participate).

Q2b. To what extent does the Evaluation Use Scale measure actual evaluation use?

The item distributions displayed in Table 12 provided evidence for the substantive aspect of validity. Overall, the majority of people who completed the survey responded that each evaluation use activity in the Evaluation Use Scale had happened, at least “a little.” And for every survey item, at least 15% of respondents responded with “no,” stating that the particular evaluation use activity or process did not happen for them. The descriptive statistics in Table 10 show the distribution of survey responses relative to each item in the Evaluation Use Scale. From this display, it is clear that all tasks had a majority of respondents performing them at least a little bit, providing strong evidence that the items in the scale represent activities in which the respondents are actually engaged.

Quotations matched to scale items: The matching of interview quotations to scale items described under the consideration of the substantive aspect of construct validity in the previous chapter shows that many interviewee comments linked to items in the Evaluation Use Scale that each existing survey item was linked to an individual’s experience in some way. These results suggest that the scale captured a fair amount of actual use in the majority of projects. However, the comments of one interviewee point to a concern about the quality of the interview data. Specifically, an interviewee commented that his/her use of the evaluation was minimal. However, when asked to provide a summary rating on level of evaluation use, this interviewee reported “extensive” use of the evaluation. Based on these findings, the available evidence for the substantive aspect of construct validity provides only marginal support for the scale’s overall validity.

Structural Aspect of Construct Validity

Q3: To what extent do the Evaluation Use Scale’s statistical factors align with the underlying theoretical and rational structures?

Validity is related to what the scale measures, in this case, the latent construct of evaluation use. To ensure that the scale is adequately measuring the construct, researchers frequently conduct factor analysis. Thompson and Daniel (1996) wrote “factor analysis and construct validity have long been associated with each other” (p. 197), and Nunnally (1978) notes: “Factor analysis is intimately involved with questions of validity” (p. 112).

The EFA resulted in an underlying structure with three factors: (1) increased knowledge/skills/ understanding; (2) increased beliefs in the importance of evaluation, and (3) future use of evaluation process, products, or findings. All three of the factors had reliability coefficients between 0.90 and 0.97 across all programs.

The evidence supplied by this factor was high. Score validity for the Evaluation Use Scale, the evidence regarding the consistency between the theoretical and statistical structures was strong.

Generalizability Aspect of Construct Validity

Q4: To what extent does the Evaluation Use Scale measure evaluation use consistently in other multi-site settings?

The evidence supporting the Evaluation Use Scale as a consistent measure of evaluation use in other settings (generalizability) is strong. The EUG survey was administered to project leaders and evaluators in four large, multi-site NSF programs.

Coefficient alpha levels were consistent – and consistently high – within factors across programs, ranging from 0.95 to 0.98 for the knowledge/skills/use factor; from 0.90 to 0.96 for the beliefs in the importance of evaluation factor; and from 0.87 to 0.98 for the use in future evaluations factor, as shown in Table 16. These high alphas indicate that the Evaluation Use Scale functions reliably and consistently in measuring evaluation use across different settings. All three factors of the Evaluation Use Scale had alpha levels between 0.87 - 0.98 across all programs. The consistently high alphas serve as strong evidence supporting the suggestion that the scales are equally reliable in a variety of multi-site program evaluation settings.

The EUG survey was also administered to another group of MSP project personnel in a subsequent data collection effort. A review of the summary data report from the subsequent fielding of the EUG Survey to another group of MSP projects shows that the means and standard deviations for the items in the Evaluation Use Scale fall in similar ranges across the two administrations of the survey.

In the first fielding, the mean responses from the MSP program respondents (N=56) for the Evaluation Use Scale items ranged from 1.65 to 2.67, and the standard deviations ranged from 0.58 to 1.30. In the second fielding to a different group of MSP participants (N=41), the mean responses for items in the Evaluation Use Scale 1.78 to 2.80 and the standard deviations ranged from 0.89 to 1.18. These ranges are quite similar, indicating comparable performance by the Evaluation Use Scale items when administered in another setting. This evidence provides only marginal additional support for the generalizability aspect of construct validity because stronger support would require access to the dataset to support calculations of coefficient alpha and other analyses.

External Aspect of Construct Validity

Q5a: To what extent does the Evaluation Use Scale correlate with expected activities?

The EUG study tested the relationship between evaluation involvement and use and expected a positive correlation between evaluation involvement and use in multi-site settings. Table 17 displays the significant, positive correlations among the three factors of the Evaluation Use Scale and the two factors of the Evaluation Involvement Scale provide strong evidence of the convergent validity of the scale.

Q5b: To what extent does the Evaluation Use Scale differentiate between rationally or theoretically different groups?

The survey includes responses from two rationally different groups of individuals involved in each of the program evaluations, namely evaluators and project leaders (who were not evaluators). Because of the different job responsibilities and distinct points of view, it is reasonable to conclude that the two groups would have distinguishable mean scores on the factors in the Evaluation Use Scale. Further, one would expect that analysis would show statistically higher levels of evaluation use among evaluators than among other project leaders.

However, no statistical differences between these two rationally different groups emerged, and evaluators and project leaders alike shared similar mean scores across all three factors in the Evaluation Use Scale, as shown in Table 18. Results from an independent samples t-test suggested no significant differences between evaluators and project leaders in terms of the levels of use reported. This lack of differentiation between groups that are logically and theoretically expected to be different does not support the strength of the external, discriminant validity aspect of the construct validity of the Evaluation Use Scale.

This analysis revealed weak evidence that the Evaluation Use Scale can differentiate levels of evaluation use between groups where differences are expected. Mean score comparisons illustrated that the Evaluation Use Scale was unable to differentiate evaluators from project leaders based on a comparison of mean scores on the three factors included in the Evaluation Use Scale.. Despite the logical analysis supporting the conclusion that project leaders differ from evaluators in terms of evaluation use, a comparison of mean use scores between evaluators and project leaders did not support that finding.

An alternative interpretation of the finding that the Evaluation Use Scale did not distinguish between evaluators and non-evaluators is that the assumption that evaluators and non-evaluators are logically and theoretically different might not hold true in the case of the programs included in this study. This is discussed further in the limitations section at the end of this chapter.

Consequential Aspect of Construct Validity

Q6: To what extent could possible interpretations or uses of the Evaluation Use Scale result in a negative impact?

There was no evidence that negative consequences were likely to arise from the use of the Evaluation Use Scale. There could potentially be bias in that the language and examples are those typical of majority Americans. However, there is little chance that this bias negatively affects groups from other cultures. The Evaluation Use Scale is a research instrument; it is not a high-stakes educational examination or a test by which to judge stakeholders or evaluators.

Summary of the Evidence

This section summarizes the evidence in each aspect of the unitary validation framework which collectively addresses the primary research question:

> *Is there sufficient evidence to support the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site evaluations?*

Table 19 presents the summary assessment of the relative strength of the available evidence for each aspect of validity, using a rating scale of strong, adequate, marginal, or weak. A rating of “Strong” indicates that there is solid evidence in support of the effectiveness and comprehensiveness of the Evaluation Use Scale in addressing the specified aspect of validity; an “Adequate” rating indicates that the available evidence was substantial, but not uniformly strong or comprehensive; a “Marginal” option signifies that there is limited or mixed information available and the available evidence is, on balance, weak; and, finally, a “Weak” rating indicates that the Evaluation Use Scale did not satisfy the validity evidence criteria.

Overall, the evidence in support of the adequacy and appropriateness of interpretations and actions based on the Evaluation Use Scale in multi-site settings is mixed. Of the six aspects of construct validity in Messick’s unitary validity framework, four aspects are “Strong” or “Adequate,” and two aspects are “Mixed” (Strong & Weak, Strong & Marginal). An argument can be made that there is sufficient evidence to support the validity of the Evaluation Use Scale as a measure of evaluation use in multi-site evaluation settings. In addition, the areas where the evidence was weakest may be partially attributable to the limitations of this research and study design, and not reflective of particular weaknesses in the scale itself.

Table 19.
Relative Strength of Evaluation Use Scale Validity Evidence

Research Questions	Data Sources	Quality	Rationale
Content Aspect			
Q1a. To what extent is the construct of evaluation theoretically sound?	Published research	Strong	Sound literature
	Think-alouds	Adequate	Low # of interviews
	Interviews	Strong	Strong interview data
Q1b. To what extent is the scale internally consistent and reliable?	Item variance	Strong	Good statistical results
	Scale alpha = 0.986	Strong	Good statistical results
Substantive Aspect			
Q2a. To what extent does the scale include all relevant activities and processes?	Expert judgment	Strong	Well-regarded experts
	Interviewee mentions of use	Marginal	Too few instances of on-point examples
Q2b. To what extent does the scale measure actual use?	Response distributions	Strong	Good statistical results
	Project interviews	Marginal	Few on-point examples
Structural Aspect			
Q3. To what extent do the statistical factors match underlying structures?	Exploratory Factor Analysis results	Adequate	EFA strong; no CFA due to insufficient data
Generalizability Aspect			
Q4. To what extent does the scale measure consistently in other multi-site settings?	Coefficient alpha for Scale and each factor	Strong	High coefficient alpha for scale & factors
	Summary of 2nd fielding of survey	Adequate	Comparable means; only summary data
External Aspect			
Q5a. To what extent does the scale correlate with expected activities?	Convergent validity: Significant, positive	Strong	Strong Correlation
Q5b. To what extent does the scale differentiate between different groups?	Evaluators compared to non-evaluators (project leaders)	Weak	Weak Statistical Results
Consequential Aspect			
Q6. How might uses or interpretations of the scale result in a negative impact?	Discussion of the possible biases	Marginal	Possible biases related to multicultural validity

Limitations

Potential study limitations include a number of issues including data collection methods, memory/recall strain, project vs. program-level confusion, and others. One of the limitations of the study relates to self-reported data collected via a self-administered, web-based survey. Self-reported data can be subject to social desirability influence, where individuals respond to survey items to characterize their behavior or participation in a more socially acceptable or attractive way. In this survey, it is conceivable that some respondents may have wanted to burnish the portrayal of their projects' performance given that the study was led by a well-known NSF evaluator and was funded by the NSF.

Memory and accurate recall on the part of respondents may have been difficult due to the amount of time elapsed between respondents' involvement in their programs and the time they completed the survey. The program descriptions in the appendix include information on dates and timing, demonstrating that this time gap could be quite extensive in some cases, making it more challenging to remember and report on levels of past evaluation use.

The evident confusion among a small number of respondents as to whether the survey was addressing program-level (national) evaluation activities or project-level (local project site-specific) activities is another possible limitation. Notwithstanding the EUG team's efforts to communicate and reinforce that program-level evaluation was the target for the research, some survey respondents

In addition, the pool of responses was not large enough to support confirmatory factor analysis (CFA) which requires a sufficient number of responses per item on the scale to support the analysis. In this case, the number of responses was inadequate to meet the minimum required number of responses per scale item to support CFA.

Finally, the analysis conducted to assess the ability of the Evaluation Use Scale to discriminate between logically or rationally different groups compared the use factor means for PIs/project leaders vs. evaluators, based on the notion that these groups represent two distinct types of program participants with different roles in the program, as well as different education, skills, and expertise. The analysis showed that the scale was not able to effectively discriminate between these two groups. A potential limitation of the study is the possibility that the differences between PIs and evaluators are not uniformly clear and distinct, and that the education, experience, and characteristics of the evaluators and PIs/project leaders might make these groups more alike than different in the case of these four NSF programs.

Discussion and Implications

Although developed for a specific academic research study, the Evaluation Use Scale could be helpful in other academic studies involving evaluation use in multi-site settings. In addition, outside of academic research, one can imagine many other possibilities for the applied use of the Evaluation Use Scale. For example, in other contexts, researchers, policy makers, government grant makers, philanthropic funders, program personnel, and many others could use the Evaluation Use Scale to quantify the levels and intensity of the use of program evaluation process, findings, and products in other multi-site evaluation settings.

In these settings, the intended application of the evaluation use measures could be to inform decisions about program improvement, continuation, or funding levels. Measuring the level and intensity of evaluation use could also be part of ongoing program administration and monitoring. For example, in service of the management adage “what gets measured gets done” a policy dedicated to making better use of its multi-site evaluation activities could potentially use the Evaluation Use Scale to track evaluation use levels over time.

Measuring evaluation use in multi-site programs – whether funded by the government, private enterprise or the public sector – could help answer accountability questions, including: Did we learn anything from our evaluation? Did anything change as a result of evaluating our program? Are we any better off for having evaluated ourselves? Was it worth the time, energy, and money?

Making use of a tool to quantify the amount and intensity of evaluation use in multi-site settings could also spur efforts to increase levels of evaluation use, to improve the use of evaluations, and to improve the quality and usability of the evaluations themselves. Evaluation use has been and remains a frequently studied and written about topic in the field. More recent publications notwithstanding, Carol Weiss’ 1998 chapter “Improving the use of evaluations: Whose job is it anyway?” offers a concise and still timely short set of recommendations for improving evaluation use. She identifies three basic strategies: (1) Produce high quality evaluations (which can be difficult, she acknowledges, due to political and financial pressures); (2) Write short, clear reports that are widely disseminated, in addition to “executive summaries that can be absorbed while the reader stands on one foot” (p. 267); and (3) Leverage data from other studies when feasible by conducting meta-analysis to synthesize results from evaluations of similar programs (Weiss, 1998).

In terms of future research, the Evaluation Use Scale could make an important contribution in terms of measuring evaluation in multi-site settings. In turn, the knowledge gained could help inform the policies that guide or regulate large multi-site evaluations in terms of participation and use.

Fielding a survey with the Evaluation Use Scale in other multi-site settings – both similar to and different from the four sites in this study – would inform a more in-depth understanding of the extent of the external aspect of construct validity of the Evaluation Use Scale and its capacity to differentiate between theoretically or logically different groups. This information would contribute to strengthening the validity argument.

REFERENCES

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

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: APA.

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

Alkin, M. C. (Ed.) (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage Publications.

Alkin, M. C. (2005). Utilization of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 435-440). Thousand Oaks, CA: Sage Publications.

Alkin, M. C. (Ed.) (2011). *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.). Los Angeles, CA: Sage Publications.

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

Alkin, M. C., & Taut, S. M. (2003). Unbundling evaluation use. *Studies in Educational Evaluation, 29*(1), 1-12. doi:10.1016/S0191-491X(03)90001-0

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). *The concept of validity. Psychological Review, 111* (4), 1061-1071. doi:10.1037/0033-295X.111.4.1061

Borsboom, D., Angelique, Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 135-170). Charlotte, NC: Information Age Publishing.

Brandon, P. (2015). Research on evaluation. *New Directions for Evaluation*, *18,* 3-6. doi:10.1002/ev.20152

Brualdi, A. (1999). *Traditional and modern concepts of validity.* (ERIC Clearinghouse on Assessment and Evaluation, ED435714).

Burke, B. (1998). Evaluating for a change: Reflections on participatory methodology. *New Directions for Evaluation, 80*, 43-56. doi:10.1002/ev.1116

Burry, J., Alkin, M. C., & Ruskus, J. (1985). Organizing evaluations for use as a management tool. *Studies in Educational Evaluation, 11,* 131-157. doi:10.1016/0191-491X(85)90020-3

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions. *American Journal of Evaluation, 28*(1), 8-25. doi:10.1177/1098214006298065

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7(*3), 309-313. doi:10.1037/1040-3590.7.3.309

Cousins, J. B., & Earl, L. M (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis, 14*(4), 397-418. doi:10.3102/01623737014004397

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331-364. doi: 10.3102/00346543056003331

Cousins, J. B., & Simon, M. (1996). The nature and impact of policy-induced partnerships between research and practice communities. *Educational Evaluation and Policy Analysis, 18,* 199-218. doi:10.3102/01623737018003199

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions for Evaluation, 80,* 5-24. doi:10.1002/ev.1114

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302. doi:10.1037/h0040957

Daillak, R. H. (1982). What is evaluation utilization? *Studies in Educational Evaluation, 8,* 157-162. doi:10.1016/0191-491X(82)90007-4

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks: Sage Publications.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons, Inc.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*, 1006-1012. doi:10.1111/j.1365-2929.2004.01932.x

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

Fowler, F. J. (2009). *Survey Research Methods* (4th ed.) Thousand Oaks, CA: Sage Publications.

Goodwin, L. C., & Leech, N. L. (2003). The meaning of validity in the new *Standards for Educational and Psychological Testing*: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36,* 181-191.

Greene, J. G. (1988). Stakeholder participation and utilization in program evaluation. *Evaluation Review, 12* (2), 91-116. doi:10.1177/0193841X8801200201

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*, 293–314. doi: 10.1177/109821400302400302

Herbert, J. L. (2014). Researching evaluation influence: A review of the literature. *Evaluation Review, 38*(5), 388-419. doi: 10.1177/0193841X14547230

House, E. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage Publications.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103,* 219-230. doi:10.1007/s11205-011-9843-4

Johnson, G., & Toal, S. (2009). *NSF second follow-up survey detailed report* (Report No. 9). Retrieved from Beyond Evaluation Use Project Website: http://www.cehd.umn.edu/EdPsych/BEU/documents/Second%20Survey%20Detailed%20Report_Rpt9.pdf

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377-410. doi:10.1177/1098214009341660

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112* (3), 527-535. doi:10.1037/0033-2909.112.3.527

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2*(3), 135-170. doi:10.1207/s15366359mea0203_1

King, J. A. (1988). Research on evaluation use and its implications for evaluation research and practice. *Studies in Educational Evaluation, 14*(3), 285-299.

King, J. A. (2006, November). *Persistent challenges in studying evaluation use and influence.* Presentation at the Annual Meeting of the American Evaluation Association, Portland, OR. Retrieved from: http://www.cehd.umn.edu/EdPsych/BEU/

King, J. A., Lawrenz, F., Toal, S., Greenseid, L., Ooms, A., & Johnson, K. (2006, April). *Patterns of involvement, use, and influence in multi-site evaluations: Evidence from four National Science Foundation programs*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. Retrieved from: http://www.cehd.umn.edu/EdPsych/BEU/

King, J. A., & Pechman, E. M. (1984). Pinning a wave to the shore: Conceptualizing school evaluation use. *Educational Evaluation and Policy Analysis*, *6*(3), 241-251. Retrieved from: http://www.jstor.org.ezp3.lib.umn.edu/stable/1163870

Kirkhart, K. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation*, *94*, 5-23. doi:10.1002/ev.1188

Lawrenz, F., & Huffman, D. (2003). How can multi-site evaluations be participatory? *American Journal of Evaluation*, *24*(4), 471-482. doi:10.1177/109821400302400404

Lawrenz, F., & King, J. A. (2009). *Beyond evaluation use cross-case report.* (Report No. 8). Retrieved from Beyond Evaluation Use Project Website: http://www.cehd.umn.edu/EdPsych/BEU/documents/CrossCase_Rpt8.pdf

Leff, H. S., & Mulkern, V. (2002). Lessons learned about science and participation from multisite evaluations. *New Directions for Evaluation, 94*, 89-100. doi: 0.1002/ev.53

Leviton, L. C. (2003). Evaluation use: Advances, challenges, and applications. *American Journal of Evaluation, 24*(4), 525-535. doi:10.1177/109821400302400410

Leviton, L. C., & Hughes, E. F. X. (1981). Research on the utilization of evaluations: A review and synthesis. *Evaluation Review*, 5, 525-548. doi:10.1177/0193841X8100500405

Lissitz, R.W, & Samuelsen, K. (2007). Dialogue on validity: A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448. doi:10.3102/0013189X07311286

Loevinger, J. (1957). Objective tests as instruments of psychological theory. (Monograph Supplement 9). *Psychological Reports, 3*, 635-694.

Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation, 10*(1), 35-57. doi:10.1177/1356389004042326

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35* (11), 1012-1027.

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*, (3rd ed. pp. 13-103). New York: Macmillan Press.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8. doi:10.1111/j.1745-3992.1995.tb00881.x

Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50* (9), 741-749. doi:10.1037/0003-066X.50.9.741

Messick, S. (1996a). *Standards-based score interpretation: Establishing valid grounds for valid inferences.* Proceedings of the joint conference on Standard Setting for Large Scale Assessments. Washington, DC: Government Printing Office.

Messick, S. (1996b). Validity of performance assessment. In Philips, G. (Ed.). *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Educational Statistics.

Messick, S. (1998). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment* (Report No: RR-98-48). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R.L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83-108). Charlotte, NC: Information Age Publishing.

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

Nunneley, R. D., King, J. A., Johnson, K., & Pejsa, L. (2015). The value of clear thinking about evaluation theory: The example of use and influence. In C. Christie & A. Vo, (Eds,) *Evaluation use and decision making in society: A tribute to Marvin C. Alkin,* pp. 53-71. Charlotte, NC: Information Age Publishing.

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 public policy making.* Lexington, MA: Heath, 141-164.

Patton, M. Q. (1997). *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003.) *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications.

Preskill, H., & Caracelli, V. (1997). Current and developing conception of use: Evaluation use TIG survey results. *Evaluation Practice, 18*(3), 209-226.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach.* (7th ed.). Thousand Oaks, CA: Sage Publications.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage Publications.

Shaukat, R. (2010) Moving beyond evaluation utilization theory to embrace a comprehensive theory of influence. *The International Journal of Interdisciplinary Social Sciences, 5*(2), 507-517.

Shaw, J., & Campbell, R. (2014). The 'process' of process use: Methods for longitudinal assessment in a multisite evaluation. *American Journal of Evaluation, 35*(2), 250-260. doi:10.1177/1098214013506780

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405-450. doi:10.3102/0091732X019001405

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *American Journal of Evaluation, 18*(1), 195-208. doi:10.1177/109821409701800121

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481. doi: 10.3102/0013189X07311609

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp.19-38). Charlotte, NC: Information Age Publishing.

Sireci, S., & Sukin, T. (2013). Test validity. In K. F. Geising (Ed.), *APA Handbook of testing and assessment in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 61-84). Washington, DC: American Psychological Association.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7*, 300-308. doi:10.1037/14047-004

Straw, R. B., & Herrell, J. M. (2002). A framework for understanding and improving multisite evaluations. *New Directions for Evaluation, 94*, 5-16. doi:10.1002/ev.47

Stufflebeam, D. L. (1968, January). *Evaluation as enlightenment for decision-making.* Address delivered at the Working Conference on Assessment Theory sponsored by the Commission on Assessment of Educational Outcomes Association for Supervision and Curriculum Development, Sarasota, FL. Accessed from http://eric.ed.gov/?id=ED048333.

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

Toal, S.A. (2007). The development and validation of an evaluation involvement scale for use in multi-site evaluation (Unpublished doctoral dissertation). University of Minnesota, Twin Cities.

Toal, S. A. (2009). The validation of the evaluation involvement scale for use in multisite settings. *American Journal of Evaluation, 30* (3), 349-362. doi:10.1177/1098214009337031

Thompson B., & Daniel, L. G. (1996) Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines, *Educational and Psychological Measurement, 56*(2) 197-208.

Vallin, L. M., Philippoff, J., Pierce, S., & Brandon, P. R. (2015). Research on evaluation articles published in the *American Journal of Evaluation*, 1998-2014. *New Directions for Evaluation, 148,* 7-15. dio:10.1002/ev.20153

Weiss, C. H. (1979). The many meanings of research utilization. *Public Administration Quarterly, 39*(5), 426-431

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Science Communication, 1*(3), 381-404.

Weiss, C. H., & Bucuvalas, M. J. (1980). Truth tests and utility tests: Decision-makers' frames of reference for social science research. *American Sociological Review, 45*(2), 302-313.

Weiss, C. H. (1998). Improving the use of evaluations: Whose job is it anyway? In A. J. Reynolds (Ed.), *Advances in Educational Productivity. Vol. 7* (pp. 263-276). Greenwich, CT: JAI Press.

Weiss, C. H., Murphy-Graham, E., & Birkeland, S. (2005). An alternate route to policy influence: How evaluations affect D.A.R.E. *American Journal of Evaluation, 26*(1), 12-30. doi:10.1177/1098214004273337

Yarbrough, D. B., Shula, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage Publications.

APPENDIX

NSF PROGRAM OVERVIEWS AND PROGRAM EVALUATION DESCRIPTIONS

The following pages describe the four National Science Foundation programs that formed the basis of the Beyond Evaluation Use Research project at the University of Minnesota. The program summaries are from the Beyond Evaluation Use Grant website. Retrieved from http://www.cehd.umn.edu/EdPsych/BEU/documents.html.

Advanced Technological Education (ATE)
Program Overview

The National Science Foundation funded the Advanced Technological Education (ATE) program in 1993 to increase the number and quality of technicians in the US by providing curriculum development, professional development, career pathways, and applied research on technological education**.** The program focused on collaborations primarily between community colleges and businesses focused on activities, such as material development, program improvement, and professional development. The ATE program evaluation collected data from ATE projects with a required annual survey along with site visits at 13 selected sites.

Period: 1993-present

Purpose: (1) Build capacity to provide advance tech education; (2) Increase number of skilled technicians.

Objectives: Projects focused on improving technical education materials, enhancing technical instruction, and providing professional development to faculty and teachers. Grantees established partnerships with high schools, two- and four-year colleges, businesses, government agencies, and professional societies.

Primary Target: Students & teachers in 2-year colleges and collaborating institutions.

Sites Funded: Between 1994 and 2005, ATE funded 674 projects and centers, making grants to 345 unique institutions of which more than 200 were two-year colleges.

NSF Program Evaluation Grant Funding: $1.3 million (1999-2002) and $1.8 million (2003-2005)

NSF Goals for ATE Program Evaluation: Provide evidence of quality of ATE program; Inform program improvement

Program Evaluation Questions:
To what degree is the program achieving its goals?
Is it making an impact and reaching the individuals and groups intended?
How effective is it when it reaches its constituents?
Are there ways to significantly improve the program?

Evaluation Personnel/Expertise:
NSF 3-person management team
Experienced evaluation team
Evaluation Advisory Committee

Collaboratives for Excellence in Teacher Preparation (CETP)
Program Overview

The CETP program responded to the national need to produce and retain well-qualified math and science teachers by improving the preparation of science, technology, engineering, and mathematics (STEM) teachers for grades pre-K -12. The improvement involved collaboration among several groups: college science & math departments, as well as community colleges and K-12 schools. Program evaluators sought input from sites in all phases of development: instrument development, data collection, and communication of results.

Period: 1993-2004

Purpose: To improve teacher preparation by enhancing science, mathematics, and educational methods courses at the college level

Objectives: Increase collaboration within and between K-12 and higher education institutions to foster the development of well-trained teachers competent in science and math.

Primary Target: Prospective pre-K–12 teachers

Sites Funded: From 1993 – 2000, NSF funded 19 CETPs, each involving 3-15 higher education institutions (including colleges, universities, and community colleges) and several school districts.

NSF Program Evaluation Grant Funding: $999,000 (1999-2004)

NSF Goals for CETP Program Evaluation: To learn to what extent the CETPs succeeded in achieving significant and systemic improvement in the science, technology, engineering, and mathematics (STEM) preparation of prospective pre-Kindergarten through grade 12 (preK-12) teachers

Evaluation Question: To what extent did the CETP program impact the collaboration and focus of university faculty on instructional issues? To what extent did the CETP program impact the instructional techniques used by university faculty? Did K-12 teachers who participated in CETP projects view their preparation programs differently from teachers who participated in other preparation programs? Were the instructional practices exhibited by K-12 teachers who participated in CETP projects different from the instructional practices exhibited by teachers who participated in other preparation programs?

Evaluation Activities: Convened meetings with CETP project personnel; Developed data collection instruments (surveys, classroom observation protocols.); Technical assistance to local CETPs for data collection and analysis; Standardized instruments were developed, but sites were free to use their own evaluation instruments or could add items to the standard instrument

Local Systemic Change through Teacher Enhancement (LSC)
Program Overview

The National Science Foundation (NSF) launched the Local Systemic Change through Teacher Enhancement initiative (LSC) in 1995 to improve K-12 science, mathematics, and technology instruction through teacher professional development. LSC intended to shift the professional development focus from individual teachers to district- or system-wide populations of teachers, based on the premise that system-wide initiatives yielded better results. The LSC was a $250 million program that funded 88 projects, reaching 4,000 schools across 31 states and 476 districts in the period 1995-2005. Some LSC projects took place in single school district and others covered multiple districts (up to 20 in some cases).

Period: 1995 – 2005 (Final year of new projects were funded was 2002.)

Purpose: Improve instruction in science, mathematics and technology through teacher professional development within entire school districts.

Objectives: Target all teachers in a jurisdiction for professional development, requiring a minimum of 130 hours of professional development; and prepare teachers to implement district-designated instructional materials.

Primary Target: K-12 teachers of science and mathematics; focus on entire school systems or districts, not on individual teachers.

Sites Funded: 88 projects current and completed projects across 31 states, and involving 70,000 teachers, 4,000 schools, 467 school districts.

NSF Program Evaluation Grant Funding: $6.25 million

NSF Goals for LSC Core Evaluation: Provide aggregated information across diverse projects to inform broader conclusions about design, quality, and impact of LSC projects**.** Assess individual projects to provide for mid-course adjustments

NSF Contracted with HRI to: Develop a data collection framework; Provide technical assistance in implementing evaluation activities; Prepare cross-site analyses of evaluation results

Local Project Evaluator Activities: Professional Development Session Observations (2,40 completed) ; Classroom Observations of mathematics and science lessons (1,620) ; Teacher Questionnaires (75,000) ; Principal Questionnaires (17,380) ; Teacher Interviews (1,782)

Mathematics and Science Partnerships Research Evaluation Technical Assistance (MSP-RETA) Program Overview

The Mathematics and Science Partnerships (MSP) program is the most recent program in the study, designed to improve student understanding through the development of partnerships with various institutions such as museums, colleges, and school districts. The NSF funded several sites specifically to provide research, technical, and evaluation assistance to the other MSP's; these sites were deemed MSP-RETAs. Although not a typical program evaluation like the other three programs, the Utah State RETA was charged with developing and assisting evaluation at the program level. To that end, the Utah State MSP-RETA provided evaluation assistance to other MSP sites who request it. The Utah State MSP-RETA provides assistance in the form of local, project-level assistance by external evaluators, as well as workshops and conferences on evaluation tailored to the MSP's. This evaluation involvement is totally participatory and in total control of the projects.

Period: 2002-present

Purpose: Research and development effort that supported innovative partnerships to improve K-12 math and science instruction, enhance student achievement, and reduce the achievement gap in math and science performance among diverse student populations.

Objectives: Supported the development of partnerships to integrate the work of higher education and K-12 STEM faculty through challenging coursework; Sought to increase the quantity, quality, and diversity of math and science teachers; and established results-oriented projects that implement evidence-based educational practices.

Primary Target: Teams of institutions: higher education, K-12, and other partners.

Sites Funded: In 2006, MSP was bringing together 150 institutions of higher education, 550 school districts, and 3,300 schools, along with corporate partners

NSF Program Evaluation Grant Funding: $1.5 million initially; additional funds of $300,000 received during year three of efforts

NSF Goals for MSP-RETA Program Evaluation: Work with the MSP projects to identify evaluation needs and develop program evaluation models based on those needs; Offered technical assistance and expert consultation via a network of evaluation consultants

Project Activities: Evaluation Capacity-building seminars, conferences, materials; provided individualized technical assistance; Developed and disseminated the Design-Implementation-Outcomes (DIO) Cycle framework

SURVEY INVITATION AND REMINDER EMAILS

Dear XXXX,

The National Science Foundation (NSF) wants to improve the use of program evaluation. As a recent participant in a [PROGRAM NAME], you are uniquely qualified to provide essential and firsthand input about your experiences with the [PROGRAM NAME] program evaluation. The survey, accessible through the link listed below, asks for your candid perceptions of your experience, which will help us better understand the evaluation process and its outcomes. This information will, in turn, enable NSF to enhance its evaluation activities. The web-based survey takes around 10 minutes to complete.

No names will be used in this study, and only aggregated results will be reported. At the end of the survey, you will be asked for permission to contact you if you are selected for a follow-up interview.

To show our appreciation for your time and thoughts, the first 50 respondents will receive a $10 amazon.com gift certificate. In addition, everyone who responds to the survey will be entered in a drawing for one of several web cameras. A webcam will add a visual dimension to your online chatting, videoconferencing, and e-mailing, allowing you to enjoy videoconferencing with friends and family all over the world with high-quality audio and video.

To participate, click on the survey link below, and you will be prompted for your pre-assigned account name and password (provided at the end of this email).

Please feel free to contact us at the email addresses listed below if you have questions or concerns about this survey. Or, if would like to talk to someone else, you may contact the University of Minnesota's Research Subjects' Advocate Line at (612) 625-1650.

We recognize the many demands made on your time, and thank you in advance for your willingness to help. Your input is critical, and we appreciate your involvement.

Sincerely,

Frances Lawrenz, Professor
Educational Psychology—University of Minnesota
lawrenz@umn.edu

Jean A. King, Professor
Educational Policy & Administration—University of Minnesota kingx004@umn.edu

Link: http://www2.education.umn.edu/EdPsy/NSF/Survey/Default.asp
Your account name and password:
This study is sponsored by NSF number REC0438545

REMINDER EMAIL

Dear XXX,

You have recently received an e-mailed invitation to complete a survey funded by the National Science Foundation (NSF). In an effort to improve the use of program evaluation, we are surveying [PROGRAM NAME] grant recipients because we feel you are uniquely qualified to provide essential and firsthand input about your experiences with the [PROGRAM NAME] program evaluation. The linked survey asks for your candid perceptions of your experience, which will help us better understand the evaluation process and its outcomes. This information will, in turn, enable NSF to enhance its evaluation activities. The web-based survey takes around 10 minutes to complete.

No names will be used in this study, and only aggregated results will be reported. At the end of the survey, you will be asked for permission to contact you if you are selected for a follow-up interview.

To show our appreciation, everyone who responds to the survey will be entered in a drawing for one of several web cameras. A webcam will add a visual dimension to your online chatting, videoconferencing, and e-mailing, allowing you to enjoy videoconferencing with friends and family all over the world with high-quality audio and video.

To participate, click on the survey link below, and you will be prompted for your pre-assigned account name and password (provided at the end of this email). Please feel free to contact us at the email addresses listed below if you have questions or concerns about this survey. Or, if would like to talk to someone else, you may contact the University of Minnesota's Research Subjects' Advocate Line at (612) 625-1650.

We recognize the many demands made on your time, and thank you in advance for your willingness to help. Your input is critical, and we appreciate your involvement.

Sincerely,

Frances Lawrenz, Professor
Educational Psychology—University of Minnesota
lawrenz@umn.edu

Jean A. King, Professor
Educational Policy & Administration—University of Minnesota kingx004@umn.edu

Link: http://www2.education.umn.edu/EdPsy/NSF/Survey/Default.asp
Your account name:
Your password:

This study is sponsored by NSF number REC0438545

CONSENT FORM - Evaluation Use Study

You are invited to be in a research study of the use and influence of program evaluation. You were selected as a possible participant because you have participated in the evaluation or are aware of the results. We ask that you read this form and ask any questions you may have before agreeing to be in the study. This study is being conducted by: Frances Lawrenz and Jean King at the University of Minnesota with the assistance of graduate students Lija Greenseid, Kelli Johnson, Stacie Toal and Boris Volkov.

Background Information

The purpose of this study is to investigate the possible influence program evaluation results.

Procedures

If you agree to be in this study, we would ask you to do the following things: Answer questions about the use of evaluation results from the [**NAME]** evaluation. Your answers to these questions will be recorded and transcribed but your name will not be directly associated with comments in any reports. The interview should take 30-60 minutes. You may also be asked to provide us with any artifacts demonstrating the influence of the evaluation on your project.

Risks and Benefits of being in the Study

The study has the risk that thinking about the evaluation results may bring up old memories or frustrations. You will also be giving up some of your time. The benefits to participation are that you will be helping us to understand the extent and type of influence of program evaluations so that they may be better designed in the future.

Confidentiality

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records. Tapes will be stored in a locked file and erased after transcription. Only the researcher and the transcriber will have access to the original tape.

Voluntary Nature of the Study

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota or the National Science Foundation. If you decide to participate, you are free to not answer any question or withdraw at any time without affecting those relationships.

Contacts and Questions

The researchers conducting this study are: Frances Lawrenz and Jean King. If you have any questions, you are encouraged to contact them at University of Minnesota, 612-625- 2046, lawrenz@umn.edu or kingx004@umn.edu. If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher(s), you are encouraged to contact the Research Subjects' Advocate Line, D528 Mayo, 420 Delaware St. Southeast, Minneapolis, Minnesota 55455; (612) 625-1650.

INTERVIEW PROTOCOL

Begin with IRB information. Get their consent to be interviewed recorded on the tape.

Introduction

Thank you for agreeing to talk with us about your experiences with the ATE program evaluation. First, so we can have it on the tape, would you please introduce yourself by stating your name and your relationship to your ATE project at the time of the program evaluation? And how long were you involved with the ATE program evaluation?

Clarifying Questions

First, we want to clarify that we will be asking questions about your experiences with the XXX program evaluation that took place between XXX. You may remember there were two evaluations associated with your grant: an individual project evaluation that you or your project was responsible for, and the broader program evaluation conducted XXX.

As part of the program evaluation you may remember the {items specific to program evaluation)

1. Is it clear that we want to focus on the program evaluation and that you understand what the program evaluation is?
2. By the way—do you remember taking a survey from us about this same topic last fall? (If yes) Thinking back on the survey, did you understand then that we were asking questions about the program evaluation?

Program Evaluation Involvement

1. We have two main areas to discuss today: involvement and impact. Let’s begin by having you describe the ways in which you and your project were involved in the program evaluation activities.
2. Could you please describe how you and your project were involved with program evaluation activities such as requesting the evaluation, designing the parts of the program evaluation, collecting data, or analyzing and reporting findings?

Probes, if necessary:

- Were you involved in the planning stages of the evaluation?
- Did you provide input on evaluation questions the program evaluation should address?
- Did you help in developing or field testing the survey instruments?
- Did your project provide data for the annual survey?
- Did you help create the site visit protocol?
- Did you review or draft any reports of the program evaluations findings?

3. What motivated you to get involved in these ways?
4. To what extent do you feel the input and/or assistance you provided was used, or wasn’t used, by the program evaluation staff?

5. Overall, how would you rate your level of involvement in the program evaluation? Would you say you were not involved, involved only a little, involved some, or involved extensively?

Impact

Now we’re going to ask you some questions about the effect the program evaluation had on your project, followed by questions about the impact on you, and then about broader impacts.

Project Level Impact

1. First, in what ways, if any, did the program evaluation have an impact on your project?

   a. Did you use the data provided by the evaluation in your project’s evaluation?
   b. Did you use any of the findings from the program evaluation to make decisions about your project? If so, which ones?
   c. Did the program evaluation findings spark any conversations within your project or among your project’s stakeholders?
   d. Did you use the findings to advocate for support for your project?
   e. Did you take any of the action steps recommended in the targeted brochures, for example, sustainability, material development, professional development, etc?
   f. Did the program evaluation process affect the way the project staff interacted?
   g. Have you used the site visit protocols or the planning guide?
   h. Were there other ways in which you used the program evaluation findings?

2. Overall, how would you rate the level of impact the program evaluation had on your project? Would you say there was no impact, only a little impact, some impact, or an extensive amount of impact?

Individual Level Impact

Next let’s focus more closely on the effect that the program evaluation process had on you (OR: each of you) personally. In what ways, if any, did the program evaluation process have an impact on you?

3. Did you learn new evaluation skills? (Probes, specific to program.)
4. Did you learn any new knowledge about your project as a result of the evaluation?
5. Did you feel differently about the importance of evaluation?
6. Overall, how would you rate the level of impact the program evaluation had on you? Would you say there was no impact, only a little impact, some impact, or an extensive amount of impact?

Factors Affecting the Program Evaluation's Impact

7. OK, so you've described the different ways in which the program evaluation had an impact on you and your project. What do you think were the reasons that the study had the impact it did?

Probes:

- Anything else that inhibited the impact the program evaluation had on you and your project?
- Anything else that enhanced the impact the program evaluation had on you and your project?

Broader Influence of the Program Evaluation

Now that we have talked about the impact that the program evaluation had on you and your project we'd like to look at the effects of the program evaluation more broadly.

1. First, however, what did you think was the goal of the program evaluation?
2. Can you think of any ways in which the program evaluation had an effect on the STEM education field?
3. What about any effects of the program evaluation on the evaluation community?
4. Were there any other broad impacts of the program evaluation, from your perspective? If no—do you think it's realistic to expect broad impacts from these types of large-scale program evaluations?
5. What do you believe it would have taken for the program evaluation to have had a greater impact on the field?

Wrapping-up

1. In sum, what do you feel is the most important lasting effect of the program evaluation, if any?

2. Are there any other issues regarding the impact of the program evaluation that you think are important for us to hear?

EXPERT PANEL MEMBERS

Dr. Jean King

Jean King is a professor in the Department of Educational Policy and Administration at the University of Minnesota where she serves as Coordinator of the Evaluation Studies Program. She holds an A.B., M.S., and Ph.D. in from Cornell University. In 1995 her professional evaluation activities using participatory methods resulted in the Myrdal Award for Evaluation Practice from the American Evaluation Association, and in 1999 she was awarded the Association's Ingle Award for Extraordinary Service.

Dr. Frances Lawrenz

Frances Lawrenz is a professor in the Department of Educational Psychology and Associate Vice President for Research at the University of Minnesota. Her major research focus is science and mathematics program evaluation. Her evaluations utilize a variety of techniques to best fit the needs of a given situation and usually involve mixing methods in a variety of ways.

Catherine Callow-Heusser, PI, MSP-RETA at Utah State University

Catherine Callow-Heusser is the director and owner of EndVision Research & Evaluation, 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 N. Gullickson, PI, ATE Program Evaluation

Arlen Gullickson has served as chief of staff of Western Michigan University's (WMU) Evaluation Center since 1991, and 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.

Iris R. Weiss, PI, LSC Core Evaluation

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.

Michael Quinn Patton, Advisory Board Member

Michael Quinn Patton is an organizational development and evaluation consultant. After receiving his doctorate in Sociology from the University of Wisconsin, he spent 18 years on the faculty of the University of Minnesota (1973-1991), including five years as Director of the Minnesota Center for Social Research and ten years with the Minnesota Extension Service. Dr. Patton has worked with local, state, federal and international organizations AND is the author of six evaluation books including Utilization-Focused Evaluation: The New Century Text (1997). The two previous editions of UFE: The New Century Text has been used in over 300 universities.

Marv Alkin, Advisory Board Member

The project is evaluated by an external evaluator, Dr. Marvin Alkin, Professor Emeritus of Social Research Methodology at the UCLA Graduate School of Education and Information Studies. His research and theoretical interests include work on evaluation utilization, evaluation theory, and problems of evaluating educational programs.

SUMMARY THINK-ALOUD INTERVIEW FEEDBACK

QUESTION/PROMPT	Interviewee 1	Interviewee 2	Interviewee 3
Was it mentally tiring?	No, it was not.	No.	Somewhat – I sometimes had to expend more effort to figure out what the nuance or slight difference was between similar questions.
How was the length?	Since it was mostly multiple choice questions, the length was OK.	The length was appropriate as the format allowed me to move through the survey quickly.	It seemed a little long but still appropriate for its purpose.
Did it make sense overall?	Yes, it made good sense.	Yes.	Yes – the topics matched what I had experienced in evaluation involvement, use, and influence.
Could you see the logical structure?	Yes, a clear logical structure.	Yes.	Yes – the sequence of the topics and the questions within each topic made sense and appeared to represent the "next level" of questions based on previous topics.
Did you have questions related to the concept of involvement? If yes, what?	No.	No.	No, I did not have questions, because I understood the concept of involvement based on my previous experience.
Did you have questions related to the concept of use? If yes, what?	No.	No.	No, I did not have questions about the concept of use for the same reason as above – I had understanding of this concept based on my previous experience.

ROTATED FACTOR MATRIX

EFA - 20 use items	Factor		
	1	2	3
25 increased knowledge/understanding of evaluation planning	**.691**	.381	.359
26 increased knowledge/understanding of eval implementation	**.769**	.285	.363
27 increased knowledge/understanding of communicating findings	**.727**	.357	.357
29 knowledge/understanding of STEM education evaluation	**.685**	.391	.261
30 knowledge/understanding of my project	**.691**	.348	.226
31 skills in planning an evaluation	**.699**	.399	.379
32 skills in implementing an evaluation	**.725**	.302	.429
33 skills in communicating evaluation findings	**.666**	.389	.465
35 skills as a STEM education evaluator	**.663**	.341	.417
36 skills for working on my project	**.703**	.397	.262
38 belief in the importance of planning an evaluation	.312	**.824**	.202
39 belief in the importance of implementing an evaluation	.303	**.848**	.227
40 belief in the importance of communicating evaluation findings	.347	**.821**	.206
41 belief in the importance of STEM education	.382	**.607**	.183
42 belief in the importance of STEM education evaluation	.390	**.717**	.175
45 used what I learned from planning in another evaluation	.346	.260	**.738**
46 used eval plan as a model in another evaluation	.219	.264	**.798**
47 used implementation knowledge in another evaluation	.395	.155	**.787**
48 used data collection instruments in another evaluation	.273	.019	**.709**
49 used communication info in another evaluation	.266	.356	**.711**

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.