1-2011

An Analog Experiment Comparing Goal-Free Evaluation and Goal Achievement Evaluation Utility

Brandon W. Youker
Western Michigan University

AN ANALOG EXPERIMENT COMPARING GOAL-FREE EVALUATION AND GOAL ACHIEVEMENT EVALUATION UTILITY

by

Brandon W. Youker

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation
Advisor: Chris L. S. Coryn, Ph.D.

Western Michigan University
Kalamazoo, Michigan
December 2011

AN ANALOG EXPERIMENT COMPARING GOAL-FREE EVALUATION AND GOAL ACHIEVEMENT EVALUATION UTILITY

Brandon W. Youker, Ph.D.

Western Michigan University, 2011

Goal-free evaluation (GFE) is the process of determining the merit of an evaluand independent of the stated or implied goals and objectives, whereas goal achievement evaluation (GAE), as the most rudimentary form of goal-based evaluation, determines merit according to the evaluand's level of accomplishment with regard to its goals. This study examines the utility of GAE and GFE from the perspective of the evaluation's intended users. In the study, two evaluation teams, goal achievement and goal-free, independently and simultaneously evaluate the same human service program. Each team produced a final evaluation report, which was read by the evaluation's users, who then responded to questionnaires regarding the reports' usefulness and later interviews. The questionnaire results were that 66% of evaluation users scored GAE more favorably versus 33% who scored GFE higher. The results of the interviews were that 40% of evaluation users found GAE more useful, with 20% claiming GFE more useful; the remaining users were undecided or felt the approaches equal. The conclusion is that differences between the two evaluation reports exist; however, it is not apparent as to whether these differences are caused by implementing GAE or GFE. Furthermore, the effects or differences that did present between the evaluations were small and not

practically significant enough to definitively claim one approach clearly more useful to these evaluation users.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ACKNOWLEDGMENTS

The development of this dissertation has benefited from both the advice and criticisms of Professors Chris L. S. Coryn, David Hartmann, Liliana Rodríguez-Campos, Michael Scriven, James Sanders, and Michael Q. Patton. Further assistance was provided by P. Cristian Guigu, Amy Gullickson, Wes Martz, Daniela Schröeter, Lori Wingate, Sally Veeder, and Mary Ramlow. I would also like to recognize the challenges and opportunities provided by the Evaluation Center at Western Michigan University.

Appreciation and recognition is due to the partnering organizations who allowed their program to be evaluated and their staff to participate, and who participated themselves.

Lastly, a heartfelt thanks goes to my family, who has supported me academically, professionally, and personally. In particular, thank you to my parents, Bruce W. Youker and Christine A. Youker, and my father-in-law, David K. Caldwell. Finally, the love and encouragement of my wife, Elizabeth, and daughter, Vera, continue to motivate me both in my scholarly pursuits and, more importantly, in life.

Brandon W. Youker

TABLE OF CONTENTS

Table of Contents—Continued

CHAPTER

Table of Contents—Continued

CHAPTER

Table of Contents—Continued

CHAPTER

Table of Contents—Continued

LIST OF TABLES

List of Tables—Continued

List of Tables—Continued

LIST OF FIGURES

CHAPTER I

INTRODUCTION TO THE PROBLEM

This chapter begins with a description of the problem studied in this dissertation including its background and situation. Next, the specific purpose of the study conducted to investigate and partially resolve the stated problem is described. Additionally, the outline presents the specific questions investigated, an explicit delineation of the research problem and the hypotheses formulated and tested, and the dissertation's importance. Finally, the chapter concludes with an outline of the dissertation's remaining chapters.

Background of the Problem

Historically, evaluation has been a normative endeavor where scholars, theorists, and practitioners prescribe untested theories of evaluation (Coryn, Noakes, Westine, & Schröter, 2011; Friedman, Rothman, & Withers, 2006; Hellström & Jacob, 2003; House, 1983; Patton, 1997; Scriven, 1973, 1991; Shadish, Cook, & Leviton, 1991; Stufflebeam, 2001; Thiagarajan, 1975; Vedung, 1997). As Tourmen (2009) states, "There is an abundant literature aimed at theorizing and prescribing evaluation practice" (p. 7); however, there are few empirical studies on evaluation. In fact, as Henry and Mark (2003) assert in their article "Toward an Agenda for Research on Evaluation,"

> Prescriptive advice and admonitions about how to do evaluation have been plentiful, filling books, journals, conferences, e-mails, and conversations. But these are generally based on personal experience, observation, and the individual's sometimes idiosyncratic beliefs and values—not on carefully gathered evidence that can be described, shared and critiqued. (p. 70)

Smith (1993) recognizes that "empirical knowledge about the practice of evaluation is essential for the development of relevant and useful evaluation theories" (p. 237) and thus "there is a need to identify which theoretical claims in fact presuppose testable empirical fact" (p. 240). Smith asks evaluators

> for increased empirical study of evaluation practice to describe the nature of actual practice; to compare the feasibility and effectiveness of alternative models, methods, and theories; to provide a basis for the development of descriptive evaluation theories; and to assess the utility of prescriptive theories. (p. 238)

Shadish et al. (1991) claim that "evaluation will be better served by increasing the more systematic empirical content of its theories" (p. 483); they add that such efforts

> have always been relatively rare in evaluation because so little effort is generally put into developing empirically testable hypotheses based in evaluation theory, and because so few evaluators are both interested in the topic and in a position to undertake such studies. (p. 484)

The goal-based evaluation (GBE) approach (sometimes referred to as objectives based evaluation) is a prime example of an evaluation approach that continues to dominate evaluation practice despite few empirical studies of its merits in comparison to non-goal-based approaches. Goal-orientation domination has existed since Ralph Tyler developed his objectives-based evaluation approach in the 1940s (Alkin, 2004a; Fitzpatrick, Sanders, & Worthen, 2004). In fact, Friedman et al. (2006) report that "as evaluation emerged as an independent field within the social sciences, it became closely identified with the measurement of goal attainment" (p. 201); thus, there is a plethora published on goal-based approaches and their methods. This point is reiterated by Mark, Henry, and Julnes (2000), who, in discussing early program evaluation, stated that GBE was the dominant methodological paradigm in evaluation as:

> Explicit program goals were converted to measurable objectives, these were tested, and then the program's performance was compared to the objectives. In

> this approach the evaluator's role was thought to be simply to test fact-based claims that originated in statements about program or policy goals; the complex issue of which outcomes should be selected for evaluation and why.…By sidestepping this issue, early evaluators implicitly preempted debate on any additional effects or side effects that might bear on the worth of the program. (p. 33)

Further demonstrating the fact that early on there was a general acceptance of objectives-based evaluation are the examples of scholars who have furthered GBE-related theories and methodologies (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Campbell & Stanley, 1963/1966; Chen & Rossi, 1983; Cook & Campbell, 1979; Cronbach, 1963, 1982; Metfessel & Michael, 1967; Popham, Eisner, Sullivan, & Tyler, 1969; Suchman, 1967, 1969).

However, there existed a handful of evaluation scholars such as Cronbach (1963), Scriven (1967), and Stake (1967), who began promoting evaluative inquiry beyond simple goal achievement and introduced some of the limitations associated with pre-specified goals and objectives. They argued that the assessment of goal achievement is only part of the evaluation process as the evaluator also has a responsibility to explore side effects (Stake, 1967). At the time, the authors failed to acknowledge the fact that the promotion of the evaluator's search for side effects, albeit logical, is itself prescriptive in nature and worthy of empirical study.

In the early 1970s, Scriven (1972) introduced a radical concept that urged evaluators to specifically avoid focusing on program goals or objectives while conducting their evaluations. He called his counter to the goal-based approaches, goal-free evaluation (GFE). Prior to the introduction of GFE, there was little challenging or questioning of the goal- and objective-oriented evaluation paradigm; there was no proposed alternative. Scriven's GFE was the theoretical alternative approach and a number of his publications

proclaimed its logical soundness and methodological strengths (Scriven, 1972, 1973, 1976, 1991).

For a few years following GFE's introduction, there was mild interest in the approach amongst evaluation scholars. The majority of the GFE literature consists of philosophical debates regarding its logic, strengths, weaknesses, and feasibility (e.g., House, 1980; Salasin, 1974; Scriven, 1972, 1973, 1974b, 1991). Even today, many evaluation textbooks contain short blurbs about GFE, primarily discussing it from a hypothetical or theoretical perspective in a single paragraph (e.g., Fitzpatrick et al., 2004; Grinnell & Unrau, 2008; Patton, 2002a). That said, articulation of specific methodologies for conducting GFE remains nearly non-existent; and nearly half a century since its introduction, GFE has remained conceptually abstract and highly theoretical in the minds of most evaluation scholars with very few known practitioners and even fewer who have written about it. There still is only one known attempt at an empirical investigation of GFE, a doctoral dissertation (Evers, 1980). So as Tourmen (2009) put it, why without scientific study "would they [evaluators], for example, prefer one method to another?" (p. 7). Thus, it can be reasonably concluded that those who support or oppose GFE do so on the basis of ideology rather than empirical evidence.

Lastly, evaluation scholars tend to agree that the evaluator maintains an ethical obligation to consider evaluation utility with every evaluation. Virtually nothing is known as to GFE's utility or lack thereof. Emphasis on evaluation use is justified based on the existing moral imperative for all evaluators to attempt to "ensure that an evaluation will serve the information needs of the intended users" (Joint Committee on Standards for

Educational Evaluation, 1994, p. 23). Discussion of evaluation utility will be furthered in Chapter II.

Statement of the Problem Situation

Without empirical knowledge regarding the comparative advantages and disadvantages of GBE versus GFE, the evaluator or the prospective evaluation user is less capable of making informed decisions in choosing the appropriate evaluation approach. Hence, the status quo is maintained without critical reflection or investigation into whether preferable alternatives exist. Program administrators and managers as well as external evaluators continue to employ sometimes highly sophisticated goal-oriented approaches regardless of whether the approach is best for providing the needed information or answering the important evaluation questions. Lastly, few studies have rigorously examined distinct evaluation approaches and their utility from the perspective of the evaluator's consumer (i.e., those who hire evaluators and who are supposed to use the results from an evaluation).

Purpose of the Study

The overarching purpose of this dissertation is to collect quantitative and qualitative data for comparing the utility of two opposing evaluation approaches and therefore contributing to the body of knowledge on both approaches. Two prescriptive theories are contrasted via an experimental analog study: (1) goal achievement evaluation (GAE), a sub-type of GBE that solely examines stated goals and objectives; and (2) GFE. Both of these approaches presuppose certain benefits to the evaluation user that are worthy of systematic scrutiny. Thus, there is a pragmatic aspect of this study which is to

examine whether GFE has value to the intended users of evaluation by examining its utility.

Secondly, this dissertation is intended to contribute to the limited knowledge about GFE in general. There is little written on GFE. When GFE is discussed, the literature lacks practical information as to the details of how to actually conduct a GFE (i.e., such writings "tell one what to do, but not how to do it" (Coryn et al., 2011, p. 206). Since Scriven introduced GFE, he has been promoting its use, claiming logical and practical benefits when compared to goal-based approaches, but only an unpublished doctoral dissertation has examined this claim empirically. Scientific investigation of these two evaluation approaches has the potential to increase evaluator credibility when the evaluator attempts to convince an evaluation client of a particular evaluation approach's potential for contributing to the betterment of the evaluation client's program or policy (Henry & Mark, 2003). Referring to such studies, Scriven (1974b) wrote, "it will take only a few such experiments…to give us a good picture of GFE. I think its value will be demonstrated if it sometimes picks up something significant at a cost that makes the discovery worthwhile" (p. 47). Unfortunately, these studies have never come to fruition. Therefore, this study is an attempt to rekindle the debate and study of GFE to determine its relative merit and worth. It is *not* designed to encourage evaluators to abandon GBE in favor of GFE. "Evaluation will not be well served by dividing people into opposing camps: pro-goals versus anti-goals evaluators" (Patton, 1997, p. 184). Rather this dissertation adds to the sparse literature on GFE while offering some balance between the two approaches through the systematic examination of both.

Objectives to Be Investigated

Three specific objectives are investigated in this dissertation. They are as follows:

1. From the perspective of evaluation users, is there a difference between GAE and GFE with regard to utility?
2. What, if any, are users' perceived differences in utility between GAE and GFE? If differences do exist, how do they differ specifically in terms of instrumental use, conceptual use, and persuasive use?
3. If differences in perceived utility exist, what explains those differences?

Thus, the first question seeks to determine whether there is, in fact, any perceived difference in utility. The second question seeks to determine what those differences are. Finally, the third question seeks to explain the reasons for any perceived differences.

Conceptual and Substantive Assumptions

There are two crucial and related postulates underlying the assumptions of this study. The first is that evaluation utility is worthy of study and has a logical connection to what it means to be a quality evaluation. The second is that goal-based approaches, such as GAE, continue to be prevalent in professional evaluation and therefore adequately serve as a reasonable comparison to GFE. The following is a description of both assumptions.

Assumption#1: Evaluation Utility Is Worthy of Study

Evaluation utility is worthy of systematic examination primarily because of its relationship with what it means to conduct a "good" evaluation. There are a number of

checklists and guidelines that assist evaluators in determining the merit(s) of an evaluation. According to Yarbrough, Shulha, and Caruthers (2004), possibly the most widely accepted checklists for evaluation is the Joint Committee on Standards for Educational Evaluation's (Joint Committee) 1994 publication, *The Program Evaluation Standards* (PES). In the PES, the connection between the use of evaluation and overall quality of that evaluation is so fundamental that utility is identified as one of only four program evaluation standards. The Joint Committee's emphasis on evaluation utility alone represents a significant endorsement for further study of evaluation utility.

Additional evidence that evaluation utility is worthy of consideration is the consistent dialog among academics regarding evaluation utility. Evaluation use was introduced in the late 1960s and the discussion has continued on relatively consistently to today. For instance, Amo and Cousins (2007); Eisner (1979b); Mohan, Tikoo, Capela, and Bernstein (2006); Patton (1988, 1997, 2007); Preskill, Zuckerman, and Matthews (2003); Scriven (1972, 1991); Shulha and Cousins (1997); and Weiss (1967, 1988) are just examples from a number of publications that consider evaluation utility.

Several of the publications cited in the preceding paragraph deal with how to increase evaluation use, yet there is an ongoing debate as to whether or not the evaluator is ultimately responsible for the actual use of an evaluation (Patton, 1988, 1997; Scriven, 1991, 2005b; Shulha & Cousins, 1997; Weiss, 1988, 1998). In fact, both Scriven (1991, 2005b) and Weiss (1988, 1998) caution against judging an evaluation (or evaluator) based on the actual use of the evaluation. For example, referring to her 1988 paper presented at the American Evaluation Association, Weiss (1998) said:

> [E]valuators should not be held accountable for failures to use their results. Even when program staff know about the findings, understand them, believe them, and

> see their implications for improving the program, many factors can interfere with their using results for program improvement. (p. 22)

Weiss' opinion above is, at least in part, analogous to the old adage that "you can lead a horse to water but you can't make him drink." She concludes that the goal is "*effective* utilization of evaluation not necessarily more utilization" [emphasis added] (1998, p. 30). Scriven (1991) also considers use from the evaluator's perspective, which led him to grow concerned that when use is highly emphasized a conflict of interest is created as the evaluator is pressured to "adjust the findings to what decision-makers are willing to do rather than what they should do" (p. 371).

Nonetheless, evaluators overwhelmingly agree that they do in fact possess some degree of responsibility for the use of their evaluations; and that "none of these cautions are meant to suggest that no one should study or attempt to increase utilization" (Scriven, 1991, p. 371). Therefore, the question is not whether evaluation utility is important to evaluators but rather how important is utility. The assumption is that evaluation utility is always of some relative importance to the evaluator and in all evaluations; and this, in and of itself, warrants the study of evaluation use.

Assumption#2: GAE Is a Commonly Used Evaluation Approach

The second assumption in this dissertation is that despite the heavy criticism of GAE in contemporary evaluation literature, GAE continues to be prevalent among evaluators and program managers. Christie and Alkin (2005) assert that:

> Objectives-based evaluation approaches continue to be used to guide evaluation practice (in addition to the relatively widespread use of criterion-referenced tests used to measure student performance). For example, the World Bank's Operations Evaluation Department uses an objectives-based evaluation approach to evaluate development work. (p. 285)

Some specific supporters of GAE include Zink (2001), and Mauk and Schmidt (2004), who promoted GAE in the evaluation of nursing; DePanfilis and Salus (1992), who used GAE for evaluating child abuse prevention programs; and Lee and Ahn (2004), who assessed housing programs via GAE. However, GAE does not only exist in the world of external professional evaluation, but nearly all internal evaluations are goal-based in nature. Therefore, GAE is a reasonable approach to compare with GFE, not only because it serves as a polar opposite to GFE, but also because it is an actually practiced approach to program evaluation. In conclusion, this study continues on the presumptions that evaluation utility does matter, and that GAE serves as a suitable comparison approach.

Analog Studies

This study contributes to the dearth of empirically-derived and tested evaluation theory via an experimental analog. Analog studies are controlled studies that are designed to approximate real-life evaluation practice settings while allowing some degree of experimental control in testing a hypothesis about a potential influence on evaluation practice or outcomes (Henry & Mark, 2003). There are six types of analog studies in evaluation which can be used in various combinations with each other. These analog studies fall into two categories: (1) evaluand-directed analog studies, and (2) outcome-directed analog studies. Evaluand-directed studies refer to whether the evaluand is mock (i.e., fake or fabricated) or real (i.e., actual), while outcome-directed analog studies refer to the use of mock evaluand outcomes and mock evaluation outcomes. In this dissertation

an analog study using a real evaluand, with real evaluand outcomes, and real evaluation outcomes is used to study GAE and GFE utility.

Probably the greatest limitation of analog studies is external validity or the generalizability of the study's findings beyond the actual study sample and to real-world evaluands and evaluations. Real outcomes differ, the evaluation situations and environments differ, the training and prior experience of evaluators differs, and the consequences for evaluands differ substantially between the field and the laboratory. Consequently, proper study protocols for maintaining the fidelity of the study are of the utmost importance. Another limitation is that when using real evaluands, the evaluand's willingness to participate in an analog evaluation study is probably systematically different than an evaluand who is not willing to participate. The willing evaluand may be more mature, more evaluation savvy, and more confident in its performance and outcomes.

Even so, there are several strengths of analog studies. The primary benefit of analog studies is the control and flexibility offered to the investigator. The investigator may select particular subject groups (e.g., graduate students, community members, theatrical actors, or a real evaluand), and standardize evaluation procedures for all study participants, and/or systematically varying outcomes. With this control, the investigator can also directly compare the effects of variations in evaluation techniques, measures, information given to subjects, analysis methods, and synthesis methods among others.

Fidelity

Fidelity is defined as "the extent to which delivery of an intervention adheres to the protocol or program model originally developed" (Mowbray, Holter, Teague, & Bybee, 2003, p. 315). In this case, the program model is the particular evaluation approach (i.e., GAE and GFE) and the intervention is the evaluation itself. In an analog study, which examines two evaluation approaches, fidelity to the specific approach is essential because if the fidelity to the respective approach is not maintained by the evaluator, the internal validity of the entire study is jeopardized.

Articulating fidelity to an evaluation model or approach is practically difficult. Smith (1994) points out that in actual evaluation practice there is a lot of vagueness in what it means to follow a model:

> Although evaluators sometimes speak of designing evaluations which follow a particular model, their language generally refers not to instrumental application of procedural specifics, but to the selection of an overall orientation or approach. Because evaluation models are not procedurally prescriptive, are subject to varied interpretations, are mute on many of the details required to implement an evaluation, and must be operationalized within the demands of a specific context of application, many decisions are left to the evaluator's professional judgment in spite of the prior selection of a given model. No one study can thus be argued to be the epitome of a given model, and many quite different studies are arguably appropriate versions of the same model. (p. 4)

If someone does not document and/or measure the evaluators' adherence to an intended model, the consequence, according to Chen (1990), is that there is no reliable way to determine whether poor evaluative conclusions reflect a failure of the evaluation model or failure to implement the evaluation model as it was intended. Smith (1993) also notes that "if evaluation theories cannot be uniquely operationalized, then empirical tests of their utility become increasingly difficult" (p. 240). Therefore, establishing fidelity

criteria, and the ability to measure adherence to them, enables interventions "to be more standardized, consistently researched, and replicated" (Mowbray et al., 2003, p. 317) as well as confirming "that the manipulation of the independent variable occurred as planned" (Moncher & Prinz, 1991, p. 247).

Statement of Hypotheses

There is one specific hypothesis investigated in this dissertation: there is no practically significant difference in evaluation utility between GAE and GFE. Subsumed under this hypothesis are three dimension of evaluation utility: instrumental utility, conceptual utility, and persuasive utility. Stated propositionally, the alternative hypotheses are that there is a practically significant difference in instrumental, conceptual, and persuasive utility between GAE and GFE from the perspective of the evaluations' users, while the null hypothesis is that there is no practically significant difference between GAE and GFE with regard to evaluation utility from the perspective of evaluation users. In conventional notation, the null and alternative hypotheses postulated for evaluation utility can be represented as:

$$H_0 : \text{GAE} = \text{GFE}$$

$$H_1 : \text{GAE} \neq \text{GFE}$$

The specific direction (e.g., GAE > GFE; GAE < GFE) of the alternative hypothesis is two-tailed as there is no prior knowledge as to whether one approach should have more instrumental, conceptual, or persuasive utility than the other. That is, one cannot reliably claim whether GFE should have either greater or lesser utility than GAE. Although, one might hypothesize that GAE should be more useful to evaluation users as

it focuses on the goals toward which the evaluation users are supposed to be directing their efforts, this is speculation and is not proposed in this study.

Importance of the Study

In this dissertation the perceived utility of GAE versus the utility of GFE from the perspective of the evaluation user is investigated. In empirical terms, GFE is *terra incognita* because so little has been published on its procedures and methods, or its actual limitations or actual benefits. Therefore, this study also seeks to generate knowledge which is substantive, theoretical, and methodological in nature, moving from *a priori* prescriptions about evaluation to an *a posteriori* position. Through providing evidence about evaluation and ultimately knowledge of evaluation itself, this study contributes to the "empirical basis for improving practice and enhancing our understanding of the types of evaluation most likely to move us toward social betterment" (Henry & Mark, 2003, p. 70).

Consequently, and more specifically, this study should provide systematically derived information that can be used to refine, revise, and extend current knowledge regarding GAE and GFE as well as potentially contribute to and influence scholarly research on evaluation and evaluation theory, methodology, and practice. By studying and attempting to answer these questions, this study seeks to help evaluation practitioners and consumers make more informed decisions. The mission of this dissertation is not to end the debate as to the utility of GAE (and other goal-based approaches) versus GFE, but rather to initiate it from an empirical basis.

Chapter Summary

The first chapter of this dissertation introduces the problem studied in this dissertation and how this dissertation aims to address this problem via an analog study of GAE and GFE utility. The specific research questions, hypotheses, and study significance are also included in this chapter. Below is a description of the dissertation's chapters.

Outline of the Dissertation

Chapter I describes the problem studied in this dissertation. In Chapter II the relevant literature, with a specific emphasis on GAE and GFE, is reviewed and synthesized. The methods used to study and contrast GAE and GFE utility are presented in Chapter III. The study's findings are presented in Chapter IV; and its conclusions, implications, limitations, and directions for future research are addressed in Chapter V.

CHAPTER II

LITERATURE REVIEW

The second chapter offers an examination of GFE found in the academic and professional evaluation literature. The purpose of this chapter is to present the history, theory, and logic of GFE; to describe the rationale underlying an analog study of GFE utility; and to identify the strengths and weaknesses of GFE. As previously stated, there are relatively few articles published specifically related to GFE; therefore, this dissertation begins by presenting a comprehensive literature review beginning with the sole empirical study of GFE.

Previous Empirical Studies on GFE

The only formal study examining GFE was a doctoral dissertation conducted by Evers (1980). Evers compared "the relative utility of operationalized versions of goal-free and goal-based evaluation techniques" (p. 2) while evaluating several four-year colleges in the Midwest and Northwest U.S. Evers randomly sampled from 31 nationally recognized evaluators until he selected three evaluators for the goal-based team and three for the goal-free team. Each team attended a one-day orientation and training; one team received training in GBE and the other GFE. Neither team was aware of the other team's existence. The study assessed the evaluators' rapport with the evaluand project director, the use of the evaluators' time, the expectations of the evaluators and the project director, the overall satisfaction of both the evaluators and the project director, and the evaluators'

confidence with implementing their respective evaluation approach. In measuring the utility of the evaluation reports from the evaluees' perspective, Evers used semantic differential rating scales. The instrument he constructed consisted of 58 bipolar adjective pairs on a seven-point scale. Respondents rated the degree to which the report was active to passive, logical to illogical, consistent to inconsistent, scholarly to ignorant, and so on. Evers chose semantic differential scaling, in part, because Osgood, Suci, and Tannenbaum (1957) demonstrated that bipolar pairs of adjectives yield reliable findings which highly correlate with alternative measures of the same attitude.

Evers (1980) found that "evaluators can be trained to use such a goal-free approach and that the training can carry over to differences in the on-site evaluation process" (p. 68). However, his overall conclusion was that the evaluee ratings of the evaluation reports' utility did not significantly differ.

One of the limitations of Evers' study was that in his investigation of evaluation utility, the evaluators did not examine the same evaluand; rather his goal-based and goal-free teams evaluated separate evaluands. In other words, one might find the GFE report to be very useful at one site yet not know whether a goal-based approach would have proven even more useful had it been the approach used to evaluate that particular evaluand. Nonetheless, Evers dissertation represents the only known attempt at systematic study of GFE.

Evaluation Utility

As previously mentioned in Chapter I, professional evaluators have been attempting to increase the use of and the usefulness of evaluation for decades. In

continuation of this tradition, an important aspect of this study is to ascertain the utility of GFE; but before delving into how an evaluation can be used, an explanation of who an evaluation is useful to is presented.

One of the most ardent supporters of evaluation's utility, Patton (1997) states that what is typically meant when one refers to evaluation use is the "actual primary intended user and their explicit commitment to concrete, specific uses" (p. 21). Furthermore, most often the intended evaluation users of particular interest are the evaluand's upstream stakeholders, or those "who have invested time, effort, money, and/or egos in the design, development, and/or implementation of an evaluand" (Davidson, 2005, p. 249). These users tend to be what Weiss (1998) refers to as "program people" or program sponsors, designers, administrators, managers, practitioners, potential users, competitors of the evaluand, public officials, the community, and civil society. The following is a brief description of key potential evaluation users:

- Program sponsors: those who pay program's bills (e.g., individuals, organizations, foundations, taxpayers, etc.)
- Program directors: program administrators at organizational, local, state, and national levels
- Program practitioners: staff or service providers who are in direct contact with the program's clients/consumers
- Other potential users of the evaluation: managers of similar programs, state and federal officials and foundation officers not directly affiliated with the program, policy makers, social scientists, other evaluators, and the public at large

The rationale behind focusing on these upstream stakeholders, rather than downstream stakeholders (i.e., consumers or other impactees directly and indirectly affected by programs), is that the upstream stakeholder is the "evaluation user who has the responsibility to apply evaluation findings and implement recommendations" (Patton, 1997, p. 21). As stated by Cousins (2004), "It is what the user chooses to do with evaluation findings that ultimately affect a wide range of others, not the least of whom would be intended program beneficiaries" (p. 392). Weiss (1998) concurs and adds that "professionals have the most direct opportunity to use results, and they are also likely to feel personally attached to the program and willing to invest the time in its evaluation" (p. 30).

In pondering evaluation utility, Weiss (1998) identifies five types of evaluation uses that lead to beneficial changes in or for the evaluand: (1) instrumental, (2) conceptual, (3) process, (4) persuasive, and (5) downstream.[1]

The first type of evaluation use is instrumental use, which represents the traditional or historic meaning of evaluation use referring to decision making, accountability, and improvement orientation. Examples of instrumental use include decisions to "end a program, extend it, modify its activities, [and] change the training of staff…" (Weiss, 1998, p. 23). To be of instrumental use, the evaluator needs to understand the evaluand and context, conduct a quality investigation, and effectively communicate the results (and, if appropriate, the recommendations).

The second type of use is conceptual use. Evaluation "findings can change the understanding of what the program is and does… [later, the program personnel can

---

[1] Downstream use has been so named by the author of this dissertation.

apply] their new conceptual understandings in instrumental ways" (Weiss, 1998, p. 24). Conceptual use leads to improved understanding of the program, the evaluation users, and its stakeholders. To be of conceptual use to the evaluand, the evaluator offers the evaluation clients generalizations about evaluand performance and effectiveness.

The third type of evaluation use is called process use and refers to the focused thought processes that are requested of the stakeholders involved with the evaluation. For example, "program staff who participate in defining and framing the evaluation begin to think more deeply about what they are trying to accomplish" (Green, 1988, as cited in Weiss, 1998, p. 25). Patton (1997) and Weiss (1998) point out that although the primary product of an evaluation is the evaluation report (and presentation), utility also refers to the evaluation process itself which includes the interactions between the evaluator and the stakeholders.

The fourth type of evaluation use, persuasive use, is based on mobilizing "support for a position that people already hold about the changes needed in the program" (Weiss, 1998, p. 24). Persuasive use is employed to legitimize positions, increase support and supporters, and rally them into action for change or for opposing a change. Furthermore, persuasive use may be used to increase stakeholders' sense of ownership within and over the program. Obviously, the integrity of the evaluation users frequently dictate whether persuasive use is employed in ways that benefit or fail to benefit the program; for example, program people can use an evaluation to persuade others to make unnecessary, unwarranted, unfounded, illogical, or short-sighted changes.

The fifth type of use might be called "downstream use." Downstream use refers to the "influence on institutions and events" (Weiss, 1998, p. 24) not directly affiliated with

the evaluand. For instance, the evaluation report may be read by a politician hundreds of miles away who promotes a legislative mandate based on the evaluation's findings; or program administrators of a critical competitor may read the evaluation report and initiate changes in their program based on another program's evaluation report. A final example of downstream uses includes the extrication of ideas or data from the evaluation report which is potentially relevant to the general public, to professionals in the field, and for scholarly publication.

Despite a lack of consensus on a singular definition of evaluation utility, nearly all agree that intended use to the intended user means that utility is context dependent. The definition of utility, therefore, somewhat relies on what the intended user finds to be or not to be useful. Furthermore, a significant barrier to defining utility is the passage of time often required before potential and actual uses of the evaluation are revealed (not to mention actual program effects); an evaluation may provide information which is immediately applicable, while some findings seem to reveal their utility only in hindsight.

There are potentially negative uses of an evaluation that are worthy of mention. According to Cousins (2004), an evaluation can be inappropriately not used and inappropriately used. First, evaluation users can inappropriately not use the data from an evaluation report when they should be using them. What Cousins refers to as "unjustified non-use" can be caused by mischievous or devious evaluation users who intentionally suppress evaluation findings or by users who simply make an honest mistake as ignorance or error leads them to disregard information that they should be using. Second, evaluation users can misuse an evaluation intentionally or unintentionally. Evaluation users mistakenly misuse an evaluation report when the users accept and use data and

conclusions from an evaluation report that is inaccurate, incomplete, incorrect, or otherwise faulty. Abuse of an evaluation occurs when the evaluation users misuse the evaluation to manipulate or coerce in hopes of producing inappropriate action or inaction in others.

Two examples of evaluation abuses and their relation with GFE are described below. These misuses can be considered unintended side effects of an external evaluation as they are not the intention of the ethical professional evaluator. A common negative use of evaluation occurs when program stakeholders use the mere existence of the evaluation as a bogus display of accountability or to present an appearance of substance and legitimacy. This false exposition or masquerading remains a significant criticism of internal evaluation in general as it represents a serious threat to internal evaluator credibility. The employment of GFE limits the evaluand and evaluation users' ability to feign legitimacy as the evaluator must be external to the evaluand and the evaluator is less likely to know how to "fake" it without possessing the knowledge of specific intentions (Scriven, 1976).

Teaching to the test, the third example, is typically considered a negative aspect of evaluation use. According to Weiss, when the evaluator chooses what specifics to investigate, s/he influences the evaluand as the evaluand tends to work only on the things or in the areas that will be on the test. In educational evaluation, this point is frequently made by those who oppose heavy reliance on standardized tests for the determining student or school achievement as they associate teaching to the test with a lack of flexibility and creativity. The other argument made by detractors of teaching to the test is that of opportunity cost. Teaching to the test encourages staff to focus only on things

deemed important even if it is at the expense of other critical things not being assessed. GFE can reduce the effects of teaching to the test as the goal-free evaluator is screened from the preordinate program objectives of which the tests are founded. Since the goal-free evaluator does not know the goals and intentions of the program, the program staff will only be able to "teach to the test" on areas where values and goals overlap (see Figure 2). Yet there is a positive aspect of teaching to the test; in essence, it emphasizes the areas that presumably matter and that are of particular importance. In other words, teaching to the test encourages the production of outcomes in the areas already deemed important. Furthermore, a secondary benefit of conducting a GFE is that it does not simply accept nor use the upstream stakeholders' goals and hence their tests rather the goal-free evaluator's inquiry helps determine whether the goals and objectives are meaningful in the first place.

Current Study

This dissertation examines immediate evaluation report utility; therefore, two of Weiss' (1998) five types of uses are excluded and therefore are considered beyond the scope of this study: (1) process utility, and (2) downstream utility. First, process use is not included in this dissertation study because of its emphasis on direct and open communication between evaluator and program staff, managers, and/or administrators, whereas the principles of GFE dictate that the evaluator maintain a high level of independence and distance from program personnel. Therefore, the goal-free evaluator is systematically prevented from fully pursuing process utility for fear of jeopardizing the goal-free nature of the evaluation. Second, although downstream use is a utility

dimension worthy of inquiry, it is beyond the scope of this dissertation study as this study exclusively examines utility which is acknowledged by the evaluation user shortly following the completion of the evaluation report. Thus, this study does not investigate use that becomes apparent much later.

For the remainder of this dissertation, the accepted definition of *immediate evaluation utility* is an evaluation that produces findings with relatively instantaneous use instrumentally, conceptually, and persuasively to the evaluation's intended users. The next three paragraphs describe and define the three latent constructs (or dimensions of utility) selected for examination. Below each dimension are its associated measured variables and definitions.

Instrumental use is the "direct, attributable use of the evaluation results to inform decisions" (Rogers, 2005, p. 74); it is implemental and represents the traditional meaning of evaluation utility. Instrumental use is measured by (1) program improvement which is a programmatic change for the better, progress in development, and/or a superior condition as compared to the previous condition; (2) decision making or the process of choosing among alternatives leading to a course of action; (3) accountability which is placing responsibility to someone or some group of people for an activity; (4) generalization about program performance which means the ability to take information, rules, and strategies learned about one situation and apply it appropriately to other similar situations about the program's operation, processes, or manner of functioning; and (5) generalization about program effectiveness which is the ability to take information, rules, and strategies learned about one situation and apply it appropriately to other similar situations about the program's ability to produce an effect.

Conceptual use is "the ways in which evaluation can have an impact on the way people think about the evaluand, on the issues" (Rogers, 2005, p. 74); it is of or relating to program concepts affecting the understanding of the program and stakeholders. Conceptual use is measured by (1) an understanding of what the program is and does; (2) an understanding of the program stakeholders and what they do; and (3) an understanding of the evaluation users' roles and responsibilities, which is comprehension of what the evaluation users do or should do with regard to the program.

Persuasive use refers to the ways in which an evaluation and it findings can have the power or influence to induce action or belief. Persuasive use is measured by (1) supporting a change which means aiding, backing, strengthening the cause or interests in making a change within the program; (2) opposing a change which is being against, contrasting, or resisting a change within the program; and (3) increasing stakeholder ownership in the program via increasing the stakeholders' sense of possession of or control within and over the program.

Finally, the above measured variables are considered *desiderata*, as opposed to *necessitata*, which means that an evaluation can have utility even when it fails to produce "useful" information in one or more of the total number of indicators (i.e., measured variables). For example, an evaluation's findings are shown to be of use in 9 out of the 11 indicators; thus, information on two of the measured variables was considered useless, or least not useful. In fact, hypothetically an evaluation may produce findings of use only on one indicator, yet if that information is believed to be particularly useful or valuable the evaluation as a whole is considered useful even though it did not produce utile findings on other variables. Nevertheless, on average, an evaluation deemed to be of use will

likely have more quantity and quality information on each measured variable as well as more coverage across all measured variables.

In summary, the previous section introduces some of the scholarly literature on evaluation utility. Additionally, it examines and describes evaluation utility and its three dimensions as employed in this study. The next section offers GFE's history.

The History of GFE

Scriven (1972) coined the term *goal-free evaluation* and formally introduced it as a potential program evaluation approach in his paper titled "Pros and Cons about Goal-free Evaluation"; yet the concept of a GFE existed unrecognized within professional evaluation for several decades in product evaluation and within informal evaluation for millennia. If one accepts the definition of evaluation as something like, the process of determining the merit, worth, and/or significance of something, and accepts GFE as an evaluation without particular reference to goals or objectives, then the history of GFE begins with the history of evaluation itself.

The next section of this chapter examines several key historical periods leading to the development and discovery of GFE. Six periods are discussed as they relate to GFE: (1) prehistory, (2) ancient history, (3) the European Renaissance, (4) Tylerian evaluation, (5) the Consumers Union, and (6) contemporary professional evaluation. This highly Eurocentric history is not meant to be a comprehensive history of all potential or actual uses of GFE but rather to establish that there is a long and relatively consistent history of GFE within the broader history of evaluation.

Prehistory

It is reasonable to speculate that the origins of an informal version of GFE predate written language in the applied fields of product and performance evaluation. Anthropologists and archeologists have documented a plethora of stone and bone tools and weapons dating back to our Eolithic ancestors. Over time and embedded in layers of sediment, these objects show refinement and improvement, hence evaluation (Scriven, 1991). A similar refinement in building quality and technique is evident in examining the construction stages of pre-historic Homo sapiens in the building of megaliths such as Stonehenge. Furthermore, Neolithic people were conducting performance evaluations when they evaluated their agricultural and animal domestication techniques. This demonstrates that the earliest Homo sapiens were most certainly evaluating their products and their processes with relatively systematic procedures relying on cultural and oral traditions for passing on their methods and techniques. Furthermore, trade among different cultures is evidenced by the dispersion of manmade items as well as building and agricultural techniques throughout geographic regions and across cultures. This fact permits the assumption that at times, objects were exchanged without the use of a common language; therefore, forms of communication were limited to techniques like gesture and demonstration. Thus, it is logical to assume that during these exchanges, the recipient, to some degree, determined the merit and quality of the tool or weapon without knowledge of the maker's specific intentions or goals, whether the goal being a lighter arrowhead, a more durable ax handle, a larger ear of corn, or healthier livestock. Consequently, it is at least possible that, on occasion, our Eolithic relatives used a very rudimentary informal version of GFE.

Ancient History

One of the earliest individuals worthy of mention for unknowingly using goal-free techniques is the great Greek physician Hippocrates of Cos (c. 460-370 B.C.E.). A contemporary of early Greek thinkers Pericles, Euripides, Aeschylus, Sophocles, and Aristophanes, Hippocrates, and what would later be called Hippocratic physicians, divorced themselves from superstition and the theurgical philosophy of disease[2] in favor of systematically observing the processes of life and working with a set of ethical principles declaring an obligation to the patient (Nuland, 1988). Hippocrates was not focused on his patients' personal or individual treatment goals; on the contrary, he was interested in carefully examining, describing, categorizing, diagnosing, and treating the person's dysfunction and symptomology (cf., Garrison, 1960; Martí-Ibáñez, 1961; Nuland, 1988). Hence, Hippocrates was concerned with what the patient needed not what the patient wanted.

To illustrate how Hippocratic medicine was primarily a goal-free endeavor, consider a patient who visits Hippocrates requesting an incantation to exorcise the demons inflicting boils and pain. Just as modern physicians, Hippocrates is minimally interested in the patient's actual goal (i.e., demon exorcism); instead, he is interested in systematically examining the total physical, mental, and environmental functioning to identify and treat the underlying causal mechanisms that limit the patient's functioning. In fact, Hippocrates used what might be called a precursor to the scientific method as he collected data on several patients with similar symptoms and studied particular treatments

---

[2] The theurgical philosophy of disease says that "illness is caused by unknowable supernatural forces, and so the cure had also to come from those same forces" (Nuland, 1988, p. 7).

and responses to these treatments, the culmination of which was to determine whether the treatment had merit in certain situations. For example, in observing the treatment of patients with bacterial infections, he assessed their symptoms, evaluated his results, and then concluded that there was merit in warming patients; subsequently this treatment for infection has been shown to stimulate the immune system by mimicking the affects of fever (Chen et al., 2006). Hippocrates accomplished evidence-based treatment all without concern for the particular and often irrelevant treatment goals and objectives of his patients. In shifting from the history of evaluation within medicine, the next example of evaluation's ancient history comes from art.

The judgment of art (i.e., art criticism) is an evaluative endeavor (see the connoisseurship model of evaluation[3]) and depending on the circumstances, can be goal-free. According to Scriven (1974b) "in the field of aesthetics it has been widely but not universally accepted that it is fallacious for a critic to consider the intentions of the artist in assessing the work of art" (p. 40). For better or worse, in many cases the art critic does not know or have access to the explicit intentions of the painter, composer, performer, writer, poet, sculptor, chef, architect, designer, etc. Instead, the critic judges the evaluand based on some determination of what the critic believes to be relevant criteria for judging the particular piece good or bad. However, art criticism is not necessarily goal-free. There are numerous painters, composers, writers, etc. who have discussed their works and their

---

[3] Elliot Eisner (1979a, 1979b, 1985, 1990, 1991) promoted the connoisseurship model of evaluation which relies on the judgment of experts who are valued for their presumed experience and knowledge (i.e., expertise) and for their shared value system. Scriven (1991) acknowledges the model's potential, yet warns that it is often subject to the fallacy of technicism and the fallacy irrelevant expertise; therefore, according to Scriven, connoisseurship is rarely the appropriate approach for a program evaluation. However, it should be noted that Eisner does not endorse connoisseurship as the sole approach to an evaluation (Fitzpatrick et al., 2004).

intentions in creating them. Modern orchestral composers often write their own program notes that describe the pieces and their intentions (e.g., melancholy, love, war, nature, etc.), for example. Scriven says that "there's nothing in there that says the background and context of the artwork cannot contribute," (p. 40); however, to be considered a goal-free art critic, the critic must avoid learning the specific goals and objectives of the artist.

The ancient Greeks were interested in art and art criticism, and no one better exemplifies the ancient Greek art critic better than Xenocrates of Sikyon (fl. c. 280 B.C.E.). According to Italian art historian Lionello Venturi (1936/1964), "Xenocrates wrote a treatise dedicated to painters and sculptors in order to give advice and principles.… Xenocrates tried to fix a relationship between his own artistic principles, as categories of artistic judgment, and some concrete artistic personalities" (p. 37). For example, Xenocrates created standards for judging sculptures of the human form based on four criteria: (1) the natural balance of a statue, (2) the variety of bodies, (3) the level of difficulty exhibited in the representing human hair in marble or bronze, and (4) the delicacy in the execution of details (Venturi, 1936/1964). For judging paintings Xenocrates established six criteria: (1) proportion, (2) color and the harmonization of color, (3) tone as unity, (4) perspective, (5) grace, and (6) variety. Xenocrates evaluates; he determines the criteria by which to judge the artwork; he systematically observes the work; and he compares the work according to his standards of goodness; and he concludes whether the work is of quality or not.

After Hippocrates, Xenocrates is one of the earliest who can be called a goal-free evaluator as he developed his criteria of merit, judged the sculptures, sculptors, and the sculptors' portfolios in addition to comparing the artwork and artists without particular

reference to the sculptors' intentions. First, Xenocrates made evaluative conclusions regarding the merit, worth, and/or significance of each evaluated piece of work and each sculptor. For instance, Xenocrates concluded that the sculptures of Polycleitos are superior to those created by previous sculptors because "his statues rest their weight on one leg"; however, the figures are "too monotonous"; and Myron surpasses Polycleitos in variety yet is unable to "represent the hair of the head" (Venturi, 1936/1964, p. 40). Xenocrates does not care if Polycleitos was attempting to create the largest statue or depict the emotion of despair on the statue's face; he does not care whether Polycleitos intended to realistically represent human hair. Rather, Xenocrates judges the work and sculptor based on the criteria which he believed to be merit-defining and then derived evaluative conclusions based on what he observed. Again, all of this was conducted without particular concern for or attention to Polycleitos' artistic goals or objectives.

Of course, modern art critics and evaluators would be quick to recognize that Xenocrates criteria are inherently biased toward the value, or disproportionate weighting, of life-like imitation in sculpture which was the dominant aesthetic zeitgeist as opposed to the stylized representations of the human form which would become prevalent in Western art centuries later. Furthermore, there are significant issues with validity, generalizability, reliability, and credibility in art criticism and connoisseurship in general; nevertheless, art criticism provides an example of a legitimate attempt at systematically determining the merit, worth, and significance of a piece of art, artist, or artist's body of work without regard to the creator's motives, instead focusing on the artist's actual outcomes or product. The history of evaluation continues into the European Renaissance with another example from art criticism.

The European Renaissance

GFE via art criticism was rejuvenated during the European Renaissance (c. 1350-1550). For example, art critic Giorgio Vasari analyzed and compared the works of Giotto, Leonardo, Raphael, and Michelangelo in *The Lives of the Artists* (1550-68). Evidence of merit determination is exhibited in one of Vasari's evaluative conclusions where he compares Raphael to Michelangelo. Vasari writes, "These things, I say, Raphael considered, and not being able to approach Michelangelo in that part of the nude, he resolved in these other parts to emulate and perhaps surpass him" (Vasari, cited in Venturi, 1934/1964, p. 103). Another art critic, Gian Paolo Lomazzo, published *Treatise on the Art of Painting* in 1584; in it, he judged paintings based on: (1) theory, (2) practice, and (3) iconography (Venturi, 1936/1964). Lomazzo is just one of the many examples of art critics who judged art based on a list of criteria while ignoring the explicit intentions of the artists; thus a goal-free evaluator. The Renaissance not only fostered changes in art and art criticism, but also a resurgence in philosophy, science, and the philosophy of science. The next section of GFE history include the years where evaluation begins to emerge as a discipline distinct from research and demonstrates that, historically, evaluation has been inextricably tied to goal attainment (Fitzpatrick et al., 2004; Madaus & Stufflebeam, 1989; Patton, 1997; Scriven, 1991).

Tylerian Evaluation (Goal-Based Evaluation)

Fathered by Ralph Tyler (1902-1994), goal-based or objective-based approaches are evaluations which are particularly concerned with the attainment of pre-selected goals and objectives. As previously stated, GBE has been the dominant approach on the

evaluation scene since its inception (Alkin, 2004a; Fitzpatrick et al., 2004).[4] Goal-based approaches are important to mention because it was the questioning of these approaches that led, in part, to the theorizing of GFE.

Madaus and Stufflebeam (1989) claim that Tyler's evaluation process was conducted in seven phases: (1) goals and objectives are determined, (2) the objectives are classified by type, (3) the objectives are refined and put into behavioral terms, (4) situations are identified when these behaviors may be observed, (5) different measures for gathering evidence are tested and chosen, (6) instruments are pilot-tested and performance data is collected, and (7) performance data and behavioral objectives are compared (Tyler, 1942, pp. 498-500). It should be noted that Tyler (1974) was not naïve with regard to goal measurement as he emphasized pilot-testing data collection instruments to refine their capacity for measuring the relevant effects as well as to critique learning objectives and their appropriateness with regard to the anticipated effects. As stated by Fitzpatrick et al. (2004), "Tyler stressed the importance of screening broad goals before accepting them as the basis for evaluating an activity" (p. 73).

Despite the significant methodological improvements in educational testing and measurement, it was not until the late 1960s before program evaluation began possessing its own theoretical framework, distinct from research methods, with evaluation-specific publications by such authors as and Guba (1969), Scriven (1967), Stake (1967), Stufflebeam (1968), and Suchman (1967). Prior to the 1960s, Tylerian evaluators drew from theories and practices in cognate disciplines for improving their methodologies in

---

[4] Objectives-based evaluation approaches are also sometimes known as objectives-oriented evaluation, objectives-referenced evaluation, and criterion-referenced tests.

quasi-experimental design, survey research, taxonomy, and ethnography, among others (Fitzpatrick et al., 2004). GFE existed in practice during the early 20th century, yet although not GFE in name. The next section introduces the Consumers Union, whose evaluations epitomize GFE in practice.

The Consumers Union

One of the first modern examples of GFE comes from product evaluation and the development of consumers' groups. In the 1920s, Frederick Schlink, a former staff member for the National Bureau of Standards, organized a consumer club in White Plains, New York; this club eventually became the Consumers Union, publisher of "Consumer Reports" magazine (Consumers Union of U.S., 2000). Setting the Consumers Union's evaluations apart from the evaluation methodologies of its contemporaries is the fact that the Union's evaluations exemplify GFE (Scriven, 1991). Further adding to the significance of Consumer Reports, the Consumers Union demonstrates a long history of the use of quantitative data collection methods in conjunction with a goal-free approach.

From its founding, the Consumers Union's inquiries and product evaluations have been based on the ideologies and principles of positivism. For example, a May 1936 report from the Consumers Union included an article grading various brands of milk (Consumers Union, 2000). The Consumers Union created a grading rubric (i.e., best buy, acceptable, and not acceptable) based on the chemical analyses of milk samples. Using experimental design methods in a highly controlled laboratory environment, the researchers examined the levels of vitamins, minerals, and fat as well as the levels of chemical and biological contaminants in each sample. Next, the Union factored in the

retail cost, compared the various brands, and judged the milk and milk companies according to the results of these quantitative analyses. By the early 1950s, the Consumers Union included survey research methods to its repertoire as it collected product repair frequency data by having its readers respond to quantitative mail surveys (Consumers Union, 2000).

The Consumers Union's evaluation model represents one of the best examples of GFE as the Union conducts its product evaluations without regard to the intentions or goals of product designers and manufacturers. To illustrate, consider a new model sport utility vehicle (SUV); regardless of the manufacturer's motives in designing the vehicle, the Consumers Union judges the SUV by first describing its various specifications (e.g., dimensions, seating height, weight, carrying capacity, cabin space, engine type, four-wheel drive, etc.). Then the vehicle is classified by type (e.g., full-size, mid-size, compact, mini, crossover) allowing for the development of quality standards by which the vehicle and its performance is compared with other vehicles of its type. Many relevant criteria of merit are known or identified for judging the quality of SUVs of this type and for ultimately deeming it a good or bad purchase for a consumer of a particular type. Often with evaluating a new product, criteria are developed iteratively by observing what the product does. Common criteria might include the vehicle's predictability and reliability, braking system, handling system, suspension system, cooling system, transmission, exhaust system, safety features, tire quality, fuel efficiency, carbon dioxide emissions ratings, top end speed and acceleration, towing capacity, torque, comfort, color options, owner satisfaction reports, availability of additional options and features, retail cost, maintenance costs, and re-sale value, among a plethora of other possibilities. The

automobile's quality, or lack thereof, is determined by measuring the vehicle's performance outcomes on the criteria deemed relevant. These results are placed into a grading rubric and weighting system which calculates a grade, rank, or score[5] that can then be compared to the designation assigned to other SUVs of its type. After synthesizing the data and completing the appropriate comparisons, the SUV is determined to be excellent, very good, good, fair, or poor. This entire evaluation process is conducted without specific attention paid to or knowledge of the designers' manufacturers' goals in creating the particular SUV.

It should be noted that the Consumers Union does not *intentionally* avoid the goals of the car manufacturer; however, the Union also does not seek out goal-related information and thus are operating in a goal-free manner. The criteria of merit are determined, standards are set, evidence is systematically collected and analyzed, comparisons are made, and the process culminates with an evaluative conclusion; all of this is accomplished without particular attention to the product manufacturers' goals. The Consumers Union has over seventy years experience with conducting goal-free product evaluation; yet, the discovery of GFE remained unrecognized by evaluation scholars.

Contemporary Professional Evaluation

Contemporary professional evaluation generally permits the inclusion of varying philosophies and ideologies with regard to gathering data (Scriven, 1991; Young, 1990). During formal evaluation's infancy, evaluators adhered to the positivist doctrine of value-free social science which asserts "that no evaluative judgments can be made with

[5] A grade, rank, or score is often assigned to components or systems of the vehicle; additionally, the Consumers Union gives a grade, rank, or score to the vehicle in its entirety as well.

scientific objectivity" (Scriven, 1983, p. 231). This doctrine of value-free science represents the influence of positivist philosophy on evaluation (Patton, 2002). However, on the contrary, Scriven (1991) argues that "the widespread acceptance of the doctrine of value-free science is a groupthink phenomenon, exploiting evaluation anxiety at the expense of rationality, since any scientist can see—once the point is made—the essential role that evaluation plays in every science" (p. 375). For example, the same positivist scientist claiming that value judgments have no place in science makes value judgments as s/he proclaims to know the difference between a quality research study and poor studies (Coryn, 2007). In discussing postpositivism and postpositivists, Patton (2002a) writes that Donald Campbell

> recognizes that discretionary judgment is unavoidable in science, that proving causality with certainty in explaining social phenomena is problematic, that knowledge is inherently embedded in historically specific paradigms and is therefore relative rather than absolute, and that all methods are imperfect, so multiple methods, both quantitative and qualitative, are needed to generate and test theory, improve understanding over time of how the world operates, and support informed policy making and social program decision making. (p. 92)

Sirotnik and Oakes (1990) state that if the proposition that inquiry is never value free is accepted, then the accumulated body of work by critical theorists direct one toward a constructivist epistemological synthesis called critical inquiry that is evaluative by its very nature. Furthermore, the positivist traditions regarding knowledge and postmodern critical social construction of knowledge are nearly always interlinked therefore evaluators should be prepared to deal with them both (Young, 1990). Thus, it

can be concluded that current and continual attacks on positivism "are not beating a dead horse, they are beating an eohippus" (Scriven, 1991, p. 270).[6]

Even though GFE is epistemologically and ontologically neutral, it is often incorrectly labeled critical theory or a qualitative method of inquiry. Again, product evaluation represents the strongest evidence that GFE can be used with positivist ideologies (see The Consumers Union example above). On the other hand, GFE can also be employed with critical theory allowing for multiple realities as is the case with GFEs that implement constructivist methodology and qualitative data collection methods.

GFE is frequently mislabeled an alternative evaluation approach assuming that it relies on critical theory and qualitative methods yet, as demonstrated above, the Consumers Union evaluations provide a commonplace counterexample as the goal-free evaluator can and does subscribe to positivist ideology while using quantitative methods. Furthermore, regardless of how one chooses to categorize GFE, today positivist, postpositivist, and critical theory ideologies are accepted by mainstream professional evaluation; therefore the professional evaluator should be equipped with knowledge of the various ideologies and their associated methodological assumptions. While this section discusses contemporary evaluation practice and GFE's relationship within it, the next portion of this chapter outlines the logic that underlies GFE.

---

[6] An eohippus is an evolutionary ancestor of the horse extinct for over 50 million years.

The Logic of GFE

The logic of GFE includes: (1) the definitions of GFE and its major concepts, (2) the nature of its relationships to other subjects and disciplines, and (3) the rules of inferences that govern it.

Definition of GFE

The next section of this dissertation defines GFE. The reason to elaborate on the definition is that it establishes boundaries on the concept. A quality definition encourages common understanding by helping to clearly state a precise meaning of a word or concept by providing the extent and the limits of the word. Deviation from the common meanings of terms frequently causes confusion and discourages persistence in understanding concepts (Scriven, 2005a).

The definition of GFE espoused in this dissertation refers to the process of systematically determining the merit, worth, and/or significance of an evaluand with the evaluator partially or fully screened from the stated (or implied) purposes, aims, and intentions (i.e., specific goals and objectives) of those who design, produce, and/or implement the evaluand. This definition is in agreement with the one from the PES which defines a GFE as an "evaluation of outcomes in which the evaluator functions without knowledge of the purposes or goals" (Joint Committee, 1994, p. 206). The presupposition in a GFE is that the evaluator intentionally avoids learning the preordinate goals and objectives of the evaluation client, evaluation users, and certain stakeholders; instead the evaluator observes and measures the actual outcomes and presents the evaluation client with an evaluation report based on all actual effects, positive, negative, and neutral. As

Scriven (1991) writes, the evaluator judges the evaluand based on definitional and functional premises about what the evaluand is and does, and on meeting the consumers' relevant needs.

Before continuing with the definition of GFE in its entirety, an investigation is warranted into the varying common dictionary meanings and definitions for the words *goal*, *free* (to a lesser extent), and *evaluate/evaluation*. In addition, the definitions from professional evaluation literature are also examined. The reason behind such depth in examining these definitions is to present GFE's philosophical and theoretical basis.[7]

Goal

There tends to be a general agreement on the definition of the word *goal* (or *goals*) in common usage as well as in the evaluation literature. Both *Chambers Concise Dictionary* and *Cambridge Advanced Learner's Dictionary* define a goal as an "aim" or "purpose." The *American Heritage Dictionary of the English Language* uses "purpose" in its definition for goal: "the purpose toward which an endeavor is directed; an objective… an intention" (Pickett, 2000, p. 752). The third definition for goal in the *American Heritage Dictionary* is "[a] place to which something moves" (Pickett, 2000), which does not imply intention; however, most of the common English dictionaries, and including the *Oxford English Dictionary* (OED), contained the word *aim* and/or *purpose* in their definitions of goal.

---

[7] *Note:* Only relevant definitions of the words from the dictionaries are given in this dissertation. For example, definitions of *goal* as it pertains to a finish line in a race or as a score in soccer are excluded. Similarly, the definitions of *evaluate* that pertain to mathematics, and definitions of *free* that relate to the absence of cost are also omitted.

Consistent with common relevant uses of *goal*, the evaluation literature also gives *aim*, *purpose*, and *intention* as synonyms. Scriven (1974b) refers to goals as "a subset of anticipated effects; they are the ones of special importance, or the ones distinctive of this project" (p. 37). Most evaluation scholars are apt to differentiate between goals and objectives according to their specificity and measurability. For example, The PES, Weiss and Jacobs (1988), and the *Encyclopedia of Evaluation* all define the word *goal* and distinguish it from an objective. The PES defines a goal as "an end that one strives to achieve" (Joint Committee, 1994, p. 206) and adds that an objective is also an aim, but is more specific than a goal. Weiss and Jacobs write that goals are "broad statements of a program's purposes or expected outcomes, usually not specific enough to be measured and often concerning long-term rather than short term expectations" (p. 528), while objectives are "statements indicating the planned goals or outcomes of a program or intervention in specific and concrete terms" (p. 533). Tucker (2005) defines a goal as a "general statement of an intended outcome" and continues by stating that a goal is "usually operationalized into a measurable objective" (p. 171).

The differentiation between a program's goals and its objectives is an important distinction for practical reasons in quality control and program management, yet the exact distinction is somewhat irrelevant in the attempt to observe all of the relevant effects on stakeholders and impactees. In other words, it is not the intention; it is the outcome. The thought does not count in this case; rather the results are what matter given fair consideration of the resources limitations. Therefore, despite the specificity or the measurability, an objective is an aim just as a goal is an aim. The goal-free evaluator

avoids all stated aims; hence, as previously mentioned, a GFE is as much an objective-free evaluation as it is a goal-free one (Scriven, 1974b).

Scriven (personal communication, February 22, 2007) claims in all actuality the evaluator typically has a basic understanding of the general purpose of the evaluand (i.e., goal) simply by knowing the nature of the evaluand; consequently, the overarching goal or goals of the program are obvious. Therefore it is a misnomer to call an inquiry of this type "goal-free" rather it should be referred to as an "objective-free evaluation." Hence, throughout this dissertation, objective-free evaluation is implied when referring to GFE. In conclusion, the common English dictionaries as well as the professional evaluation literature mostly concur that synonyms for goal include aim and purpose and the generality with which one refers to the aims is the distinguishing factor between a goal and an objective and for this reason, an objective is subsumed under a goal.

Free

In common English, there is a general consensus regarding the definition of *free*, something like, without the restraint of or to rid. The *American Heritage Dictionary* defines *free* as "not controlled by obligation or the will of another" and "not affected or restricted by a given condition or circumstance" (Pickett, 2000, p. 700). There is no specific definition for *free* in the evaluation literature; rather, it is usually used in conjunction with other terms to make combinations such as goal-free, cost-free, bias-free, for example. Thus, the "free" in GFE refers to the theoretical independence of the evaluator from the evaluand's stated goals and objectives which were established by its designers, manufactures, implementers, the evaluation client, and/or the evaluation user.

Evaluate/Evaluation

The definition of *evaluation* is important to the study of all evaluation and, therefore, the inquiry into the various definitions requires more depth. Although there is relative agreement on the definition of "evaluation" in everyday language, there is less of a consensus among evaluation scholars. Different approaches to evaluation use different definitions of evaluation (Patton, 1997). Below is an examination of a few vernacular definitions as well as definitions from professional evaluation literature. This section concludes with a synthesized definition.

Definitions from Common Usage

Although there is a layman's understanding of the word *evaluation*, Scriven identified approximately 60 context-dependent synonyms for evaluation including such examples as *appraise, analyze, assess, critique, examine, grade, inspect, judge, rate, rank, score, test,* and so on (Patton, 2002). *Chambers Concise Dictionary* defines *evaluation* as: "to form an idea or judgment about the worth of something" (Editors of Chambers, 2004, p. 398), while the *Cambridge Advanced Learner's Dictionary* defines *evaluate* as "to judge or calculate the quality, importance, amount or value of something" (Walter, 2005, p. 425). The *American Heritage Dictionary* offers the definition: "to ascertain or fix the value or worth of" and "to examine and judge carefully; appraise" (Pickett, 2000, p. 615); *Webster's Third New International Dictionary of the English Language* defines *evaluation* as: "to estimate or ascertain the monetary worth of: value" with a second definition of, "to examine and judge concerning the worth, quality, significance, amount, degree, or condition of: appraise, rate" (Grove, 1986, p. 786). Often

considered the gold standard for English dictionaries, the OED only provides definitions of *evaluate* that refer to mathematical applications. Despite the lag in keeping current with the common usages of the term, portions of the OED's definition include "to work out the value of" and "to 'reckon up' and ascertain the amount of" (Simpson & Weiner, 1989, p. 447). The OED offers a separate definition for *evaluation* and defines it as "the action of appraising or valuing (goods, etc.); [and is] a calculation or statement of value" (Simpson & Weiner, 1989, p. 447). The last dictionary definition examined, *Random House*, defines *evaluate* as "to determine or set the value or amount of; appraise," with a second definition of "to judge or determine the significance, worth, or quality of; assess," (Random House, 2001, p. 670). In examining these standard dictionaries, frequently used synonyms for evaluate become apparent. A synthesis of these synonyms and definitions is to judge, ascertain, and/or appraise the quality, worth, amount, and/or value of something.

Since the early 1990s, growth in Internet access and usage helped create a new type of dictionary, a wiki-dictionary.[8] Possibly the Internet's most frequented wiki site is Wikipedia.com, whose submissions are, for the most part, created and edited by anyone who wants to and has access to the Internet. There is also an evaluation-specific wiki site, Evaluationwiki.org, which raises submission and editing standards by requiring all who submit or edit an entry to identify themselves and their credentials. Even though Internet-based sources are typically considered less credible, the wiki sites are usually updated

---

[8] There are numerous other Internet-based sources for definitions of *evaluation* that are not examined in this dissertation as they do not significantly deviate from definitions previously discussed. Online sources include pre-existing dictionaries adapted for the Internet such as The Oxford English Dictionary Online and Merriam-Webster Online, and dictionaries that exist solely online like Princeton University's WordNet, YourDictionary.com, and Freedict.com.

much more frequently and thus have the potential for being more current and relevant for the user. One other major benefit of wiki dictionaries is that the wiki dictionary's collective nature means that a term also represents the *zeitgeist* in which it was defined. In summary, wiki sites facilitate an open and public discourse regarding a word's definition and the result of this online collaboration leaves readers with an accumulative definition that is characteristic of the time from which it was extracted.

> According to Wikipedia (October 24, 2008 en.wikipedia.org/wiki/Evaluation):[9]
>
> Evaluation is the systematic determination of merit, worth, and significance of something or someone using criteria against a set of standards. Evaluation often is used to characterize and appraise subjects of interest in a wide range of human enterprises, including the arts, criminal justice, foundations and non-profit organizations, government, health care, and other human services.

Evaluationwiki.org begins its page "what is evaluation" by citing the American Evaluation Association's (AEA) definition of *evaluation*. According to Evaluationwiki.com (October 24, 2008 www.evaluationwiki.org/index.php/Evaluation _Definition:_What_is_Evaluation%3), "evaluation involves assessing the strengths and weaknesses of programs, policies, personnel, products, and organizations to improve their effectiveness." A subsequent definition for evaluation from Evaluationwiki is "the systematic collection and analysis of data needed to make decisions, a process in which most well-run programs engage from the outset." The two wiki dictionary definitions for evaluation are similar to the definitions from the common English dictionaries and, as will be presented in the next few paragraphs, agree with some definitions from professional evaluation yet do not agree with others.

---

[9] As of September 2008, both Wikipedia.org and Evaluationwiki.org offer a definition of *evaluation* but neither defines *evaluate*.

Definitions from Professional Evaluation Literature

There is no uniformly agreed upon definition for *evaluation* found in the professional evaluation literature; in spite of this, the next few pages are dedicated to presenting selected definitions of evaluate by separating them into two types. The first group consists of five definitions that closely resemble the aforementioned common definition as they define *evaluation* as the determination of value, merit, or worth, while the latter definitions deviate from these common definitions and consequently are critiqued.

Definitions Group One: Common-Professional Definitions. A sound definition for *evaluation* should focus on what evaluators do that distinguishes them as evaluators (Scriven, 2005a). This distinguishing feature, according to Scriven (1991), is that evaluators are concerned with determining value; therefore, he defines evaluation as the determination of merit, worth, and value of something, or the product of that process.[10] The *Encyclopedia of Evaluation* defines evaluation as "the systematic assessment of the worth or merit of some object" (Bickman, 2005, p. 141). Fitzpatrick et al. (2004) agree with the determination of worth or merit of an evaluation object but also add, "the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit) in relation to those criteria" (p. 5). Stufflebeam (2001) supports the definition of *evaluation* as a study designed and conducted to assist some audience in assessing an object's merit and worth. Fetterman and Wandersman (2005) define *evaluation* as "the formulation of judgments about the merit, worth, or

[10] In subsequent versions, Scriven replaced "value" with "significance" (Scriven, 2005a).

significance ... on the basis of systematic inquiry" (p. 204),[11] while the definition from The PES is "the systematic investigation of the worth or merit of an object" (Joint Committee, 1994, p. 205). The last example definition is that of Fournier (2005), who states that evaluation is a process of applied inquiry in which the evaluator collects and synthesizes evidence that will be used by the evaluator drawing conclusions about the state of affairs, value, merit, worth, significance, or quality. In conclusion, a simplified and synthesized version of the preceding definitions from the professional evaluation literature eliminates several of the synonyms to conclude that evaluation is the process of systematically determining merit, worth, and/or significance of something.

There are some professional evaluators who provide definitions that are congruent with the common definition for *evaluation* yet have non-substantial differences. For example, Davidson's (2005) definition does not include the exact words *merit* or *worth*; many of these authors, Davidson included, do offer alternative phrasings or synonyms such as *quality* and *value*. Regardless, the definitions derived from the common dictionaries and the five definitions from the aforementioned scholarly publications agree on evaluation's purpose of determining merit. To distinguish between the common use of *evaluation*, which includes the rudimentary evaluation that we do daily, professional evaluators regularly add the words *systematic* and/or *objective* to instruct the reader how the determination of merit should occur. Of course, there are alternative definitions that do not agree with the above definitions and several of these are discussed next.

---

[11] In the definitions of evaluation, words like *program* are sometimes removed in order to create a general definition of evaluation, one that is not specifically related to fields of evaluation (e.g., program, product, personnel, policy, proposal, performance, portfolio, and process). Instead, a sound definition should encompass all of the fields and domains of evaluation including intradisciplinary evaluation, metaevaluation, the common more rudimentary and idiosyncratic evaluation that humans do daily.

Definitions Group Two: Alternative Professional Definitions. The following examples hail from the second grouping of definitions: the alternative definitions. In this section, some of the limitations and fallacies that accompany these alternative definitions are explained. The definitions from most of these alternatives tend to be too narrow in scope or lose sight of evaluation's distinguishing characteristics.

Below are examples of the numerous evaluation authors who define *evaluation* in too narrow of terms. Rossi (Alkin, 2004b) states that evaluation "consists essentially in the application of the repertory of social research methods to provide credible information that can aid in the formation of public policy, in the design of programs, and in the assessment of the effectiveness and efficiency of social policies and social programs" (p. 127). Similarly, Rossi, Freeman, and Lipsey (1999) write that "program evaluation is the use of social research procedures to systematically investigate the effectiveness of social intervention programs" (p. 4). These two definitions are too narrow as they relegate evaluation to the social sciences and research methods despite the fact that evaluators use many tools that come from outside of applied social science research like logic, axiology, and ethics, to name a few (Scriven, 2005a). Evaluators, like social scientists, collect factual datum; however, in accordance with the definition subscribed to in this dissertation, the evaluator has not completed the job until an evaluative conclusion or conclusions (i.e., a judgment of merit, worth, and/or significance) is achieved. Furthermore, Rossi (2004) defines *evaluation* according to its uses (e.g., the assessment of effectiveness and efficiency), when the actual purpose of evaluation is to make a judgment about whether something is good or bad, valuable or

invaluable, and/or significant or insignificant.[12] Lastly, Rossi's and Rossi, Freeman, and Lipsey's definitions ignore a major field or practice area of evaluation, which is often an essential subevaluation in conducting a program evaluation: product evaluation.

Some definitions are simply too wordy. Grinnell and Unrau's (2008) definition of *evaluation* is an example of such; it focuses solely on the implementation of research methods, and bases its definition on evaluation's uses rather than what it does. Grinnell and Unrau write that an evaluation is

> a form of appraisal using valid and reliable research methods; there are numerous types of evaluations geared to produce data that in turn produce information that helps in the decision-making process; data from evaluations are used to develop quality programs and services. (p. 546)

Like Grinnell and Unrau's (2008) and Rossi's (2004) definitions from above, Patton's (1997) definition also emphasizes the ways an evaluation can be used. According to Patton, evaluation is "the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future programming" (p. 23). However, as previously stated, the utility of an evaluation is a required consideration for all professional evaluators yet the degree to which the evaluator is responsible and accountable for the use of the evaluation by the evaluation's stakeholders is debatable. A definition of *evaluation* that focuses on how the evaluation will be used has potential to result in bias with regard to what is investigated during the evaluation and what is actually recommended by the evaluator, when the real focus

---

[12] In this example, effectiveness and efficiency may be criteria for determining the merit of the evaluand.

should be on gathering the appropriate relevant evidence necessary for determining the evaluand's merit.

Royse, Thyer, Padgett, and Logan's (2006) definition is another example of a definition that is too narrow; they characterize program evaluation as "applied research used as part of the managerial process" (p. 11). Despite the aforementioned problem with defining *evaluation* as research, program evaluation is not limited to the managerial process or the management department. To offer a real-world example, the U.S. News and World Report magazine's ranking of undergraduate programs is conducted independent of, and not intended to support, management of any university. Rather, the magazine's rankings are primarily conducted to determine the merit and worth of the programs for the purpose of generating knowledge for the consumer, and secondarily for holding universities accountable for their performance and costs via influencing prospective student and parent applicants.[13]

Schalock and Thornton (1988) define *program evaluation* as a structured comparison, which is another instance of a definition that is too exclusive. Evaluators certainly deal with structured comparison, yet it is not evaluators' defining characteristic. Evaluators also read texts, use statistics, apply management information systems, and likely drink soda; but it would be absurd to define evaluation based on any of these activities. A quality definition of *evaluation* must offer a distinction between evaluation and all other disciplines or applied fields. Furthermore, not all evaluations include comparisons; often the evaluation is concerned solely with the absolute merit of the evaluand. Lastly, the above definition also fails to acknowledge the making of judgments

---

[13] Of course another motive of the magazine is to sell magazines and advertisements.

and determining merit, which are the distinguishing characteristics that separate evaluation from the rest.

According to Weiss and Jacobs (1988), program evaluation is "a planned review of a program [that] attempts to answer questions of concern to the group that initiated or requested the evaluation" (p. 534). The major problem with this definition is that it states that the purpose of evaluation is to answer stakeholder questions. In accepting the definition of evaluation as the determination of merit, the stakeholders' questions may be, and sometimes are, irrelevant. Focusing solely on the questions posed by the evaluation client leads the evaluator toward an emphasis on the program targets; discourages concern for all actual outcomes and effects; creates circumstances rife with opportunities for hiding, disguising, embellishing, etc. actions, outcomes, and effects; as well as distracts the evaluator from the true task at hand, the determination of merit. For example, a school district may want to know whether students and parents are satisfied with the new educational curriculum offered when in fact the relevance of satisfaction is debatable as it can be argued that satisfaction is unrelated to the students learning anything significant with the new curriculum, learning more in comparison to an alternative curriculum, or whether the particular type of program or curriculum is warranted in the first place.

Newcomer, Hatry, and Wholey (2004) also omit the judgment and determination of merit in their definition which states that "program evaluation is the systematic assessment of program results and, to the extent feasible, systematic assessment of the extent to which the program caused those result" (p. xxxiii). This may be true; however, not only do evaluators gather data on outcomes and determine causality, but they

subsequently take this information and use it to make a judgment about the program determining whether the program was good or bad, worthwhile or worthless, and significant or insignificant. Omitting the role of judgment in evaluation is the same problem found in the theory-driven evaluation and the empowerment evaluation approaches. The theory-driven approach defines *evaluation* by its aim at solving and articulating how and why programs work or do not work, while empowerment evaluation defines *evaluation* as the use of evaluation to empower self-determination (Coryn, 2005).

In conclusion, for many of the reasons outlined above, the following extended definition for *evaluation* is subscribed to throughout the remainder of this dissertation:

> Evaluation is the process of systematically and objectively determining, or judging, the merit (i.e., quality), worth (i.e., value especially related to monetary and non-monetary cost), and significance (i.e., importance) of an evaluand (i.e., the thing being evaluated; the object of the evaluation); or the product of that process.

In short, evaluation is the systematic determination of merit. After examining definitions for *goal*, *free*, and *evaluation*, a synthesis definition of GFE is possible (see Table 1).

Table 1

Synthesis Definitions of Goal, Free, and Evaluation

Goal	Free	Evaluation
Aim; purpose	Independent of; not affected by the obligation to and the obstruction of	The determination of the merit, worth, and/or significance of something

Below are two phrasings for the definition of GFE.

Phrasing 1: GFE is the process of systematically and objectively determining the merit, worth, and/or significance of an evaluand conducted partially or fully independent of the stated (or implied) specific purposes, aims, and intentions of the evaluand or its upstream stakeholders.

Phrasing 2: GFE is the determination of the merit, worth, and/or significance of an evaluand independent of or not affected by the obligation to and the obstruction of the upstream stakeholders' stated or intended aims or purposes.

Nature of GFE's Relationships

The nature of GFE's relationships refers to its connections with other subjects and disciplines. GFE has a relationship with GBE, it has a philosophical relationship, and it relates to several fields of evaluation.

Goal-Based Evaluation Principles

This chapter began by stating that a good definition places limits on a concept; hence, an articulation of GFE's parameters is offered through an explanation of what GFE is not. The characteristic that defines or typifies GFE is its relationship to GBE and therefore a discussion on the nature of GFE's relationships begins by briefly examining GBE.

All evaluations can be placed in one of three categories according to their position regarding goal orientation: (1) entirely goal-based, (2) entirely goal-free, or (3) a combination goal-based/goal-free. A GFE is not an entirely GBE. A GBE is "any type of evaluation based on and knowledge of—and referenced to—the goals and objectives of

the program, person, or product" (Scriven, 1991, p. 178). According to Hezel (1995), GBE refers to

> cases where programmatic goals have been clearly established during the program's formation, the goals and subsequent concrete and precise objectives become the criteria for measuring the "success" of the program. The goals-based approach is particularly useful for evaluating those aspects of the program that are circumscribed by goals established for the program. In this case, the goals established for the program articulate in a general way the outcomes expected from the program. In turn, the expected outcomes form the basis for the measurement of actual outcomes. (p. 47)

GBE is a categorized by Fitzpatrick et al. (2004) as an objectives-oriented evaluation approach, which is described in the following statement:

> The objectives-oriented evaluation approach has caused program directors to reflect about their intentions and to clarify formerly ambiguous generalities about intended outcomes. Discussions of appropriate objectives with the community being served have given objectives-oriented evaluation the appeal of face validity—the program is, after all, merely being held accountable for what its designers said it was going to accomplish, and that is obviously legitimate. The objectives-oriented evaluation approach is one that directly addresses Standard U4, Values Identification, in *The Program Evaluation Standards* (Joint Committee, 1994). Its emphasis on clearly defining outcomes as the basis for judging the program helps evaluators and others to see the value basis for judging the program. (p. 82)

The philosophical rationale underlying GBE is that program stakeholders have selected the goals and objectives and toward these goals, the upstream stakeholders direct their efforts. Goals and objectives can be created with careful reflection and thoughtfully adapted over time, and, as Vedung (1997) points out, the goals represent the desires of the key and most influential parties involved with the program and therefore "are not haphazard wishes or incidental desires" (p. 61). As previously mentioned, when discussing teaching to the test (see p. 22), one of "the strength of goals is that they direct programs by focusing actions on specific outcomes" (Friedman et al., 2006, p. 202); furthermore, in public programs, the goals have often been adopted by democratically

elected politicians thus giving them a special status as the politicians have procedures established for decision making and theoretically are representing the interests of their constituents.

In a GBE, the evaluator focuses efforts on examining the preordinate goals and objectives and then measuring whether they were achieved. A goal-based evaluator—and to a lesser degree a goal achievement evaluator—may validate the goals as accurate and representative of administrators' intentions (Hezel, 1995). Furthermore, the goal-based evaluator may—but rarely does—conduct a needs assessment to investigate the relevance of the goals and level of difficulty in attaining the goals, a comparison of alternative ways of achieving the same goals, a search for side effects, and an examination of program's processes. However, these processes are not required of the goal-based evaluator and therefore are often foregone, seen as ancillary since the primary emphasis is always the measurement of goal achievement.

GAE is GBE in its most rudimentary form. GAE is a monitoring system with the lone task of determining whether the evaluand met or is meeting its goals and objectives (Scriven, 1991). The defining difference between a GAE and a GBE is that GAE is restricted to solely examining goal achievement whereas GBE is free to examine other areas in addition to goal attainment. The specific principles of GAE are as follows:

1. Identify the evaluand's goals and objectives.
2. Operationalize the goals and objectives.
3. Measure performance on the goals and objectives.
4. Compare the evaluand's performance with the achievement of the goals and objectives.

Thus, the ultimate question for the goal achievement evaluator is whether the evaluand achieved the goals.

As stated earlier, the primary argument in favor of GAE is that a program is designed to do certain things in a certain way; hence, a program should be judged by comparing what it is designed to do with its actual performance outcomes (and to a degree, the outcomes of its consumers (see Scriven, 2005b). Program managers and staff must monitor their efforts yet have to deal with their limited ability to collect relevant data from relevant sources; furthermore, there are the ever present issues of credibility with internal evaluations. So they hire an external goal achievement evaluator who offers an independent analysis as to whether or not, and the degree to which, the goals and objectives are being or were met.

The goal achievement evaluator accepts the stated goals and objectives. At best, they may conduct a brief goal alignment (i.e., an elementary assessment of the relationship between the program's actual activities and intentions with its official stated goals) and adapt the official stated goals only when necessary. Sometimes the goal achievement evaluator meets with one or more key program people and/or evaluation clients to ask them whether these goals are representative of what the program does and is trying to do. This is done simply to ensure that the official stated goals are not outdated or inaccurate, and then the evaluator adjusts the official stated goals and objectives as deemed appropriate.

The bottom line is that GAE is useful for internal management-oriented evaluation, but is rarely justified as the sole source for external evaluation, in part, because the evaluator simply adopts the goals and objectives as stated by the evaluation

client leading to several critical potential weaknesses. Among the negative results is the frequent failure to consider the current relevance of the goals and objectives or to investigate ways of achieving the same goals with fewer resources or different methods. Thus, there is value in frequently questioning and confronting the underlying assumptions of program goals and strategies (Argyris & Schon, 1978). An evaluator using a GAE furthermore neglects to pursue any effect unless it has been stated as a goal or objective and therefore the goal achievement evaluator may be missing positive or negative side effects or side impacts. An additional limitation shared by of both GBE and GAE is that neither is equipped to compensate for contextual or environmental changes which result in the adaptation of the program's goals or resources (Scriven, 1991). Nevertheless, GBE has its strengths and weakness as does GAE as does GFE. However, GAE alone is rarely the appropriate evaluation approach for external evaluators.

The Philosophy of GFE

GFE is ontologically and epistemologically neutral (i.e., it does not subscribe to any one ideology over another) and, therefore, this neutrality enables other evaluation models and approaches to be adopted or adapted for working with GFE. Therefore, GFE can be used with methods derived from positivist, post-positivist as well as constructivist and critical theory ideologies.[14] A short list of example methods/methodologies that theoretically can be used in complement with GFE includes quantitative methods and qualitative methods, *ex ante* designs (e.g., in proposal evaluation) and *ex post facto* designs (e.g., in product evaluation), and survey methods and experiments. Other

[14] Examples of adaptable evaluation models include medical, social science, transdisciplinary, connoisseurship, judicial, adversarial, responsive, naturalistic, utilization focused, and many others.

methods used with GFE include document analysis, statistical methods, ethnography, and, of course, the blinding of the investigator (e.g., the specific avoidance or screening of the designers' or implementers' intentions).

There are two sets conditions that must be met for a model to be adapted for using with a GFE. The conditions for implementing a GFE include: (1) the suitability of adapting a particular evaluation model to use as or with a GFE, and (2) that the GFE is conducted independently of any GBE. First, GFE requires that the chosen evaluation model used with GFE is flexible enough so that it can be adapted (if need be) to work in conjunction with GFE. Second, GFE excludes only the models which dictate that the initiation of the evaluation be in a goal-based manner[15] or are heavily goal-oriented throughout. Even so, according to Scriven (1991), a GFE may be used with a GBE as long as the GBE begins after the GFE, i.e., at the conclusion of the goal-free portions of the evaluation. A GBE can also be employed simultaneously to evaluate the same evaluand as a GFE as long as the goal-free evaluator remains completely independent of the goal-based evaluator. Therefore, GFE is methodologically goal-free yet neutral with regard to many other methodological decisions except goal orientation. To reiterate, one of the following two conditions must be met to effectively incorporate GFE with GBE:

1. The evaluation must begin entirely goal-free and after the data collection (and likely some analyses and possibly reporting) has been completed, the evaluation becomes goal-based.

---

[15] For example, program theory models of evaluation (e.g., Chen, 1990; Rogers, 2000; Weiss, 1997) typically require goal-orientation.

2. The GFE must be designed and conducted completely independently of the GBE and by separate evaluators who do not communicate with the GFE team anything implying preordinate goals or objectives.

Fields of Evaluation

According to Scriven (2005a), there are numerous field of evaluation; the seven most common fields are referred to as the "seven Ps": program, product, performance, policy, portfolio, personnel, and proposal. Program evaluation, product evaluation, and performance evaluation have the most easily recognizable relationship in terms of employing with GFE.

Program

Program (project) evaluation "receives the most attention and has the most well-developed principles, procedures, and practices" (Coryn, 2007, p. 61). Most discussions and debates regarding GFE revolve around its applications in program evaluation. In program evaluation, GFE is analogous to liability insurance protecting the insured from claims of inappropriate actions and negligence as the goal-free evaluator searches for all actual effects (i.e., outcomes and impacts) and is unaware whether it is an intended effect or a side effect. Especially when the evaluator uncovers and reports unintended negative effects, the evaluator provides the evaluand the potential to avoid circumstances that may have, in the future, been considered inappropriate or negligent.

Product

Product evaluation generally refers to the evaluation of artifacts or physical objects that have been created or developed by another (Scriven, 1991). As noted earlier, GFE has existed within product evaluation with consumer magazines and book reviews epitomizing the goal-free approach to evaluation. Rarely does the product evaluator need to know the intentions of the product manufacturer or designer in determining the merits of the product. For example, based on a general understanding of the definition and function of a cellular telephone, one can evaluate it without asking LG, Ericsson, Kyocera, Motorola, or other phone manufacturers what they intended when creating the device.

Performance

Performance evaluation, which refers to the evaluation of a particular achievement (Scriven, 1991), is often goal-free in nature. A symphony conductor holding auditions for a new principal cellist is unaware of the musician's specific goals and intentions when s/he plays a self-selected piece; rather the conductor judges the performance based on the performance outcomes and compares them to the conductor's criteria of merit and performance standards.

Policy

Regarding policy evaluation, Coryn (2007) states that "the evaluation of policy is normally either retrospective evaluation of implemented policy or prospective evaluation of possible policy or comparison of alternative policies" (p. 65). Typically a policy

evaluation requires that the evaluator have a firm understanding of what the policy is intending to do. Therefore, GFE is rarely suited for policy evaluation as the various policy statements reflect the policymakers' goals and objectives and thus is heavily goal-based in nature. However, there is some potential for using GFE with retrospective policy evaluation; for example, an evaluator without extensive knowledge of the details of the No Child Left Behind Act of 2001 may conduct an evaluation of the policy at a particular school district by examining all student outcomes and by knowing that the basic aim is improved student academic performance. GFE may also be used with two or more policies to determine the ranking of the policies according to their actual outcomes and effects; comparing the outcomes of harassment policy X with the outcomes from harassment policy Y, for example. GFE is not suited for prospective evaluation as GFE only searches for actual outcomes (ones that have occurred or are occurring) not predicted outcomes.

Portfolio

For artists, architects, musicians, teachers, etc., portfolio evaluation usually refers to "a body or selection of professional achievement" (Coryn, 2007, p. 66). In some cases, portfolio evaluations are goal-free such as when an architect provides a prospective client with a portfolio (e.g., drawings, blue prints, and photos of completed structures); the client evaluating the portfolio may be unaware of and unconcerned with the architect's particular intention. Rather, the client examines such criteria as mass, form, space, texture, volume, light, shadows, materials, cost, functionality, originality, construction materials, and aesthetics, for example.

In business and finance, a portfolio typically refers to the investment portfolio, which is the aggregate investments of an individual or institution (e.g., stocks and bonds, mutual funds, real estate, and art collections). In general, the ultimate aim of the investor is explicit, the reduction of risk through diversification; however, the investor's specific intention in diversifying is less apparent leaving the possibility of using GFE.

Personnel

Personnel evaluation refers to the evaluation of a "person's qualifications or performance in relation to a role and larger defensible purpose" (Stufflebeam, 2005, p. 308). Personnel evaluation can be goal-free. Typically, administrators and managers use personnel evaluation during the hiring, promoting, and firing processes. When the evaluator is the employer or prospective employer, the evaluand's (i.e., the prospective employee or current employee) work goals and objectives are nearly completely irrelevant. In this case, the evaluator's goals and objectives are usually the only ones that matter.

Proposal

Proposal evaluation generally refers to a document submitted to an individual or institution requesting monetary or non-monetary support. Commonly, a proposal is written by a program and sent to a prospective funding organization. A proposal evaluation is an assessment of the requester's ability to perform the perspective tasks and whether the program's anticipated outcomes are worthwhile (Coryn, 2007). Similar to personnel evaluation, proposal evaluation is nearly always goal-free as the proposal

writer's goals are irrelevant and the funding organization's mission, purposes, and intentions are the basis for judging the proposal.

In summary, GFE has a least some applicability with each of the seven fields of evaluation. This demonstrates the ubiquitous potential for conducing GFE. Still, program, product, and performance evaluation fields are generally where the application of GFE is most hotly debated.

Rules of Inference Governing GFE

GFE is subject to the same rules of inference governing other evaluation approaches and models.[16] According to Youker (2010), the general logic of evaluation that crosses all evaluation fields and approaches also applies to GFE. "It is the basic reasoning that specifies what it means to evaluate something" (Fournier, 1995, p. 17) and hence provides general rules of inference for all evaluations. There are four operations all of which require systematic and justifiable methods and conclusions. First, if one is to judge something, one must determine the criteria by which to judge it. Second, standards describing how performance should look are constructed or found (e.g., poor through excellent). Third, the evaluand's performance is measured on the identified criteria and then compared to the performance standards. Fourth, synthesis, or the combining of evidence on several dimensions or subdimensions, is employed to draw one or more evaluative conclusions.

Although both goal-based approaches to evaluation and GFE adhere to the above general logic of evaluation, there is a rational basis for inferring that GBE and GFE are

[16] For a general discussion of the rules of inference, see Coryn (2007), Scriven (1991, 2007), and Davidson (2005).

distinct. The difference between them is in the application of the general logic, what Fournier (1995) calls *working logic*. Consider the difference in how an evaluand is operationalized; in the GBE approach, an evaluand is a series of goals and objectives, whereas in GFE, an evaluand is something designed to meet the needs of a particular consumer group in a particular context. This dissertation poses that GBE and GFE are in fact distinct. Presented below is a further elaboration of GFE's principles.

Principles of GFE

A GFE refers to an evaluation conducted with the evaluator unaware of the upstream stakeholders' stated intentions. In a GFE, the evaluator intentionally avoids learning the official or stated goals and objectives of the evaluand, evaluation client, and other upstream stakeholders; rather, the evaluator observes and measures the actual outcomes and judges the evaluand according to broad-based outcomes founded in logical premises such as the program's performance in meeting the consumers' needs. From a teleological perspective, the goal-free evaluator begins by focusing on the examination of definitional premises or what Aristotle calls the *essential properties* of the evaluation object. Aristotle taught that "all things had two kinds of properties: essential properties, without which they wouldn't be the particular *kind* of thing they were, and accidental properties, which were free to vary within the kind" (Dennett, 1995, p. 36). To illustrate, the goal-free evaluator evaluating a literacy program for functionally illiterate adults first tackles the essential properties which might include criteria like gains in reading and successful utilization of adult learning techniques. Accidental properties of a reading

program for adults might include the quality of the particular textbook, teaching style, affordability of the program, and convenience of the location.

GFE's rationale is that stated goals and objectives are unnecessary noise for the external evaluator (Scriven, 1972). If one accepts the definition of evaluation as the systematic determination of merit and since the program was designed to meet some relevant needs of a target consumer, the evaluator sees that the program's intentions are not required in determining what makes the program good or bad. In fact, goals and objectives often prevent the recognition of relevant unintended positive and negative side effects and side impacts. Thus, to aid the evaluator in conceiving and then observing all possible areas for relevant actual outcomes, the goal-free evaluator is screened from specific goal-oriented information. Scriven (1991) also points out that if the program is doing what it intends, then many of the criteria identified by the goal-free evaluator should match the program's stated goals. Thus, the determination of the criteria of merit is prescriptive (i.e., what should be), whereas the goal-free evaluator also attempts to describe what is.

The specific principles of GFE are as follows:

1. Identify relevant effects of which to examine without referencing goals and objectives.
2. Identify what occurred without the prompting of evaluand goals and objectives.
3. Determine if what occurred can logically be attributed to the intervention.
4. Determine the degree to which the effect(s) are positive, negative, or neutral.

Thus, the ultimate question for the goal-free evaluator is: What occurred that can be attributed to the evaluand?

Bias Control

GFE, in theory and in method, is designed to increase the evaluator's ability to control potential biases. The reduction of bias is a fundamental concern in all systematic inquiry; typical bias reduction methods include sampling, randomization, blinding, statistical controls, and triangulation (Henry, 1998). Before delving in bias reduction methods, a few definitions are needed.

Bias refers to prejudice, partiality, unfairness, and/or subjectivity (Mathison, 2005a), while *bias control* is "an attempt to limit the influence of unjustified views, e.g., premature or irrelevant views" (Scriven, 1991, p. 69). GFE aids in controlling bias through its independence from stated goals. According to Kushner (2005), independence is "a stance for an evaluation that is not subject to the control of or that does not provide privileged access to any particular stakeholder group or constituency" (p. 198), and in this case the evaluation is not controlled by the evaluation's upstream stakeholders and their intentions. The way this is accomplished is through the process of blinding.

Blinding is a primary tool of the goal-free evaluator in controlling bias related to goal-orientation. In a GFE, blinding is attempted by systematically blinding (i.e., concealing) stated program intentions from the view of the goal-free evaluator. According to the *American Heritage Dictionary*, "to blind" is "to deprive of perception or judgment"; and, suitably analogous in evaluation, the definition for *blinder* is "one that blinds; a pair of leather flaps attached to a horse's bridle to curtail side vision" (Pickett,

2000, p. 188). So, therefore, the evaluator uses the goal-free approach to eliminate the blinders that lead toward stated goals.

The analogy of the removal of racehorse's winkers allowing the viewing of her periphery for "side" effects, as opposed to tunnel-vision toward the goal (i.e., finish line), is akin to the goal-free evaluator's removal of goal-orientation to prevent the tunnel-vision guiding the evaluator toward seeing only effects related to preordinate goals. This perceptual blindness biases the evaluator and contaminates the evaluator's ability to see the evaluand's "true" outcomes and "true" merit." Tunnel-vision toward goal-orientation can heavily influence program administrators and practitioners as well. Scriven (1972) says, it is not a matter of honesty but rather one of failing to see the forest for the trees; and Patton (1997) agrees and adds that the "difficulties in clarifying a program's goals may be due to problems inherent in the notion of goals rather than staff incompetence, intransigence, or opposition to evaluation" (p. 180). Nonetheless, in GFE, the independent external evaluator's ignorance of specific goals is deemed a positive; thus, the evaluation approach intends to maximize this independence. Consequently, one of the main determinants of whether GFE is appropriate in a given situation is whether the evaluators are, in fact, independent and external to the evaluand and other stakeholders; and, more importantly, whether the prospective goal-free evaluators can be considered *tabula rasa* in terms of their awareness of the program's goals and objectives.

In his early writings on GFE, Scriven (1974b) makes the analogy between GFE and the double-blind pharmaceutical study. The goal-free evaluator, like the pharmaceutical evaluator, does not need to know the direction of the intended effect or the intended extent of the outcomes as they are hardly relevant in determining merit. In

pharmaceutical studies, a double-blind study refers to an experiment where neither the individuals being studied nor the researchers know who belongs to the control group or the experimental group. In the case of GFE, the evaluator does not know which effects are goals or objectives and which are side effects. Only after all the evaluation data have been recorded (and in some cases analyzed, i.e., triple-blind GFE) do the inquirers learn which effects are which. Thus, screening the intended effects from the evaluator is a critical part of this double-blind research design if the goal-free evaluator is to examine *all* relevant effects.

It is important that the evaluator and the upstream stakeholders agree to adherence to the rules of blinding by willingly participating with GFE's screening requirements; if this cannot be agreed upon or logistically arranged, GFE may not be the most appropriate evaluation approach. The methodological requirements of GFE dictate that in the majority of cases someone considered impartial (i.e., not assigned to GFE design and data collection) is required to serve as the screener like an administrative assistant, a third party, or the client (Youker, 2005a, 2005b). Therefore, one of the first orders of business is for the goal-free evaluator and evaluation client/evaluand to appoint a screener. According to Evers (1980), a screener is

> an individual who assists the goal-free evaluator during early stages of the evaluation both in terms of editing materials and serving as a liaison to the project staff. This person serves as a critical buffer between the evaluator and sources of bias while the goal-free evaluator is trying to employ strategies of discovery and investigation to uncover actual effects. (p. 40)

The screener's role is to conduct the initial meetings with the evaluation clients/evaluand to omit any and all goal-oriented communiqués and documents from the goal-free evaluator. The screener searches all documents and archival records to keep the

evaluators from the program's goals and objectives; however, some sources have a greater or lesser relative likelihood of requiring omissions in blinding the goal-free evaluators from the goals. For instance, this goal-oriented information is often found in program promotional materials, grant proposals, progress reports, staff training materials, and evaluation reports; and found by communicating with program administrators, managers, staff, funders, and clients. It is worth noting that simply learning the names of the cooperating organizations may lead one to infer the evaluand's general aims; however, identifying the program's specifically stated objectives is not so obvious. Furthermore, even if someone accidentally tells the evaluator a goal or objective, it does not mean that s/he accurately stated it (Scriven, personal communication, February 22, 2007).

Table 2 below is a summary of potential sources for finding goal-oriented information. The material is based on the writings of Scriven (1973, 1974b, 1991) and uses the format provided by Evers (1980). This table is intended to serve as a general tool for recognizing common potential evaluation-related situations and materials and their relative likelihood of being a source of goal-related information, thus requiring omission in screening them from the goal-free evaluator. On the column on the left side of the table are the "sources for goal-based information" (i.e., situations and documents). These sources are divided into categories based on chronological stages of evaluation. On the right side of the column is the screening level, which essentially is a prescriptive rating of the relative level of attention, effort, and thoroughness required by the screener in the screening processes. Below the table is a description of the level of caution or screening recommended for maintain the goal-free nature. The screening level is highly debatable

and should be considered an approximation, as it is beyond the scope of this dissertation to empirically investigate the actual frequency of goal-based information from these sources. However, adding further credibility to the table is the fact that it was examined and accepted by Scriven (personal communication, February 22, 2007).

Table 2

Sources of Goal-Oriented Information and Requiring Screening Level

	Source of Potential Goal-Based Information	**Screening Level**
1.	**Pre-Site Visitation**	
A.	Initial Contacts (e.g., phone calls, emails, face-to-face, etc.)	High-Level
B.	Parts of the Program Proposal	
	1. Overview of the problem	General-Level
	2. Introductory Passages	Moderate-Level
	3. Program Descriptions	Moderate-Level
	4. Client profiles	General-Level
	5. Needs assessment data	General-Level
	6. Mission statement	Screened Entirely
	7. List of partnering organizations/programs & relationships	Moderate-Level
	8. Goals & objectives(other advance organizers)	Screened Entirely
	9. Proposed strategies	Screened Entirely
	10. Proposed activity plan(s)	Screened Entirely
	11. Proposed staffing plan	High-Level
	12. Summary Passages	Moderate-Level
	13. Proposed budget	General-Level
C.	Target Group/Evaluator Interactions	
	1. Check target group (i.e., consumer) needs	Moderate-Level
	2. Check target group treatment effects (outcomes)	Moderate-Level

Table 2—Continued

	Source of Potential Goal-Based Information	Screening Level
D.	Representative Project/Program Materials	
	1. Curricular- texts, study guides, & pre-posttests	Moderate-Level
	2. Program brochures & promotional materials	Screened Entirely
	3. Program training materials	Moderate-Level
	4. Non-curricular- environmental or experiential or "gestalt"	Moderate-Level
	5. Staff/Employee rosters & demographic information	General-Level
	6. Program staff job descriptions & responsibilities	Moderate-Level
	7. Policy manuals	Moderate-Level
	8. Organizational flowchart	General-Level
	9. Client flowchart	Moderate-Level
	10. Client eligibility requirements	Moderate-Level
	11. Contracts/agreements between the program & consumers	Moderate-Level
E.	Process Observation of Treatment	Moderate-Level
F.	Internal Evaluation Data	High-Level
G.	External Evaluation Data	Moderate-Level
H.	Historical/Archival	
	1. Minutes of staff meetings	Moderate-Level
	2. Budget status reports & annual reports	Moderate-Level
	3. Internal staff correspondence	Moderate-Level
	4. Correspondence between project & funding agent	Moderate-Level
	5. Miscellaneous progress reports	Moderate-Level
	6. Client demographic information from intake forms	General-Level
I.	Overview of Research/Literature in Area of Investigation	General-Level
2.	**On-Site**	
J.	Staff/Evaluator Interactions	
	1. Staff introductions to the project	High-Level
	2. Staff public relations tours	Screened Entirely
	3. Final debriefings	Moderate-Level
	4. Data about long & short-term effects or benefits	Moderate-Level

- General-Level Screening: Refers to situations or documents with minimal likelihood of requiring significant goal-oriented omissions, thus requiring basic-level screening such as having the evaluator send an email to program people to remind them of the goal-free nature and having a screener conduct a once-over of the document.
- Moderate-Level Screening: Refers to situations or documents with moderate likelihood of requiring significant goal-oriented omissions, thus requiring more diligent screening such as having the evaluator send multiple emails to program people, having the screener re-read (i.e., re-screen) documents, or having multiple screeners for reading the same document.
- High-Level Screening: Refers to situations or documents with high likelihood of requiring significant goal-oriented omissions, thus requiring robust screening such as having the evaluator ask program people to make a screened version of the entire document, specifically targeting the goal-based material by sending internal memos to or meet with program people to ensure they understand what to share with evaluators, and/or using multiple screeners of documents.
- Screened Entirely: Refers to situations or documents that are goal-specific in nature, thus requiring entire omission or only carefully selected excerpts are given to the goal-free evaluator.

There are numerous sources of evaluation bias and attempting to list them all is beyond the scope of this dissertation; however, offered below are common types of evaluation bias, which GFE serves to reduce or control through its double-blind approach.

- Observer Bias: Observer bias occurs when the researcher or observer knows the goals of the program and allows this knowledge to influence what gets observed and the depth of the observations during the evaluation. Goal-free evaluation is designed to shield evaluators from the program goals and, consequently, the associated social, perceptual, and cognitive biases (Scriven, 1991).
- Experimenter Bias: Experimenter bias is any source of error introduced in an evaluation in the way the evaluation is designed, the data are collected and analyzed, and conclusions are drawn. GFE is designed to correct the error of neglecting relevant criteria and outcomes that are critical in determining merit.
- General Positive Bias: General positive bias refers to the tendency of evaluators to turn in more favorable results than justified (Scriven, 1991). GFE is designed to enhance evaluator independence and reduce the propensity for general positive bias in that the evaluator does not know which results or effects are goals and which are side effects.
- Financial Relationship Bias and Organizational Relationship Bias: Financial relationship bias and organizational relationship bias are closely associated and, like general positive bias, are related to evaluator independence. Financial and organizational biases are introduced through the relationship between the evaluation client/evaluand and the evaluator. Whenever an evaluator is contracted to evaluate a program (and even before), a relationship begins; or whenever the evaluation is based on a

> pre-existing relationship between the evaluation client/evaluand and the evaluator, potential threats to evaluator independence are inherent.[17] In the real-world, an evaluation is often an audition for a second evaluation or at least for a positive referral; thus, there is always a motive for downplaying poor results and exaggerating positive results. However, as mentioned in general positive bias, GFE limits the evaluator's inclination and ability to placate the evaluation clients by "giving them what they want" because the evaluator does not know specifically what it *is* they want.

Consumerism

GFE is founded in a consumerist ideology as its underlying philosophy emphasizes a balance between consumer and administrator, an examination of consumer need, and a consideration of the *cui bono(?)* principle. The goal-free evaluator ignores the program's intentions in favor of its consumers' outcomes (see *Consumers Union*, p. 34), which is often labeled the consumer-oriented approach to evaluation. Stufflebeam (2001) says that these consumer-oriented approaches are designed to protect the consumer from poor programs, practices, and products, and assist consumers in choosing the highest quality services in meeting their needs. Fitzpatrick et al. (2004) state:

> The consumer-oriented approach to evaluation is predominantly a summative evaluation approach. Developers of products have come to realize, however, that using the checklists and criteria of the consumer advocate while the product is being created is the best way to prepare for subsequent public scrutiny. Thus, the

---

[17] Scriven (1991) notes that lack of independence is not proof of bias, rather higher probability of bias; for example, when the evaluators have a pre-existing relationship with an evaluation client, it does not mean they are biased, just that there is a higher probability of bias than if there were no pre-existing financial or organizational relationship.

> checklists and criteria proposed by "watchdog" agencies have become tools for formative evaluation of products still being developed. (p. 101)

Stufflebeam (2001) adds a succinct summarization of the consumerist philosophy underlying GFE and, although he was solely referring to consumer-oriented approaches in general, the following statement applies to GFE in particular: "The approach regards a consumer's welfare as a program's primary justification and accords that welfare the same primacy in program evaluation" (p. 58).

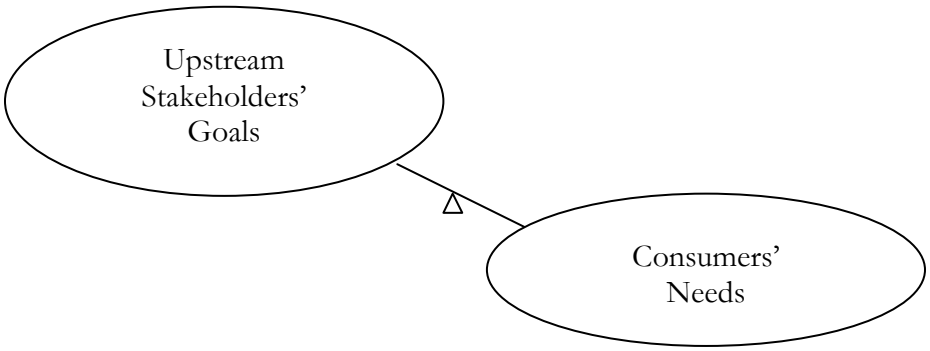
There are some frequently practiced methods and techniques among the consumer-oriented approaches. The consumer-oriented approaches use advance organizers (Woolfolk, 2001) such as "societal values, consumers' needs, cost, and criteria of goodness in the particular evaluation domain" (Stufflebeam, 2001, p. 59). Additional common practices include the application of varying methods such as "checklists, needs assessments, goal-free evaluation, experimental and quasi-experimental designs, modus operandi analysis, applying codes of ethical conduct, and cost analysis" (Stufflebeam, 2001, p. 59).

The evaluator who subscribes to the consumerist ideology is attempting to equalize the power between consumerism and managerialism in an evaluation. One way the evaluator accomplishes this is by balancing the program's needs and wants with the needs and wants of the consumers as well as by balancing the power among the evaluator, the upstream stakeholders and evaluation client (Scriven, 1974b). By its design, GFE shifts power from the program (i.e., the evaluand) to the evaluator in that the program's goals are omitted and the evaluator judges the evaluand according to independent and justifiable criteria based on actual outcomes.

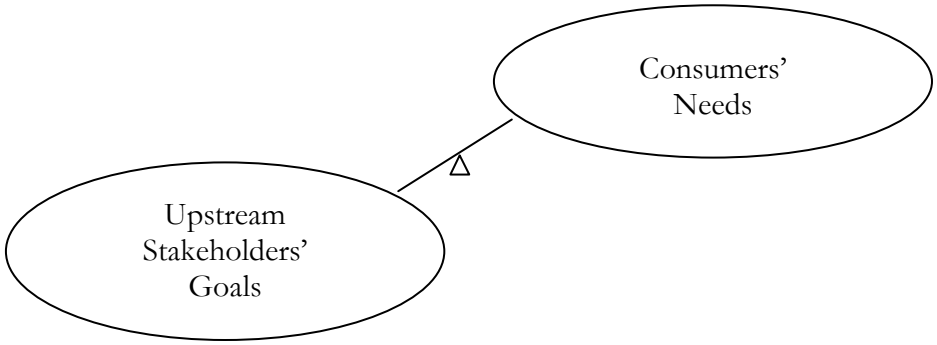
Figure 1 below is the theoretical representation of the balance of power at three stages—before, during, and after implementing a GFE. The first box (i.e., Before Implementing GFE) illustrates that prior internal and external evaluation efforts are generally goal-based and reflect the upstream stakeholders' goals. The consumers' needs are usually important to the program but tend to be secondary relative to the stated goals of the evaluand. However, during GFE (i.e., second box) the balance of power shifts from the upstream stakeholders to the goal-free evaluator via the evaluator's screening from these stakeholders' goals and, thus, the goal-free evaluator tilts power from the managerial intentions of the upstream stakeholders to the needs of the evaluand's consumers. Finally, after GFE (i.e., third box), the evaluation balance of power equalizes and is closer to level compared to before and during the GFE. After GFE, the evaluation client and other upstream stakeholders regain much of their power in terms of deciding what to do with the evaluation findings, e.g., what changes to make and not to make and what to publish or not, etc. Ideally the GFE produces some useful findings regarding the evaluand's actual outcomes and, if so, power levels when the program people use some of this information to improve the program or use it to enhance future monitoring and evaluation efforts. Lastly, in counterweighing managerialism, the consumer needs assessment is frequently used in developing the goal-free evaluator's criteria of merit for judging the evaluand, thus directing the evaluator in the search for relevant effects rather than upstream stakeholder goals.

GFE has a relationship with needs assessment. Davidson (2005) says that "goal-free evaluation is sometimes called needs-based evaluation because needs assessment is one of the primary tools used to identify what effects (both positive and negative) should

Before Implementing GFE



During Implementing GFE



After Implementing GFE

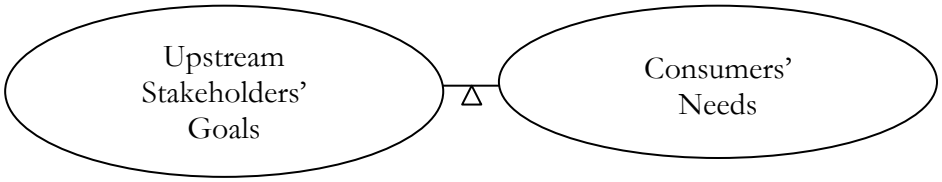


Figure 1. Conceptual Representation of Balance of Upstream Stakeholders' Goals with Consumers' Needs Before, During, and After Implementing a GFE

be investigated" (p. 241). In fact, needs assessment is not the sole property of GFE as it can be a critical part of conducting any evaluation. Sanders (2006) says that even for mature evaluands "it is important to take stock, to do a rolling needs assessment, every year or so to guide the development [of a program]" (p. 58).

A needs assessment is the process for uncovering the facts about the function, or lack thereof, of organisms or systems, while the definition of a need is anything essential for a satisfactory level of performance or mode of existence (Youker, 2006). In GFE, the orientation is toward the primary needs of the consumer as the consumer is *raison d'être* for the service deliverers and delivery systems (Altschuld & Witkin, 2000; Scriven, 1991). Weiss (1998) says the following about the involvement of program consumers in evaluation efforts:

> Program *clients* also have a big stake in the program. I wish their inclusion were more widespread, especially clients from marginal groups in society. Difficult as it is to engage them in evaluation activities, they are apt to have different interests and concerns from those of staff, and addressing their questions would broaden the scope of the study. (p. 30)

Consumerism is an ideological response to the *cui bono(?)* question, i.e., who benefits from this? Who is the true or actual beneficiary in an evaluation? There are three realistic "suspects": (1) the evaluator; (2) the evaluation client, evaluation user, and other upstream stakeholders; and (3) the evaluand's consumers and impactees. Is the goal-free evaluator the real winne, or are program funders, administrators, managers, and staff the beneficiaries of the evaluation? Do the true benefits of GFE befall the program's consumers? Theoretically, a sound evaluation should, to varying degrees, benefit all three.

The following are a few considerations in answering *cui bono*. Some criticize GFE claiming that it is easier than a GBE because, in the same way it is difficult for the evaluator to cheat in favor of the program, the goal-free evaluator cannot be "wrong" as the "correct" goals, objectives, and outcomes are not known to the evaluator. Therefore, given quality inquiry and justified decisions and conclusion, whatever the goal-free evaluator claims to observe and then submits in the evaluation report is "right." However, according to Scriven (1991), GFE requires increased effort on the part of the evaluator and therefore identifies evaluator incompetence. For argument sake, accepting the view that GFE is more difficult than GBE, it seems easier for the evaluator to avoid introducing GFE altogether in favor of sticking exclusively with a GBE instead of employing the challenging GFE only to produce a vainglorious report. Continuing this line of thought, it is always possible that an individual evaluator has malicious intent, a political objective, or some other conscious or unconscious bias influencing him/her, but barring ulterior motives, incompetence, and the normal sources of bias, GFE may be the more difficult approach and, if so, the goal-free evaluator is unlikely the ultimate beneficiary of the employment of this approach.

Both the program people and the program's consumers are intended beneficiaries of GFE. Upon introduction, GFE may be viewed with skepticism or as a threat by program staff and administration. However these upstream stakeholders are the theoretical prime beneficiaries of a GFE as the program people become the evaluator's consumer, while the program's consumers are the evaluator's downstream impactees or secondary consumers. Although the consumers needs are *raison d'être* for the program, the program people's needs are *raison d'être* for the evaluator. This does not imply that

the consumer-oriented evaluator uses the program's goals; rather the evaluator accepts an implied program mission, the fulfillment of the consumers' relevant needs. To illustrate, similar to a workshop for elementary school educators, the workshop participants are the reason the trainer exists; however, the teachers in the workshop and school district's needs in educating the students are the reason the workshop is needed and hence the trainers/facilitators. To offer another example, the poor do not require an evaluation; the programs of which the poor are (potentially) involved require the evaluation. Not until the individual involves him/herself in pursuing services from an evaluand (i.e., program, policy, product, etc.), or is impacted by the evaluand, does the person become a consumer (or potential consumer) and an evaluator's potential subject of inquiry.

Another reason to adopt caution in terms of stated goals is the "trendiness" of products and programs. For example, there is a "green company" fallacy where numerous product manufacturers claim to be green (i.e., environmentally sound) yet some companies have done nothing to change their products or product model; they simply changed their marketing and image strategies. Similarly, nearly every American is aware of a fly-by-night weight loss program that becomes the fad of the day but later is abandoned or discredited. GFE is designed for the program's consumers to be the long-term actual beneficiaries of the evaluation through their involvement with a high quality, accountable, constantly improving program. In inquiring into the rules of inference governing GFE, this section discussed bias control and consumerism; and next, concluding this section, is an examination of physicality, design, and intentionality.

Physicality, Design, and Intentionality

The goal-free evaluator does not discount the intentions of the evaluand entirely; on the contrary, the evaluator often attempts to ascertain the program's actual intentions by observing actual practice and actual outcomes. In doing so, the evaluator inquires into the program's physicality, design, and intentionality. Dennett (1987) developed a three-part classification of stances of which all Homo sapiens are said to subscribe when they attempt to comprehend and predict the behavior of entities such as animals, other humans, and machines. The three stances are (1) the physical stance, (2) the design stance, and (3) the intentional stance.

According to Dennett (1987), in principle, the physical stance will always prevail since everything ultimately adheres to the laws of physics or laws of nature; however, for the inquirer, finding answers in this way is arduous and the prediction of the object's behavior is rarely accomplishable in a realistic or timely fashion. For something that is actually designed like a car, a skyscraper, a portfolio, or an educational program, the process of predicting the behavior of the object can be expedited by circumventing physics and appealing to the design of the object. For example, a great leap in understanding occurs when the assumption that the pancreas is designed to aid in the digestion process and hormone production. Scriven (1991) uses the analogy of a wristwatch, saying that what is needed to evaluate the merits of a watch is a shared "understanding of the meanings of the terms describing the evaluand" (p. 217). For Scriven, the criteria of merit are founded on what the watch is designed to do: keep accurate time, be legible, be durable, and be aesthetically pleasing. Dennett also uses

product evaluation and a time-keeping device in his analogy of the design stance. Dennett (1987) writes:

> Almost anyone can predict when an alarm clock will sound on the basis of the most casual inspection of its exterior. One does not know or care to know whether it is spring wound, battery driven, sunlight powered, made of brass wheels and jewel bearings or silicon chips—one just assumes that it is designed so that the alarm will sound when it is set to sound. (p. 182)

The views of Dennett mirror that of Scriven and underscore the ideological and methodological arguments for the legitimacy of GFE.

The highest order stance, the intentional stance, is one in which the object is not only designed for an aim but also consists of an agent with intentions that direct the entity's behavior. When a hiker sees a black bear, the hiker attempts to predict the bear's probable action; in doing so, valuable time is wasted on ascertaining the bear's physical or design characteristics. The bear's physical make up of cells, blood vessels, organs, two eyes, and fur as well as the fact that the bear's jaws and claws are designed to devour are irrelevant when face-to-face with one in the wild. Rather, the hiker would be wise to forego pondering the first two stances in favor of deciphering the bear's intentions, more specifically whether the bear intends to accost him/her or not. Continuing with a second analogy, consider the following scenario. A bank alarm sounds; you see a masked man dressed in black with a grappling hook in one hand and a duffle bag in the other running down the alley away from the bank. It would be preposterous to stop the individual to inquire as to his intentions as they seem quite apparent. Even if you did slow him down long enough for him to respond to you, could you trust his answer? Now imagine a man chasing the robber, brandishing a gun, and screaming for him to "freeze"; the man in pursuit is obviously someone who intends to stop this individual. In the case of program

evaluation, the true agent is often vague hence complex to define and even if an agent *is* specified, it is another daunting task to determine that agent's actual intentions. Although the immediate intentions of the man with the gun chasing the robber is clear, i.e., he intends to stop the bank robber; it is unclear as to his "true" intention (e.g., to steal the money from the robber, to kill the robber, to arrest the robber, to be heroic, etc.), again demonstrating the difficulty of deriving true long-term intentions (the same goes for the alleged bank robber's long-term true intentions). A true intention should be demonstrated via the actions (or inaction) of the program and its stakeholders and therefore is potentially observable; furthermore, if the action is not evident to the evaluator, it is possible it is not a "true" intention, or the observable effects produced by the program toward these intentions and the consumers' outcomes are trivial. In other words, the theory is that if the program does not produce observable outcomes that are detected by the goal-free evaluator, then the program should examine both its intentions, and the alignment of its attention and effort toward producing positive effects with regard to these intentions.

Just as the design stance can be used with things that are not actually designed, the intentional stance works for entities which both have and lack conscious intention. This point is fundamental in the argument for the legitimization of GFE. In accordance with Dennett's intentional stance, the goal-free evaluator avoids knowledge of the explicit intentions of the entity (i.e., evaluand) as the intentions (i.e., goals and objectives) of the entity are secondary in determining merit; and it is resource consuming to ascertain the true intentions behind the evaluand. Identifying true intentions is an imprecise task as an evaluand's actual intentions includes an accumulation of the

upstream stakeholders' "true" individual personal and professional intentions and motivations. Moreover, a large program is similar to what Dwight Eisenhower (1960) coined, the military-industrial complex; it consists of numerous intricate systems that work quite independently of each other and sometimes toward undefined, conflicting, or contradictory goals; a situation reminiscent of the adage: the left hand does not know what the right hand does. With real-world resources and contextual limitations, it is frequently inefficient for an evaluator to spend valuable energy articulating whose goals to use and which combination of goals to use. The point here is that it is arguable whether a program ever has or is capable of conscious knowledge of its own true intentions; nevertheless, goals are unnecessary for the evaluator in determining merit. Consequently, the goal-free evaluator dismisses the imprecise official, (i.e., stated) intentions of the program; instead, the evaluator relies on an analysis of the program on the assumption that it lacks consciousness of its own intentions. The goal-free evaluator assesses the physical and design properties of the program as well as observes the program's actions and outcomes to determine its intentions. In cases when the goal-free evaluator is asked to assist the evaluation client with program goal alignment, this information is used to develop various hypotheses regarding the program's observed actual intentions (see *The Goal-Free Evaluator's Criteria as a Tool for Goal Alignment*, p. 93).

In conclusion, a continuance of Dennett's bear analogy, it is arguable whether animals—or black bears, to be specific—possess the ability to be conscious of their own intentions; anyway, it is not important especially when trying to decide whether that bear wants a human snack. Furthermore, it would be preposterous to question mama bear as to her goals, yet it is very possible to collect data on what she plans to do by seeing what she

does. The hiker may determine that the bear is quarrelling with other bears; she is wandering her territory; she is foraging for berries; she is preparing to hibernate; and so forth. Basic safety requires the hiker to assess risk by first considering the most imminent threat: does the bear intend to attack me? As the hiker further observes, two small cubs come from the brush, mama bear turns around and frolics with her cubs. In light of this new information, the hiker reconsiders the earlier hypothesis that the bear intends to accost him/her. As the hiker watches the bears play, s/he slowly backs away, makes the way to the trail, and heads for his/her cabin. With basic knowledge of the creature's physical prowess and predatory design, the hiker assesses the situation paying particular attention to the most serious predictions with the most critical outcomes. After some observation, the hiker is able to predict that in the immediate, the bear does not intend to eat him/her; rather, she intends on entertaining her young. Of course, the hiker realizes the short-term nature of this prediction and takes appropriate measures in response.

The rules of inference governing GFE are founded in bias control; consumerism; and physicality, design, and intentionality. Next in the examination of the rules of inference governing GFE is a conceptual framework of the outcome scenarios between the goal-free evaluator's criteria of merit and the program's stated goals following a pre-evaluation assessment. This chapter concludes with a summary of GFE's potential benefits and the criticisms of GFE with responses to these criticisms.

Conceptual Framework of the Outcome Scenarios Between the Goal-Free Evaluator's Criteria and the Program's Stated Goals

The goal-free evaluator creates a list of criteria of which to observe evaluand performance and the criteria from this list may or may not overlap with the stated goals of

the program. Often an evaluator conducts a needs assessment to identify the consumers' relevant needs and uses this information during the determination of the evaluand's criteria of merit. Throughout this process, there are five relationship possibilities between the goal-free evaluator's established criteria and the program's stated goals and objectives. Below, in Figure 2, the five scenarios are conceptually represented via Venn diagrams. A more in-depth illustration of the second scenario is offered as it is by far the most plausible for all GFEs while Scenarios One, Four, and Five are hypothetically plausible but much less probable, and Scenario Three is fairly improbable.

In all five of the scenarios, there is the assumption that the program and/or evaluation client possess a working list of stated goals and objectives, and that the goal-free evaluator will identify relevant criteria. Furthermore, it is assumed that terminology, rhetoric, and jargon differs between the goal-free evaluator and the program yet via literature review and content analysis, the evaluator is capable of reasonably determining which of the terms used by the goal-free evaluator as criteria are similar or synonymous to terms used by the program as a goal or objective.

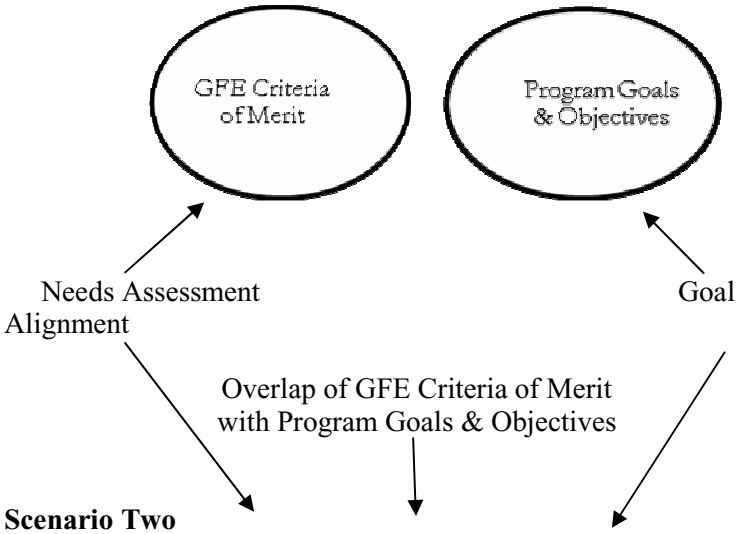
To introduce the Venn diagrams, a brief understanding of the meanings behind the ovals, arrows, and labels is provided. First, there are five series of sometimes overlapping ovals representing the various relationships possible between the goal-free evaluator's criteria and the program's goals. Overlap is significant. Although, it may not be justified in all cases, but where there is overlap (i.e., represented in Scenarios Two, Three, Four, and Five) between the goal-free evaluator's criteria and the program's stated goals, we might congratulate the program for choosing these goals, for working toward these goals, and for possibly producing outcomes on these goals; we might also praise the

goal-free evaluator for detecting the effects and recognizing these outcomes. On the other hand, criteria identified by the goal-free evaluator but not stated as program goals theoretically represent criteria where the program should at least consider targeting its attention, or it represents an evaluator who chose irrelevant criteria, whereas the criteria stated by the program as goals but were not identified by the goal-free evaluator represent criteria where the program might seriously bolster its efforts, or adjust programming, or reconsider whether the criteria should be a goal. The arrows between Scenario One and Scenario Two merely acknowledge that both the goal-free and goal-based evaluators often derived their respective criteria from pre-evaluation assessment methods. The goal-free evaluator uses the consumer needs assessment while the goal-based evaluator (with general exception of the goal achievement evaluator) conducts a goal alignment. Below are the models of the five possible scenarios.

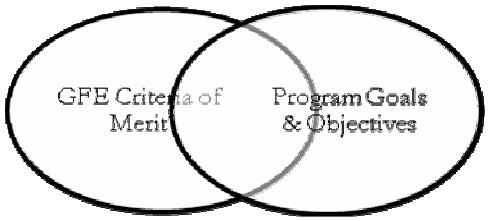
The first possible outcome is displayed in Scenario One where the criteria of merit as defined by the goal-free evaluator (i.e., left oval) is completely independent of the official preordinate criteria stated in the program goals and objectives (i.e., right oval). Scenario One illustrates a situation where none of the goal-free evaluator's criteria or observed effects match, or are synonymous, with the criteria stated by the program in its goals. If a goal-free evaluator was to determine criteria and none are congruent with the program's goals, this represents a situation where (1) the evaluator's capacity and impartiality should be examined, and/or (2) a program's goals and activities should be examined for possible alignment.

Scenario Two is what most goal-free evaluators can expect; the second scenario depicts a situation where the criteria of merit, as identified by the goal-free evaluator,

Scenario One

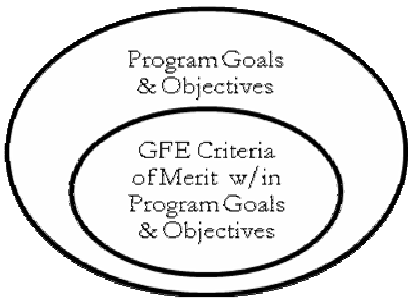


Scenario Two



Scenario Three

Scenario Four



Scenario Five

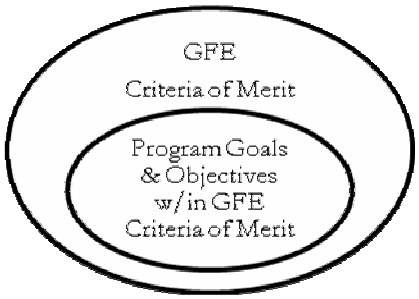


Figure 2. Conceptual Framework of the Outcome Scenarios Between the Goal-Free Evaluator's Criteria and the Program's Stated Goals

overlap with some of the program's goals and objectives. This means that some but not all of the goal-free evaluator's criteria are the same as stated in the program goals and objectives. Possible causal explanations include that this is: (1) a program with "paper" goals without the effort or resources provided to achieve effects on these goals (e.g., the goals are outdated or are in dispute); (2) a program that produces effects on some goals but the effects are too small, fleeting, or trivial to detect; and/or (3) the evaluator fails to detect the goals which means that the evaluator or evaluation approach is deficient.

The third scenario is a hypothetical result where the criteria identified by the goal-free evaluator and the program's goals completely overlap; the goal-free evaluator finds only and all of the program's goals. This is illustrated in Scenario Three with one oval superimposed over the other to create an entirely shaded oval. This scenario is highly theoretical as it represents the ideal program, evaluator, and evaluation or it indicates that bias was introduced into evaluation. If the goal-free evaluator identifies the same criteria—no more, no less—as stated in the upstream stakeholders' goals, this represents (1) a program whose goals are perfectly aligned with its outcomes and an evaluator using

an ideal evaluation model who was able to identify each and every need and goal, and/or (2) an evaluator who was contaminated by the goals or with other forms of bias, hence whose credibility is threatened. The latter scenario is more likely.

The fourth scenario illustrates a situation where the total list of criteria or effects identified by the goal-free evaluator are program goals; however, the evaluator missed some effects or failed to detect some stated goals. Therefore, the oval representing the evaluator's criteria is inside the larger oval representing the program's goals. This illustration shows (1) a program whose goals are well aligned with the consumers' needs and desires, and the program's effects; (2) a program with too many goals or trivial goals and therefore fails to produce detectable effects on some of its stated goals; (3) consumers who do not demonstrate need related to goals stated by the program; and/or (4) a less than competent evaluator who fails to recognize all critical effects, needs, and/or outcomes.

The fifth scenario is the inverse of the fourth. In Scenario Five, the goal-free evaluator uncovers all of the program's goals plus finds other criteria; thus, the oval representing the program's goals and objectives is located inside of the goal-free criteria of merit oval. This scenario represents a program that is producing outcomes on all of its goals as these effects are detected by the goal-free evaluator, yet the program fails to acknowledge other potentially relevant effects. From an internal program evaluation or managerial perspective, the evaluand is likely performing well as effects are uncovered in all goal-related areas; however, according to the goal-free evaluator, the program is missing some outcomes on potentially important criteria and should consider expanding

its goals and programming to cover a broader range of criteria and effects. Of course, an additional possibility is that the evaluator's list of criteria is too broad.

It should be noted that the ovals in Figure 2 are not drawn to scale. For instance, a program has 10 stated goals and the goal-free evaluator identifies 30 criteria. Of the 30 criteria identified by the goal-free evaluator, only 2 criteria were stated by the program as being a goal. If the ovals in Scenario Two were drawn to scale relative to the numbers just described, the oval of *goal-free criteria of merit* should be drawn three times larger overlapping with the program goals by 6% and the much smaller *program goals & objectives* oval overlaps with the goal-free evaluator's criteria by 20%. The scale described above is approximated below in Figure 3.

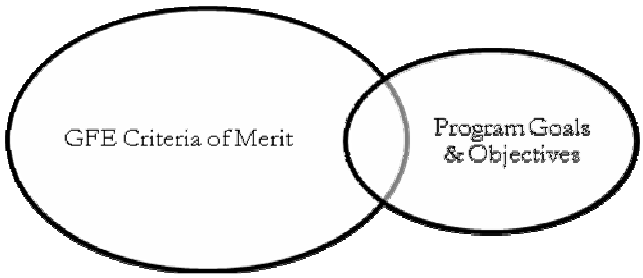


Figure 3. A Scale Example Conceptual Approximation of the Relative Number of Criteria Versus Goals

The relativity of the ovals sizes is applicable in all five scenarios and, obviously, there are three possible scenarios that may exist independently or in combination with one another: (1) the goal-free evaluator finds more criteria than is stated in the program's goals; thus, the criteria oval is larger than the program goals oval like in Figure 3; (2) the program's goals cover more criteria than is identified by the goal-free evaluator, represented by a larger program goal oval and a smaller GFE criteria oval; and/or (3) the

goal-free evaluator identifies the same number of criteria as were stated in the goals and objectives thus the ovals are identical in size. Same-sized ovals does not imply that the evaluator's criteria and the program's goals were aligned or matched; rather, it simply means that there is an equal number of criteria identified by the evaluator as are stated as program goals.

As previously stated, the most probable of these outcome scenarios is Scenario Two. Below in Figure 4 is a simplified example of the second scenario using a hypothetical substance abuse program given to illustrate the goal-free evaluator's criteria and the program's goals. The three stated goals of the substance abuse program are: (1) to reduce substance use, (2) to enhance social skills through team-building activities, and (3) to connect the consumer to appropriate community resources. Imagine that the goal-free evaluator finds three effects: (1) reduction of substance use, (2) connection with financial assistance during consumers' recovery, and (3) positive time and personal attention from program staff members. Thus, the evaluator adopts these as evaluation criteria. Notice that in this example (Figure 4), the actual effects and the stated goals agree only on the first criterion regarding the reduction of substance use, which is represented by the middle overlapping portion.

In summary, there are several possible relationships between the goal-free evaluator's criteria of merit identified during the evaluation, and the preordinate goals and objectives created by the past and present program stakeholders. A couple of these scenarios are more or less feasible or probable than others. The next section examines how the criteria that were identified by the goal-free evaluator can be used to help the program align its goals with its actual outcomes and the needs of its consumers.

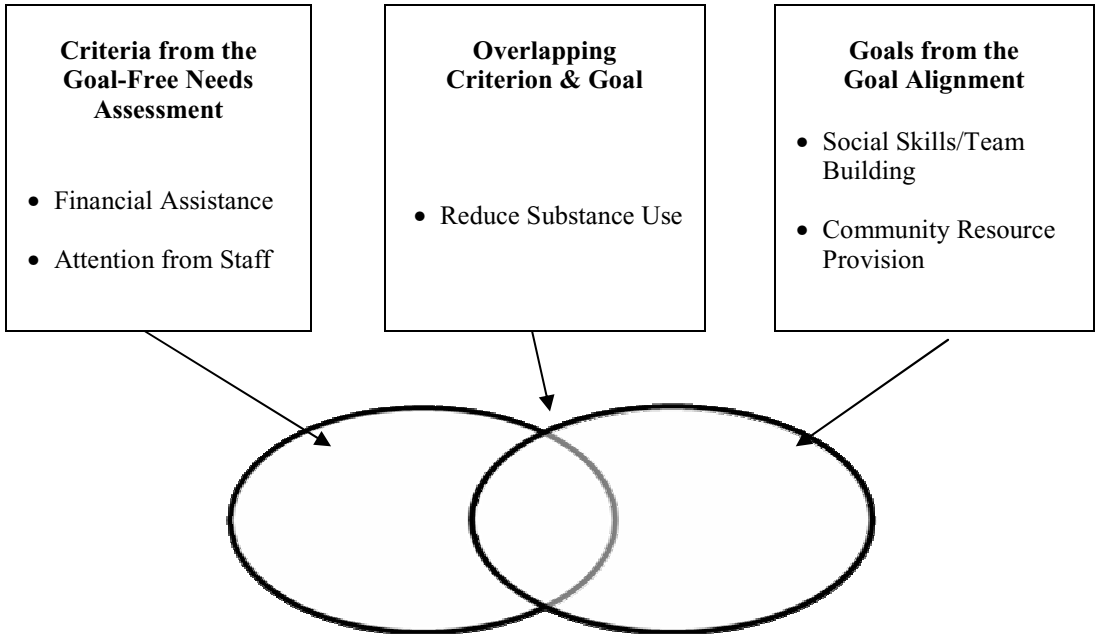


Figure 4. A Hypothetical Substance Abuse Program—The Relationships Between Program Goals and the Goal-Free Evaluator's Criteria

The Goal-Free Evaluator's Criteria as a Tool for Goal Alignment

According to Patton (1997), a "result of goal-free evaluation is a statement of goals … a statement of operating goals becomes its outcome" (p. 182). These criteria become a reference tool for goal alignment as they have potential to become one of the program's official goals or objectives. However, Patton says that Scriven discourages the determination of "true" program goals as an outcome of GFE because Scriven feels that GFE's outcome is the determination of merit with an emphasis on the consumers' relevant needs. Scriven is concerned that changing the focus to goal alignment confuses the goal-free evaluator by again shifting inquiry toward goal-orientation.

However, Scriven (1974b) states, "a crucial function of good formative evaluation is to give the producer a preview of the summative evaluation" (p. 35). If Scriven's

statement is accepted then it seems that offering the upstream stakeholders the criteria used by the goal-free evaluator during the evaluation is not only justified but of potential use to the evaluation users. It seems reasonable that program people use the goal-free evaluator's criteria as goals, goals for basing objectives and outcome measures for future internal evaluations and program monitoring. Besides, through the course of the evaluation, the goal-free evaluator has already developed a list of criteria for judging the evaluand; thus, it seems a relatively effortless task for the goal-free evaluator to adapt these criteria into a format familiar to the evaluation users such as goals and objectives. Patton (1997) claims that Scriven is concerned with the goal-free evaluator losing sight of the focus on merit determination and the meeting of consumers' needs. However, in adherence to GFE's principles of bias control and blinding, the adaptation of the criteria into goals and objectives should occur only after the completion of the data collection and analysis, and typically before the program's stated goals are revealed to the evaluator. Therefore, if the evaluator finds it prudent to offer this service and evaluation stakeholders request it, GFE's process of goal alignment represents a secondary task for the goal-free evaluator (the primary task being the determination of the evaluand's merit). Therefore, the evaluator provides the program people with the following information for goal alignment: (1) the goal-free evaluator's criteria of merit, and (2) the evaluator's adaptation of the criteria into a format for upstream stakeholders. In conclusion, GFE can be useful in aligning a program's goals with its actual activities and performance resulting in the total relevant criteria for judging the evaluand's merit.

Potential Benefits of GFE

There are numerous theoretical benefits of GFE. Bulleted below is a brief listing of six of these potential benefits. Following the bulleted list is a more in-depth description of each.

GFE's benefits are based on:

- Controlling goal orientation-related biases
- Uncovering of side effects
- Avoiding the rhetorical game of "true" goals
- Reducing swamping by trivial objectives
- Adapting to changes in consumers' needs
- Supplementing GBE

As discussed previously in this dissertation, one of the main benefits of GFE is the ability to control evaluation biases related to goal-orientation. Scriven (1991) claims that by reducing interaction with program staff and by screening the evaluator from goals, GFE is less susceptible to social, perceptual, and cognitive biases than GBE. Again, potential biases are introduced by trying to satisfy the evaluation client because it is not explicit in what the client is attempting to do; similarly, it offers fewer opportunities for evaluator bias or corruption because the evaluator is unable to clearly determine ways of cheating (see *general positive bias*, p. 73) (Scriven, 1991). Scriven (1974b) uses the analogy of trial juror who is approached by an interested party and offered a prestigious position or a large sum of money. Even if the juror is not swayed, the mere possibility and suggestion of bias threatens the juror's credibility in yielding an impartial judgment. The judicial system has established protocol for minimizing this bias (i.e., juror

sequestering) and a juror who communicates with a party of interest in violation of these rules and procedures faces repercussions.

GFE is more likely than GBE to identify unintended positive and negative side effects; particularly of interest to upstream stakeholders is the identification of serendipitous outcomes and contextual information (Thiagarajan, 1975). In his analogy between GFE and pharmaceutical studies, Scriven (1974b) justifies searching for side effects stating that "no evaluation of drugs today can avoid the search for side effects from the most remote area of the symptom spectrum" (p. 43). Goal orientation may cloud the evaluator in his/her search for side effects as "the knowledge of preconceived goals and accompanying arguments may turn into a mental corset impeding her [the evaluator] from paying attention to side effects, particularly unanticipated side effects" (Vedung, 1997, p. 59). Therefore, with a GFE, the "negative connotations attached to the discovery of unanticipated effects" is reduced (Patton, 1997, p. 181); terms like side effect, secondary effect, and unanticipated effect become meaningless because the evaluator does not care whether effects are intended or not (Scriven, 1974b). Moreover, when goals are poorly founded, the goal-based evaluator will miss critical effects that may have been detected by the goal-free evaluator. As stated by Fitzpatrick et al. (2004), "It is tragic when all resources go to goal-directed evaluation on a program when the stated goals do not even begin to include all of the important outcomes" (p. 85).

GFE circumvents the difficult rhetorical and often contaminating task in traditional evaluations of trying to identify true current goals and true original goals, and then defining and weighting them. Historically, goals were couched in professional fads, current jargon, or lists of priorities, according to Scriven (1974b), and "the rhetoric of

intent was being used as a substitute for evidence of success" (p. 35). In some cases, the goal setting process instigates a civil war where stakeholders battle for control of the evaluand's direction (Patton, 1997). Besides, Scriven adds:

> There is just no way around the fact that every evaluator has to face those "thousands of possibly relevant variables" and decide which ones to check in order to determine side effects. Having three or four or ten identified for you is scarcely a drop in the bucket. (p. 50)

Another benefit of GFE is that it acknowledges the effect of swamping, a situation where numerous "trivial objectives mask the true intent" (Thiagarajan, 1975, p. 39). In his early writings on GFE, Scriven (1974b) provides a real-world example of swamping while rating numerous products during an evaluation; he writes that one of the products "finished up in the 'Top Ten' in spite of zero results with respect to its intended outcomes because it did so well on an unanticipated effect" (p. 34). It is implied, in Scriven's example, that if the goals were the sole concern of the evaluator, the top 10 finishing products may have finished lower because the results of the intended outcomes would have swamped the positive unintended outcomes.

While GBE is static, GFE can adapt to the sporadic changes in consumer needs, program resources, and program goals (Scriven, 1991). There is little the goal-based evaluator can do when a program's goals change except for start the evaluation over, overhaul the evaluation design and data collection, or create excuses for the irrelevance of the evaluator's evidence. On the other hand, the goal-free evaluator who is not relying on goals and objectives continues his/her inquiry despite changes in program goals and as long as these changes are reflected in the program's actions and outcomes, the goal-free evaluator should recognize and record these effects.

GFE is—by design—capable of supplementing and informing GBE. One way to accomplish this is based on the fact that GFE is reversible. An evaluation may begin goal-free and later become goal-based using the goal-free data for preliminary investigative purposes, and this, according to Stufflebeam and Shinkfield (1985), ensures that the evaluator still examines goal achievement. Therefore, GFE findings can be used as baseline information for a GBE. Another way GFE informs GBE occurs when GFE is used as a complement to GBE. GBE and GFE "can be conducted simultaneously by different evaluators" (Stufflebeam & Shinkfield, 1985, p. 317). Thus, when used as a supplement to GBE, GFE serves as a form of triangulating evaluation approaches, evaluators, data collection methods, and data sources. Lastly, as previously mentioned, GFE produces criteria that can be used for goal alignment.

Usually when evaluators choose GFE, it is because of these potential benefits but not everyone is so convinced. While some evaluation scholars and practitioners are slightly hesitant with regard to using GFE, others are completely skeptical. This section looked at some of the potential benefits offered by GFE, the next considers frequent criticisms of GFE.

Criticisms of GFE and Responses

Several authors and evaluators have presented criticisms of GFE. The purpose of this section is to introduce the criticisms by presenting the arguments for and against GFE, thinking both theoretically and pragmatically. Below are eight criticisms of GFE and responses to these criticisms.

Criticism One

GFE is so independent that it becomes no longer of use to the evaluation's intended user.

A criticism, waged by Stufflebeam (2001), of consumer-oriented evaluation in general is that it "can be so independent from [program] practitioners that it may not assist them to better serve consumers" (p. 60).

Response to Criticism One. In response, Scriven (1991) suggests that if this is a concern regarding GFE specifically, the goal-free evaluator might use GFE as supplement to other goal-based models. Scriven (1974b) states that it is possible for GFE to "improve GBE in certain sites, not replace it" (p. 47), and if this is the case, "I am arguing for GFE as only part of the total evaluation battery" (p. 49). Second, if GFE identifies criteria and outcomes that are "so independent from practitioners," there is likely an issue either of cohesion between the program's mission, goals, objectives, practices, and actual outcomes, or the evaluator potentially lacks evaluation competencies (see Figure 2). Lastly, the evaluator is rarely encouraged to impose an evaluation approach on an evaluation client; ideally, the evaluator should offer several evaluation models and approaches appropriate for collecting the evaluation data and that suit the evaluation users' and program consumers' needs. The purpose of this dissertation, in part, is to address whether the information from GFE is useful for assisting the program's consumers.

Criticism Two

GFE is iconoclastic and requires extremely competent evaluators.

Stufflebeam (2001) calls the consumer-oriented approach "iconoclastic" as he feels that it is "heavily dependent on a highly competent, independent, and 'bulletproof' evaluator" (p. 60). He also refers to the goal-free evaluator—somewhat tongue-in-cheek—as an "enlightened surrogate consumer" (p. 58).

Response to Criticism Two. What quality evaluation does not require a "highly competent" evaluator? Is GFE, by its very nature, more dependent on a quality evaluator than any other advocated approach or model? Each evaluation model and approach has its educational, technical, and experiential prerequisites. The professional evaluator is supposed to be enlightened as compared to program managers, in areas like the logic, theory, methodology, ethics, and practice of evaluation. It should be noted that Stufflebeam's does acknowledge that to claim GFE requires a more competent evaluator is a prescriptive claim and itself a claim worthy of study.

Criticism Three

Goal-free evaluators substitute their goals and values in place of the upstream stakeholders'.

The third criticism is that the goal-free evaluator simply substitutes his/her own personal values into choosing evaluation criteria and goals (Patton, 1997; Stufflebeam, 2001). The implication is that the goal-free evaluator's criteria of merit are created subjectively, i.e., in an idiosyncratic or arbitrary manner. The argument posed by Patton (1997) is as follows: "using needs instead of program goals implies entertaining a

prescriptive instead of a descriptive view of valuing" (p. 62). Restated, Patton believes that the goal-free evaluator imposes his/her values as to what counts as meritous instead of characterizing or describing existing values.

Response to Criticism Three. On the contrary, the evaluator should only use sensible and defensible criteria and values based on what the evaluator actually observes and on justifiable logical and definitional premises. Davidson (2005) offers a response to this criticism, supporting her position that meeting consumers' needs as the source for GFE criteria is not based on her personal opinion but rather commonsense. She writes:

> As for the contention that goal-free evaluation involves applying the evaluator's personal preferences to the program, this would be true only if the evaluation were not being conducted competently. Another term for goal-free evaluation is needs-based evaluation. So, the standards used to determine program quality or value should be mostly the actual documented needs of consumers (along with several other relevant sources of value) and *not* the "personal preferences" of the evaluator. Of course, the evaluator needs to make sure that the sources of values used for the evaluation are valid and defensible ones. But replacing those with the preferences of program staff is not a great solution. (p. 234)

The goal-free evaluator may unknowingly use one or more program goals inadvertently while observing effects or documenting needs, but simply to use the evaluator's own personal preferences would jeopardize any evaluation's legitimacy and credibility. Scriven (1974b) concurs:

> Another commonly connected error is to think that all standards of merit are arbitrary or subjective. There is nothing subjective about the claim that we need a cure for cancer more than a new brand of soap. The fact that some people have the opposite preference (if true) doesn't even weakly undermine the claim about which of these alternatives the nation needs most. So GFE may use needs and not goals, or the goals of the consumer or the funding agency. Which of these is appropriate depends on the case. But in no case is it proper to use anyone's goals as the standard unless they can be shown to be the appropriate one and morally defensible. (p. 38)

Criticism Four

GFE only eliminates the needs of program staff.

The basis of the fourth criticism is that program people and other upstream stakeholders are the only ones whose needs and wants are not considered by the goal-free evaluator. Patton (1997) criticizes Scriven's question of whose goals will be evaluated, and Patton concludes that

> Scriven's goal-free model eliminates only one group from the game: local project staff. He directs data in only one clear direction—away from the stated concerns of the people who run the program. He addresses an external audience, such as legislative funders. But, inasmuch as these audiences are ill defined and lack organization. (p. 182)

Vedung (1997) reiterates that goals "are not haphazard wishes or incidental desires" (p. 61); they are no less arbitrary than any criterion a goal-free evaluator claims relevant. At least with GBE, the evaluator assesses the areas that the evaluand's stakeholders have already determined important. Furthermore, goals and objectives represent the desires of intelligent, experienced, and influential people who are involved with the evaluand and have a vested interest in the program. They are usually created with careful reflection and adapted over time "focusing actions on specific outcomes" (Friedman et al., 2006, p. 202).

Response to Criticism Four. Davidson (2005) offers a statement that serves as a nice response to Criticism Four:

> It is true that evaluations need to be designed and conducted in ways that address the information needs of program staff and other upstream stakeholders. However, the primary reason why any program or project is put into place is to meet the needs of a particular group of potential program recipients. Therefore, their needs and concerns are paramount, whereas those of the program staff are not. A good evaluation will, in any case, meet the information needs of the

> program staff; these may well be different from what the staff's wants and/or concerns might be... (p. 234)

Furthermore, the evaluator may identify and examine some of the program's goals; the goals that the program demonstrates through its actions, outcomes, and impacts. As previously stated, if the goals are important and the program is putting forth significant effort and creating significant outcomes with regard to its goals, the effects should be blatant and the goal-free evaluator should detect them. Last is the reminder that no evaluation approach should be forced on an unwilling evaluation client especially an alternative approach that appears obviously more appropriate for determining evaluand merit and serving the information needs of the evaluation's stakeholders.

Criticism Five

GFE is not really goal-free; rather it simply implements a broader understanding of what it means to be considered a goal and a wider decision audience.

Alkin (1972) made this point soon after Scriven's introduction of GFE; Alkin writes: "by 'goal-free' Scriven simply means that the evaluator is free to choose a wide context of goals … goal-free evaluation is not really goal free at all, but is simply directed at a different and usually wide decision audience" (p. 11). Grinnell, Unrau, and Gabor (2008) also argue that GFE's greatest limitation is that it is "not goal-free at all but rather focuses on wider context goals instead of program-specific goals" (p. 531). Cronholm and Goldkuhl (2003), referring to GFE, write "the involvement of a wide range of stakeholder groups is essential to this approach of evaluation" (p. 3).

Response to Criticism Five. First, most GFEs begin by focusing on wider context goals, yet when a program is doing what it says it does, it should take little time before

the competent evaluator identifies and thus may focus on program-specific goals. Therefore, it is debatable as to whether the use of broader context goals is in fact a limitation. If the non-stated goals are relevant and if evaluation resources and evaluator expertise permit, it seems that the inclusion of broad goals is not a limitation but rather a necessity. Moreover, Grinnell et al. (2008), in their criticism, fail to mention whether focusing on "wider context" rather than "program-specific" goals has merit in its own right. Therefore, saying that a GFE focuses solely on broad contextual goals connotes that looking beyond previously stated goals is undesirable; thus, a search for side effects is also an investigation of the non-stated. Nevertheless, even if one were to accept this criticism, the evaluator is still able to identify and investigate program-specific goals, the evaluator just does not know that the goals have been adopted officially.

Youker (2005b) offers an example of a GFE unintentionally observing program goals where he was a goal-free evaluator for a middle school's summer school program for "at-risk" students and a simultaneous GBE was also conducted on the same summer school program. Based on the observed needs of the middle school students and the actual program effects, Youker identified and then investigated the students' *interest, motivation, and participation in the learning process*. This was very similar to a stated program goal *to instill the desire in students to extend their learning*. The point to this example is that with a GFE, the criteria and values are often developed iteratively and although the initial list of criteria, or potential relevant outcomes, may be broader than the program states, a quality GFE examining the program's ability to meet the relevant needs of its consumer and the actual effects produced may investigate outcomes related to a stated goal without knowing that it is a goal.

Second, and more critical, this is a forest for the trees issue. Because we call it a dog and it is really a car does not diminish the car's nature, quality, or function as a car. In fact, GFE's possible misnomer was discussed previously in this chapter under the definition for *goal*. The knowledge of the overall purpose of the evaluation does not constrain GFE as an approach, which is illustrated below in the following two examples, the first from product evaluation and second from an educational program evaluation. Consumers' groups do not appear hindered when they publish their reports on automobile quality and suggest best buys even though the group is well aware of the general overarching purposes of a vehicle when it evaluates them. The aim is always the same, something like: to design a powered wheeled passenger vehicle that carries its own motor. Furthermore, the consumers' groups know that most automobiles move both backward and forward when prompted, turn in either direction when the driver rotates a steering wheel, and stops when the operator applies the breaks. However, the consumers' groups do not know, and often do not care too much about knowing, the specific aims of the car manufacturer and engineers. Manufacturers, designers, and engineers have specific goals and objectives such as producing a car that meets the particular wants and needs of average American family; one that appeals to the senses of youthful car buyers; a high-end car that offers the cutting edge in comfort, technology, and luxury; or a vehicle with high reliability, low maintenance cost, and affordable retail price. Likewise, the overarching goals of educational programs are knowledge acquisition and application; nevertheless, an evaluator does not need prior knowledge of the specific ways and means (i.e., objectives) the educator and educational program intend to foster this in its students. Whether the educators' objectives are to use hands-on learning techniques, incorporate

significant portions of visual and audio learning content, use only original texts, provide content through traditional methods and lecture, maintain strictly disciplined classrooms, and so forth, the evaluator collects data on the performances of the students, the educators, and program in determining whether the students demonstrate the acquisition and application of knowledge and whether these results are of merit, worthwhile, or significant. In summary, GFE is intentionally free from the specific-goals and objectives; yet goal-free evaluators are free and often able to infer the broad-scale aims of the evaluand, which does not significantly inhibit the evaluator from collecting relevant and useful data regarding performance outcomes, impacts, and merit.

Criticism Six

It is "methodologically much more difficult to elicit needs than to map results and let recipients do the valuing" (Patton, 1997, p. 62).

The gist of the preceding quote by Patton is that the rhetoric of needs is more difficult to specify and then observe than that of goals. Needs are too hard to determine, so instead the evaluator should skip any needs assessment in favor of reporting all evaluation results and outcomes and letting the evaluation users decide on whether their consumers' needs were met. The view that the needs assessments is unnecessary in GFE is echoed by Coryn (personal communication, July 28, 2008) as he advocates the determination of actual effects, offering the evaluation client a positive and negative outcomes profile, and leaving the determination of overall merit to the evaluation users.

Response to Criticism Six. Scriven might respond to this criticism by referencing his distinction in immediacy between what the country needs, better soap, or a cure for

cancer (Scriven, 1974b); however, the above criticism is legitimate especially given limited evaluation resources. However, if no evaluative conclusion is reached it becomes questionable as to GFE's status as an evaluation approach rather than a form of social science inquiry. In the real world, a description of all the evaluation findings could easily be scores of pages and in attempting to make an evaluation useful and a report succinct, the evaluator synthesizes this information. Synthesis, or the process of combining factual and value premises into one or more evaluative conclusions, is also one of Fournier's (1995) four stages describing the logic of evaluation and thus, for Fournier, it is fundamental in calling an inquiry of this type an evaluation.

Criticism Seven

> *"Under a pure goal-free approach, program staff need only wait until the goal-free evaluator determines what the program has accomplished and then proclaim those accomplishments as their original goals" (Patton, 1997, p. 193).*

To restate this criticism, once the GFE is complete and the evaluation users have received the report, the program administration and staff pat themselves on the back and say that the GFE's findings were their goals all along.

Response to Criticism Seven. This criticism assumes that (1) the program has subversive intentions; (2) the goals were not stated, documented, or conveyed prior to the GFE; and (3) the program has not disseminated the results of any prior internal or external evaluation. Furthermore, Patton's criticism neglects to consider the case of GFE being employed as a complementary evaluation approach as the GBE approach holds the program accountable for their original goals. If this unfortunate situation does occur, it

does not seem reasonable to scrap GFE in its entirety because of how people abuse it, essentially *throwing the baby out with the bath water*.

Criticism Eight

GFE excludes goals adopted by elected politicians who posses a special status as they have procedures established for decision making and are representing the interests of their constituents (Friedman, Rothman, & Withers, 2006).

The goals adopted by elected officials are different than the goals from program funders, administrators, and staff. First, in a representative democracy, the politician is chosen by the populace for his/her views, values, judgments, etc. Once elected, the politician *is* the voice of the people; thus, the politician's goals are of the utmost importance. The central issue boils down to being an issue of morality. The elected representative is chosen by the majority. In a representative democracy, no idea can prevail without the support of the majority (Thoreau, 1849); this morality is dictated by the socio-political *zeitgeist* of the majority. However, the question remains: are politicians elected for their judgment in making decisions based on their conscience or are they elected to represent the views of their constituents? In other words, is the politician elected to do what s/he ultimately believes is right, or is what is right the reflection of the constituents' opinions? To illustrate, this dilemma was central in the U.S. 2008 Democratic presidential nomination between Barack Obama and Hillary Clinton as the nomination came down to superdelegates. Should a Democratic superdelegate vote for whom s/he believes to be the best option, or should the superdelegate's sole concern be reflecting the majority position of his/her constituency?

Response to Criticism Eight. The goal-free evaluator might be inclined to answer these questions by saying that the representatives should vote their conscience. Thoreau (1849), in "Civil Disobedience," claims that morality is established by the current majority; however, there are countless examples of changing and shifting morality of the majority (e.g., from pro- to anti-slavery or pro-child marriage to anti-, pro-corporal punishment to anti-, etc.). Does that mean any progressive individual on the front end of a cultural shift in morality is morally incorrect for being anti-slavery before the majority has fully transitioned to agreement? This begs the question of the existence of moral absolutes.

Are there moral absolutes? Is slavery always wrong? For some pragmatic goal-free evaluators, there are certain principles and values that serve as close approximations to moral absolutes. For example, in any program, the needs of the consumers are first and foremost; and all programs have an ethical obligation to prevent known potential physical and psychosocial harm for all involved with the program, upstream and downstream. The evaluator's morals are founded on principle based in logic, ethics, resource efficiency, and the needs of the consumer, among others.

In answering these moral questions, the goal-free evaluator turns to the *cui bono* principle which concludes in the belief that the ultimate benefits should befall the program consumer. In the description of the superdelegates in the Democratic nomination process, many superdelegates and pundits alike were split. Several politicians voted their conscience going against the view of their constituency. Furthermore, the role of the politician includes mediator and compromiser; therefore, some of the theoretical best interests of the populace may not be represented in the politician's goals since they were

objects of compromise. Lastly, there are many examples of politicians tainted by corruption and personal gain attempting to further their own interests. In summary, there are numerous factors that may influence the elected official in determining goals and objectives, and because of these influences, the evaluator is justified in searching for outcomes beyond the politician's goals and objectives.

In conclusion, the preceding criticisms are frequently encountered discussions of GFE. Some of these criticisms are logical yet lack an empirical basis. Despite the concerns of these critics, none of them go so far as to suggest that GFE is useless and that it should be stricken from the evaluator's toolbox. Rather, the criticisms are concerned with whether GFE provides any real benefit for the evaluation users; and, if so, what is the nature of the differences and what explains those differences? The objectives of this dissertation were designed to examine these claims and criticisms.

Chapter Summary

The second chapter examines the academic and professional evaluation literature with regard to GFE. The literature review presents the history of GFE as well as its logic. This chapter also contains an identification and articulation of GFE's potential benefits and presents GFE's main criticisms. The next chapter describes the methodological specifics of the study.

CHAPTER III

METHODOLOGY

In this chapter the methods used to accomplish the study's objectives that were described in Chapter I are presented, specifically, the methodological approach used to assess GAE's and GFE's differences on evaluation utility from the perspective of the evaluations' intended users. Topics covered in this chapter include the study's approach and design; selection and characteristics of evaluand, evaluation users, and evaluators; study setting; materials created by the evaluand, investigator, evaluators; instrumentation; data collection and recording; and data processing and analysis. The chapter concludes by identifying some of the study's methodological limitations.

Description of the Approach

In this dissertation study, the investigator attempted to create an experimental setting analogous to actual GAE and GFE practice and then ascertain each approach's utility as determined by actual evaluation users. An analog study is an inquiry which resembles a different situation, (e.g., the conditions of a "real" evaluation) and the design in this analog study uses a real human service program, real program outcomes, and real evaluation outcomes.

As described in Chapter I, three specific objectives were investigated in this dissertation. They are as follows:

1. From the perspective of evaluation users, is there a difference between GAE and GFE with regard to utility?
2. What, if any, are users' perceived differences in utility between GAE and GFE? If differences do exist, how do they differ specifically in terms of instrumental use, conceptual use, and persuasive use?
3. If differences in perceived utility exist, what explains those differences?

Research Design

The primary methods used to investigate GAE and GFE utility are two independent one-group posttest-only designs and semi-structured telephone interviews. This design, therefore, consists of two independent variables (i.e., GAE and GFE) contrasted against the same dependent variable (i.e., evaluation utility). Two evaluation teams, one trained in GAE and the other in GFE, worked independently and simultaneously to evaluate the same entity, a human service program. Following the evaluation, each team produced its own evaluation report to present the evaluation findings. After selected evaluation users read each evaluation report they were asked for their perceptions regarding the utility of the report's findings first questionnaires and then to an interview. Therefore, this study consists of a treatment (i.e., GAE) and an observation of utility in addition to a simultaneous yet separate treatment (i.e., GFE) and observation of utility. Propositionally, the null hypothesis is that there is no practically significant difference in utility between GAE and GFE. Notationally, these hypotheses are expressed as:

H_0 : GAE = GFE

H_1 : GAE ≠ GFE

The study is also a mixed-method investigation in that three independent methods are used to investigate the primary research questions. One argument for mixed-methods research is that by using more than one method, the biases of individual methods are reduced. The three methods used to investigate the primary research questions are:

1. Semantic differential rating scales asking evaluation users to rate perceived differences in utility between GAE and GFE
2. Semi-structured interviews with evaluation users to investigate perceived differences between GAE and GFE in terms of instrumental use, conceptual use, and persuasive use
3. Content analysis of the GAE and GFE evaluation reports

Each of the three methods used to investigate the proposed research questions are described in greater detail throughout the course of this chapter.

Subject Selection and Characteristics

Three sets of subjects were necessary to conduct this analog study: (1) an evaluand, (2) evaluation users, and (3) evaluators. Each of these groups is described in the following sections.

Evaluand Selection and Characteristics

The evaluand, a human service program, initially began as a cooperative program among three organizations operating in a county in southwestern Michigan: (1) Agency X, (2) Agency Y, and (3) Agency Z. The program was selected via convenience sampling

in that the program had a preexisting relationship with one current and one former Evaluation Center evaluator and an ongoing request for external evaluation with the Center or an affiliate. The director of the Interdisciplinary Ph.D. in Evaluation (IDPE) program, which is housed in the Evaluation Center, notified the student investigator of the prospect of conducting a study of this particular program.

Two historical characteristics of this program lent themselves to a field-based investigation of GAE and GFE: (1) the program's maturity, and (2) the program's prior relationships with the Evaluation Center-affiliated evaluators. First, the program is a fairly established program as it has been operating since 2001, with its most recent external evaluation completed in 2006. Previous internal and external evaluation efforts have investigated the program's outcomes according to its stated goals and objectives; therefore, program administrators were willing to examine a potentially broader range of criteria and outcomes for this evaluation. Second, the program has a three-year history contracting with individuals affiliated with the Evaluation Center and IDPE. Both the evaluators and the program administrators reported positive experiences working with each other; thus, to a degree, a relationship and rapport were already established. The combination of the ongoing rapport with the Evaluation Center-affiliated evaluators along with the maturity of the program likely influenced the program administrators regard their evaluation preparedness, their willingness to be involved with this study, as well as their willingness to try GFE, a lesser known approach to evaluation.

The stated overarching goals of the program are to provide housing stabilization, and employment retention and job development services while reducing dependence on public assistance for persons moving from welfare to work. According to Agency X's

website, the program intends to serve clients ranging from the "newly homeless or precariously housed to the chronically homeless and unemployed." As stated by one of the evaluators who conducted a previous program evaluation, "The program administrators consider the proximal goal of the program to be reducing chronic homelessness and unemployment, with the distal goal (or mission) of reducing dependence on public assistance" (Coryn, personal communication, March 20, 2008). According to the program, the specific services it offers its consumers include (1) supportive services to assist in avoiding housing loss or to assist homeless households in obtaining replacement housing, (2) housing crisis resolution action plans to address housing needs and barriers, and (3) subsidies available to participants based on their initiative in taking action toward stabilizing their housing situation. The program claims that it also offers services related to (1) financial/household management, (2) employment, (3) education and job training, (4) transportation, (5) childcare, and (6) interagency referrals and collaboration, among others.

The program theory underlying the program's efforts is found in the following statement extracted from Agency Z's website:

> The ___ Program attempts to dissolve barriers between the stand-alone housing and employment "silos." Given, an isolated service delivery system can never garner the duplicate mainstream resources required to alleviate poverty and its debilitating symptoms such as homelessness. [The program] is a wrap-around service delivery model clearly demonstrating the interrelatedness of stable housing to stable employment, and vice versa.
>
> [The program] focuses on bridging gaps in mainstream programming contributing to chronic unemployment and homelessness. Many programs and services regularly operate in isolation from one another creating layers of conflicting requirements. Often unwittingly penalizing persons in need as they strive to navigate multiple systems thereby limiting positive outcomes.

Evaluation User Selection and Characteristics

The intended evaluation users are those responsible for the program and in applying evaluation findings. As Patton (2002b) states, "The primary intended users are people who have a direct, identifiable stake in the evaluation" (p. 2). Program evaluation users were selected primarily via convenience sampling, criterion referenced sampling (i.e., purposive sampling), and snowball sampling. The selection was convenient in that potential users were chosen based on their accessibility and willingness to participate, as well as their availability in completing both utility questionnaires and interviews. The selection of prospective evaluation users was criterion referenced in that certain characteristics were of particular importance, namely their position of authority (i.e., the power to give orders, make decisions, and make judgments) and influence (i.e., the power to affect persons, things, or events) within or over the program and in applying the findings from an evaluation. Finally, evaluation users were also selected via snowball sampling as selected key program personnel identified additional potential evaluation users to be included in the sample.

Agency X's ___ Director is a program administrator and served as the contact between the investigators, evaluation teams, and the program staff. This key stakeholder attended a pre-study meeting with the investigators to discuss the details and logistics of the study. Via email, the ___ Director distributed a questionnaire and instructions to all the administrators and directors of the other participating agencies asking the program administrators to identify other program staff who have authority and influence over the program and its evaluation (i.e., snowball sampling). Fifteen evaluation users were identified. However, in the summer of 2009, the student investigator learned that two

program staff from Agency Y left the agency without and would not be replaced, and, furthermore, Agency Z opted not to bid the contract, thus leaving nine remaining individual evaluation users.

Evaluator Selection and Characteristics

The goal achievement and goal-free evaluators were responsible for conducting the program evaluations. Evaluators were selected from a pool of doctoral students from two distinct and well respected Ph.D. programs in evaluation at Western Michigan University (see Academic Analytics, 2008): (1) IDPE, and (2) Evaluation, Measurement and Research (EMR). Potential student-evaluators were recruited via email and by classroom announcements from the IDPE program director; the student investigator also visited two different evaluation classes to recruit. The communiqués explained that the student and principal investigators[18] were recruiting doctoral students in evaluation to conduct an evaluation and prior to this evaluation, a four-hour training would be provided. The potential evaluators were also told that they could spend a few hours per week for up to approximately 24 weeks to complete the evaluation. For their efforts, the student-evaluators would receive graduate-level evaluation field experience between three to six credit hours per semester. Evaluation supervision would be provided by the student investigator and IDPE program director. It was also explained that two small teams would be involved as this was a comparative study of evaluation methodologies. It was emphasized, during the recruitment, that it was imperative to abstain from

[18] For clarification, the student investigator refers to the author of this dissertation while the principal investigator refers to the individual who is both IDPE program director and chair of this dissertation committee. In addition, when this dissertation refers to "the investigator," it refers to the student-investigator.

interactions between opposing team members especially in relation to this study as it would jeopardize the study's integrity.

The selection of evaluators was based on non-probability sampling methods such as convenience sampling and criterion-referenced sampling. The sample was convenient in that the student-evaluators were chosen based on their willingness to participate, their availability to attend trainings, their ability to commit to the study's timeframe, and their willingness to adhere to the study's requirements. The evaluators were selected for academic competence in evaluation[19] and therefore, the selection was criterion-referenced as each prospective evaluator was required to submit an academic transcript to the IDPE program director. These materials, in addition to the approval of the IDPE program director, were used to discern adequate educational background and competence in evaluation. The only exclusionary criterion for student-evaluators was if they did not meet one of the requirements below.

The specific criteria for selecting the evaluators were as follows:

1. Academic Standing: Prospective evaluators must be in good standing (i.e., GPA of 3.0 or above) in a doctoral program in evaluation as stated by the program director after examining the students' transcripts and holistically assessing student preparedness for conducting the evaluations.
2. Course Completion Minimums: Prospective evaluators must have successfully completed (i.e., GPA of 3.0 or above) the following graduate-level research

---

[19] It should be noted that the educational level of IDPE students may be relatively representative of the larger population of program evaluators. For instance, the American Evaluation Association's (AEA) résumé search for evaluators in the U.S. finds that there are roughly the same proportion of evaluators who listed their highest degree as a master's degree as those with a doctorate, and even a few with only a bachelor's degree (AEA's *résumé search* retrieved July 11, 2008 from http://www.eval.org/career _center/resumes/resume_found.asp?where=US&sort=years).

and evaluation courses (or their equivalent) as stated in the academic transcripts verified by the IDPE program director:

a. Foundations of Evaluation
    - EVAL 6000 Foundations of Evaluation

b. Advanced Evaluation
    - EVAL 6010 Advanced Seminar in Evaluation

c. Basic Research Design
    - EMR 645 Elementary Statistics

d. Advanced Specialized Research and Analysis
    - EMR 655 Research Design

3. Background Evaluation Literature: Prospective evaluators must have read and studied both the "Key Evaluation Checklist" by Michael Scriven (2007) and *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation* by E. Jane Davidson (2005).
4. Time Commitment: Prospective evaluators must be able to commit a few hours per for 24 weeks to the task; this includes attending a one-day (four-hour) training at the Evaluation Center and biweekly debriefings with the student and principal investigator.

Six students both showed interest in participating and met eligibility requirements and thus were confirmed by the IDPE program director as suitable for the study. However, had there been more than six potential-student evaluators that meet the criteria for inclusion, the investigator planned to randomly selected six for the study. Once it was confirmed that the prospective evaluators met all other eligibility criteria for participating in the study, they were randomly assigned to teams, three evaluators to the GAE team and three to the GFE team. The assignment of evaluators to evaluation teams was

reviewed and approved by the IDPE program director, and according to his holistic judgment of each student-evaluator, the teams were deemed approximately equivalent (see Table 3 below). The principal investigator assigned each team a team leader; the appointments were based on the principal investigator's knowledge of and experience with these students in his courses. The evaluation teams were instructed that they had approximately six months (i.e., February 2009 to July 2009) to complete their evaluations and submit their reports.

At the beginning of the evaluators' trainings, the investigator collected the following self-reported demographic information from the goal-based and goal-free evaluators: age, gender, years of research and/or evaluation experience, and their perception of their evaluation experience. The mean age of the student-evaluators was 41.5 years (*SD* 9.3). The GAE team's average age was 41.3 (*SD* 13.0), while the GFE team averaged 41.7 years old (*SD* 6.6). Each had a team of three, and the goal achievement team had two male evaluators, while the goal-free team had two female evaluators. The average number of years of research experience that was reported for all evaluators was six years (*SD* 5.2). The GAE team reported an average of two years (*SD* 2.0) while the GFE team reported an average of 10 years of research experience (*SD* 4.0). Combined, the evaluators averaged 3.8 years of evaluation experience (*SD* 4.0). The goal achievement team reported 2.5 years of evaluation experience (*SD* 0.8) and the GFE team self-reported 6.7 years of evaluation experience (*SD* 3.8). Two of the GAE team members claimed minimal evaluation-specific experience, while one self-reported moderate experience. One person on the GAE team reported moderate experience, the other two reported having minimal evaluation-specific experience, while all three goal-free

evaluators reported that they considered themselves to have moderate evaluation-specific experience.

Table 3

Evaluator Demographics

	Evaluator	Age (Years)	Gender (Male/Female)	Research Experience (Years)	Evaluation Experience (Years)	Evaluation Experience Rating
GAE Team	GAE-1	54	M	0	0	Minimal
	GAE-2	28	F	4	1.5	Minimal
	GAE-3	42	M	2	1	Moderate
GFE Team	GFE-1	49	F	6	4	Moderate
	GFE-2	36	M	10	5	Moderate
	GFE-3	40	F	14	11	Moderate

Study Setting

The two settings used in this study were the program's locations and the Evaluation Center. The questionnaires given to evaluation users were administered onsite at the agencies' three locations in southwestern Michigan, where the program is housed and where the evaluation users can be found. The agencies' locations are also where the GAE and GFE were conducted as it is where the program staff and participants can be found. The Evaluation Center in Kalamazoo, Michigan, was the second setting for the study as the evaluator trainings, debriefings, and team meetings were held there. The Evaluation Center is a research and development center with a mission to advance the theory, practice, and utilization of evaluation; the IDPE program is housed within the Evaluation Center.

Instrumentation and Materials

Numerous materials were used in conducting this study. The various materials were categorized based on their source. All materials were created, produced, or possessed by three distinct parties: (1) the evaluand, (2) the student investigator, and (3) the evaluators.

Evaluand-Created Materials

The first grouping of materials includes the brochures, documents, and records collected from the program. In the initial pre-study meeting with a key program administrator, the investigator requested that the evaluand provide preexisting program materials to offer to the evaluators as background and contextual information. The investigator requested that these materials be sent to the investigator, not the evaluators, so they could be screened for goal-related information before disseminating them to the GFE team.

It should be noted that not all of these materials were actually "created" by the evaluand; for example, completed external evaluation reports were written by previous external evaluators. However, they are the property of the program and for the purposes of this dissertation are included as evaluand-created materials.

Investigator-Created Materials

The investigator is the source of the second category of materials used in this study. These are the materials considered necessary for conducting the study. Further descriptions of these materials can be found in the *Instrumentation* and *Procedures*

section later in this chapter. Below is a description of the primary instruments used for data collection and other forms and documents created by the student investigator.

- *Introduction to the Study: A Handout* is a basic description of the study and information for program administrators, managers, and staff. A hardcopy and emailed version were distributed to a key administrator to be disseminated during a meeting between the program and the study's investigators.
- *Email to Prospective Student-Evaluators* is an evaluator recruitment letter describing the study, the evaluator's role, evaluator eligibility requirements, the benefits and risks of participation, and how to contact the study's investigators.
- *Identification of Evaluation Users* questionnaire is the instrument used in snowball sampling and that was given to key evaluation users asking them to identify other evaluation users with authority and influence. It was sent to and received from program administrators via email.
- *GAE & GFE Evaluator Training Curriculum Handbooks* are the guidebooks that were used for training the evaluation teams in their respective approaches. The handbooks, which included the evaluators' logs (see below) and screened materials and documents, were distributed during the evaluators' training.

- *Evaluator's Time Log* is a form for evaluators to track time and activities spent on evaluation related tasks.[20]
- *Evaluator's Communication Log* is a form for evaluators to record communication between the evaluation team and the program's stakeholders.
- *Evaluation Team's Log of Potential Threats to the Goal-Free Nature* is a form for evaluators to record threats to the fidelity of the goal-free approach.
- *Evaluation Team's Log of Potential Threats to the Goal Achievement Nature* is a form for evaluators to record threats to the fidelity of the goal achievement approach.
- *Evaluator Demographic Questionnaire* is the instrument used to collect demographic information on evaluators. It was administered and collected by the student investigator during the evaluators' training.
- *Evaluator Informed Consent Form* outlines the study, the participants' roles, and confidentiality; this form culminates by asking participants to sign indicating whether they understand the risks and benefits of participation and, if so, whether they choose to be a volunteer evaluator.
- *Evaluator Contract* is the form signed by all evaluators stating that they will reasonably and ethically attempt to maintain fidelity to the assigned evaluation approach including avoiding communication with evaluators from the other team.

---

[20] Due to the fact that the evaluators were also acquiring field experience credits from the IDPE program director, there was an incentive to over-estimate or fabricate time spent on evaluation activities. To reduce the extrinsic motivation for inaccuracy, the time logs maintained by each individual evaluator were collected exclusively by the student investigator and only after the student-evaluators received their pass-fail grade from the program director were their times reported to the IDPE program director.

- *Evaluation Utility Questionnaire* is a self-administered survey given to the evaluation users. It primarily consists of a series of semantic differential rating scales designed to ascertain the evaluation users' attitudes with regard to the utility of each team's evaluation report.
- *Interview Protocol* is a description of the investigator's process and questions for interviewing the evaluation users.
- *Methodological Comparison of the Reports* is the form used during the investigator's content analysis of both of the teams' evaluation reports.
- *Approach Fidelity Checklist* was created as a guide for each evaluation team and the student investigator. The relevant portions of this checklist were included in the evaluators' training handbooks to aid them throughout the evaluations. The checklist was also used by the student investigator while supervising the evaluation teams.

Evaluator-Created Materials

The third category of materials is evaluator-created, referring to the documents and materials that were designed by either the GAE or GFE team. The primary product made by the evaluators is the full-length final evaluation report. To provide a relatively consistent evaluation and reporting format, the following guidelines on headings and number of (single-spaced) pages were to be approximated in the evaluation report:

- Executive Summary: 2 pages
- Introduction: 3-5 pages
- Methodology: 5-10 pages

- Findings: 5-10 pages
- Conclusion and Recommendations: 3-5 pages
- Appendices: No limits on page numbers or content

Each evaluation team developed, or produced, specific materials in the process of evaluating the program. Evaluators typically design materials during the process of gathering performance data on the evaluand and its consumers, documenting evaluation efforts, and justifying evaluation-related decisions and conclusions. This type of information's inclusion was mandatory in each team's full-length report, in either the body of the text or in the appendices. In particular, each evaluation team was required to document the evaluation processes and decisions by providing information on the following, as they applied to the tools, instruments, and/or procedures used for: deciding on criteria of merit, determining standards (and grading rubrics), determining or weighting importance, measuring and observing the evaluand's and/or consumers' performance outcomes, and synthesizing criteria (and subcriteria) into a one or more conclusions.

Instruments

There were several instruments created for this study (see *Investigator-Created Materials* above). In the initial stages of the study, the investigator conducted a literature review to begin operationally defining evaluation utility and developing *the Evaluation Utility Questionnaire, a self-administere*d post-evaluation utility instrument for collecting the evaluation users' attitudes regarding the utility of each evaluation report. Additionally, the investigator created a protocol for interviewing all of the evaluation

users following the dissemination, completion, and collection of the utility questionnaire. The investigator also developed a checklist for each evaluation team intended to guide each team and ensure that they maintain fidelity to the designated evaluation approach; lastly, an instrument was created for analyzing the two evaluation reports.

Utility Measures

Triangulation of methods is intended to reduce errors that may be inherent in any one method. There were three primary methods employed in this study: (1) survey research (i.e., the *Evaluation Utility Questionnaire*); (2) semi-structured telephone interviews with evaluation users; and (3) content analysis of the evaluation reports, evaluator logs, and approach fidelity. Below, each is described in further detail.

Evaluation Utility Questionnaire

To obtain the evaluation users' perspective of, or attitude regarding, the evaluation reports' utility, the investigator developed the *Evaluation Utility Questionnaire,* which primarily consisted of semantic differential rating scales. These rating scales are commonly used by attitude researchers in measuring the connotative meaning of objects, events and concepts (Osgood et al., 1957). The scales consist of bipolar adjective pairs, e.g., useful-useless, good-bad, careful-careless. In this case, the connotations are used to determine the evaluation user's attitude regarding the utility of each evaluation report. This is justified in part by Evers (1980), who previously demonstrated that "written evaluation reports met the assumptions presented by Osgood for selection of a concept to be rated with the semantic differential" (p. 50). However, according to Himmelfarb (1993), a limitation with the semantic differential scale is that

the properties of the level of measurement are unknown. Statistically, an approach is to treat it as an ordinal scale. However in this dissertation it is treated as an interval scale arguing that the neutral response serves as an arbitrary zero and intervals between the scale values are equal.

In creating the questionnaire, an original list of over 80 adjective pairs was developed and then reduced to 25 pairs. The respondents (i.e., evaluation users) were asked to complete the 25 adjective pairs and then to complete an open-ended question asking them to explain why they felt the evaluation report was or was not useful.

To counter potential order effects randomization was employed. The order effect refers to perceptual differences arising from serial order in which the measurements are taken or differences in positions in a list. In other words, it is "the influence that one set of questions (or answer categories) may have on the answers respondents provide to later sets of questions" (Bourque & Fielder, 2003, p. 242). Furthermore, random assignment helped controlled for testing effects such as potential respondent survey fatigue and satisficing (Krosnick & Alwin, 1987).

Three versions of the same questionnaire were developed and administered at random. The evaluation users were randomly assigned one of the three versions of the utility questionnaire of which to respond. Any one version of the questionnaire contained the exact same content as the other two versions the difference among them was the order in which the adjective pairs were presented to the respondent as she proceeded through the questionnaire. A random number generator was used to assign a position to an adjective pair within a set of scales. Although randomly assigning the positive and negative directions of the adjective pairs may reduce the potential for response sets, it

was decided that the questionnaire was already of a suitable length and demanding enough of the respondents' concentration, thus all of the adjective pairs read positive on the left to negative on the right. To further reduce the order effect, the evaluation users were randomly assigned which report, GAE or GFE, they would receive and respond to first. Once the evaluation users completed and returned the first *Evaluation Utility Questionnaire*, they were given the other team's evaluation report and the utility questionnaire. This assignment process also dictated the order in which the two evaluation approaches were introduced during the evaluation users' interviews.

Pilot Studies/Pretesting

A non-probability sample of professional evaluators and evaluation students who were affiliated with the Evaluation Center and/or IDPE were asked to pilot-test the *Evaluation Utility Questionnaire*. The selection of pilot-test respondents was convenient as respondents were students who attended a graduate course in metaevaluation on a given day. The previous week, the professor of the metaevaluation course assigned these students a 24-page online evaluation report[21] to read and to discuss at the next class period. Prior to distributing the pilot questionnaire, the students were given a brief background on this dissertation study. To simulate actual conditions while completing the pilot questionnaire, the respondents were asked to pretend that they were upstream stakeholders of the program described in the online evaluation report that they read. Following the administration of the questionnaire, the respondents gave their feedback

---

[21] Committee on Institutional Cooperation. (2008, October). CIC Summer study in Mexico program evaluation report. Retrieved February 14, 2009, from http://www.cic.net/Libraries/ProgramEvals/Guanajuato2008.sflb.ashx

regarding the questionnaire and its instructions mentioning such topics as the length, clarity, and ease of use of the instrument as well as offered other comments, criticisms, and recommendations.

Including the demographic data, the draft *Evaluation Utility Questionnaire* was four single-sided pages with three sets of rating scales per questionnaire. With the addition of the questionnaire instructions, the total survey packet was five pages. Each pilot questionnaire consisted of the three sets of scales of 28 adjective pairs per set. The rationale behind including three sets of scales per questionnaire was to collect data on all three versions of the sets of scales from each respondent. To assess whether fatigue or a testing effect influenced pilot-test respondents, the respondents recorded their times and reported how long it took them to complete each set of scales. In designing the format of the three questionnaires, an assumption was held that if respondent fatigue were to occur, it should be recognizable in the time it takes the respondent to progress through each of the three sets of scales. Presumably, more time is spent on earlier sets of scales and progressively less on latter sets; since the order of the adjective pairs within each set was randomized, there should not be a testing effect among versions of the rating scale as the respondent is not able to "learn" the order that the adjective pairs are presented in the sets.

There were 10 students who participated in the pilot-test; however, 2 students did not complete one of the sets of scales on the questionnaire and therefore 28 sets of scales were completed and timed. Overall, the average time it took to complete each set of scales (i.e., 28 adjective pairs) was 2.1 minutes (SD = .99), both the median and mode

were 2 minutes. Table 4 below presents the central tendency for pilot-test respondents' times on each of the three sets of scales.

Table 4

Central Tendency During Pilot-Testing of Questionnaire

	Set 1 (n = 9)	Set 2 (n = 10)	Set 3 (n = 9)
Mean	2.0	2.5	1.8
Mean in Minutes and Seconds	2 m	2 m 30 s	1 m 48 s
Standard Deviation	0.9	1.0	1.1
Median in Minutes	2	2	1
Mode in Minutes	3	2	1

This assessment for the existence of a fatigue effect was inconclusive. The least amount of time was spent by the respondents on the third set of scales which was anticipated; however, the most time was spent on second set of scales, in excess of a minute longer. The result is somewhat unexpected as the times on the various sets did not progress in a linear fashion as the first set of scales did not take the respondents the longest complete. A possible explanation might be that the pilot-test respondents spent time assessing whether the adjective pairs and/or the instructions were the same or different among the set of scales. However, the randomization of the order of adjective pairs in the three versions of the sets of scales were employed to ensure that should survey fatigue occur, it does not systematically influence the responses to specific adjective pairs. Moreover, since the order of the adjective pairs within each scale was randomized, there should not be a testing effect between the two administrations of the questionnaire (i.e., GAE and GFE) as the respondent is less likely to "learn" the order that the adjective pairs are presented in the scales.

Pilot-testing the *Evaluation Utility Questionnaire* led to changes in the final questionnaire. For instance, as a result of the pilot-test, three adjective pairs were eliminated from the final version of the questionnaire; in addition, portions of the pre-questionnaire instructions were eliminated, edited, and condensed. The final version of questionnaire was not retested after pilot-test revisions as it was assumed to be adequate for the purposes of this investigation. The finalized questionnaire can be found in Appendix E.

Instructions for the evaluation users were sent in a voicemail message and by email. The evaluation reports and questionnaires were sent both through the postal service and email. The questionnaire was self-administered; therefore, the evaluation users were asked to complete and return each questionnaire within a week of receiving it. This process was repeated for both versions of the evaluation report and questionnaire dissemination. The dissemination of evaluation reports and questionnaires occurred in January 2010, while the last questionnaire was received by the investigator in July 2010.

Interviews with Evaluation Users

Following the collection of the questionnaires, the investigator scheduled semi-structured telephone interviews with all evaluation users who completed both of the utility questionnaires. The interviewees were told that the interview should take between 20 to 30 minutes and that the primary purpose of the interviews was to gather further qualitative depth on evaluation utility. The interviews occurred between July and September 2010.

Content Analysis of Evaluation Reports, Logs, and Approach Fidelity

The investigator conducted content analyses of the completed evaluation reports, the evaluators' logs, and the approach fidelity checklist. In conducting the content analysis of the two evaluation reports, the investigator created an instrument for coding, analyzing, and comparing the GAE report and the GFE report. On the left side of the instrument was a column with methodologically-related situations and decisions associated with the two evaluations and on the right was space for summaries and extraction from the reports as the comments pertained to evaluation utility (see Chapter IV: *Comparison of the GAE & GFE Reports' Contents*). Thus, the investigator compared both reports according to their inclusion and exclusion of statements relating to the methodology as well as the breadth and depth, quantity and quality of this information from the reports.

There were three types of logs completed by evaluators that were later analyzed by the investigator: the time log, the communication log, and logs to record threats to the goal achievement or goal-free nature of the evaluations. Time logs were used to assess the amount of time each team spent in evaluation-related activities, to see whether the teams differed in the overall amount of time spent on evaluation-related activities, and to see whether differences in time spent occurred. The communication logs were used to assess differences between teams with regard to the type, nature, and amount of communication with the evaluand's stakeholders. Lastly, the threats to the nature of GAE and GFE were evaluator-reported supplements to the *approach fidelity checklist* (see below). These logs were to investigate whether situations existed that potentially jeopardized the evaluators' independence and fidelity to their respective approaches; if

so, the quantity and severity of these threats were examined. Blank copies of all of these logs can be found in the Appendices in the evaluation teams' training handbooks.

This analog study represents a systematic investigation of idealized prescriptive theories of practice (i.e., what theory says we should do) rather than descriptive practice (i.e., what may actually occur); therefore, the delineation of the idealized versions of GAE and GFE was necessary. An idealized version of an evaluation approach refers to "the ideal (never achievable) evaluation theory [that] would describe and justify why certain evaluation practices lead to particular kinds of results across situations that evaluators confront" (Shadish et al., 1991, p. 31). In attempts to outline an idealized version of both approaches, the investigator created two lists of dos and don'ts, one for GAE and the other for GFE. The appropriate fidelity checklist was provided to the evaluation teams via their training handbook; additionally, the checklist was used by the evaluator for assessing the teams' apparent fidelity to their respective approach.

A formal pilot-testing of the approach fidelity checklists was not feasible because of resource limitations and the rarity of GFE in practice. Instead, the investigator generated criteria for approach fidelity by reviewing the literature on GAE and GFE and sought expert opinion on the initial list of ingredients for inclusion and exclusion. After the initial list of dos and don'ts for each approach was established, the investigator requested in an email, that over a dozen selected evaluation experts assess the importance of each ingredient to determine which are in fact essential for identifying an evaluation as GAE and GFE.[22] After receiving feedback from five Ph.D.s in evaluation and two

---

[22] The Approach Fidelity checklist was critiqued by Chris Coryn, Ph.D.; Wes Martz, Ph.D.; Michael Q. Patton, Ph.D.; James Sanders, Ph.D.; Daniela Schröeter, Ph.D.; Amy Gullickson, Ph.D.-ABD; and Lori Wingate, Ph.D.-ABD.

doctoral candidates in evaluation, the investigator considered their criticisms and finalized the checklist. Offered below (Figure 5) are the combined final versions of the two fidelity checklists.

This checklist is for performing Goal-Free Evaluation and Goal Achievement Evaluation approaches. The judgments about the fidelity of the evaluation approach can be made as follows: **Ø** Unacceptable Evaluator Performance, and √ Acceptable Evaluator Performance. It is recommended that the evaluation approach's fidelity fails if it scores unacceptable on one or more of the items.

Goal-Free Evaluation is the process determining merit with the evaluator maintaining partial or full independence from the stated (or implied) goals and objectives of those who design, produce, or implement the evaluand.

Goal Achievement Evaluation is the process of determining merit by analyzing whether the evaluand met or is meeting its goals and objectives.

A **goal** is a broad or general statement of a program's or intervention's purposes usually constituting longer-term expectations.

An **objective** is a specific, concrete, measurable statement of a program's or intervention's purpose usually constituting shorter-term expectations; it is the operationalization of a goal.

Goal-Free Evaluation	**Goal Achievement Evaluation**
Dos	**Dos**
□ Identify anduse a screener (i.e., an intermediary who ensures that no goal- or objective- based information is communicated to the goal-free evaluators).	□ Review program plans and meet with program staff to determine goals and objectives/identify the program's stated goals and objectives. If the program's goals are vague, translate them into measurable objectives.
□ Referall communiqués to screener and involve the screener throughout the evaluation to protect from potential contamination.	□ Determine that the goals and objectives are reasonably accurate, current, feasible, and specific.
□ Have all written material screened for references to program goals or objectives prior to evaluator receipt.	□ Identify or create standards/benchmarks based on the goals and objectives.

Goal-Free Evaluation	Goal Achievement Evaluation
Dos	**Dos**
□ Stop program staff if they begin talking aboutgoal-oriented information.	□ Measure performance related to goals and objectives.
□ Identify potential areas in which to search for effects (in part through a needs assessment) and use these as the basis for criteria to be measured.	□ Compare factual information with performance standards/benchmarks on the set goals and objectives and determine the extent to which the program achieved its goals and objectives.
□ Identify and select justifiable tools to measure performance and actual effects (i.e., tools that are reasonable with adequate groundsfor use).	□ Report on the program's performance in relation to its goals and objectives.
□ Measure performance and actual effects/ experience (observe) the program as is.	**Don'ts**
□ Compare factual information about the program effects/experiences withpre-identified needs to assess the program's impact on consumer needs.	□ Search for, measure, or report on side effects.
□ Offer a profile of the positive and negative effects.	□ Conduct a needs assessment.
Don'ts	
□ Communicate with program staff regarding goals or objectives.	
□ Attempt to find stated goals and objectives.	

Figure 5. Approach Fidelity Checklist

Procedures

The members of the GAE and GFE evaluation teams were trained by the principal and student investigators in the general principles and logic of their respective evaluation approach. The first team received training in GAE. Following the training, the GAE team conducted a rudimentary goal alignment by communicating with a few program

administrators to elicit program goals and objectives and to assess whether they are generally agreed upon. The program's goals were finalized and were then adopted as the criteria of merit by which the program's performance would be judged. Conversely, the opposite team was trained in GFE. The GFE team used direct observation and consumer interviews to observe program effects and outcomes related to their relevant needs and the extent to which the program met or was meeting those needs.

Each team prepared a written evaluation report. Before disseminating them to the evaluation users, the Evaluation Center editor reviewed the reports to control for egregious differences in writing skill, style, and format, etc. The evaluation users received the teams' evaluation reports one report at a time. After the evaluation user received and reviewed a report, the investigator distributed the *Evaluation Utility Questionnaire* to obtain the user's perceived utility of the evaluation's findings. The utility questionnaires were analyzed both qualitatively and quantitatively and then the investigator added to the qualitative depth on utility by interviewing the users.

The recruitment of student-evaluators was initiated in early November 2008 and was completed roughly a month later. Student-evaluators were trained in early February 2009 and began designing and conducting their evaluations shortly thereafter. The evaluators were instructed that they had roughly 24 weeks for conducting the evaluation and submitting the report. The GAE report was submitted on time, July 2009, while the GFE report was overdue, submitted in late September 2009.

The study's investigators conducted an initial meeting with a key program administrator in late November 2008, who discussed the study on evaluation to her coworkers. The program was evaluated by a GFE team and a GBE throughout the months

of February to September 2009. The evaluation users were given the two evaluation reports beginning in January 2010; both reports were read and both questionnaires completed and returned by July 2010. The follow-up interviews were conducted between July and September 2010. The total duration of the evaluation users' commitment to this study was nearly two years, November 2008 to September 2010.

The following is a brief description of chronological sequence of study design and implementation. The study occurred in four phases: (1) the pre-evaluation phase, (2) the evaluation phase, (3) the utility study phase, and (4) the utility analysis and reporting phase.

Phase One – Pre-Evaluation Phase

During the pre-evaluation phase, the investigator designed and developed the study. The investigator developed the hypothesis of study; operationalized evaluation utility; created data collection methods, tools and instruments, and procedures; and determined the methods for accessing and analyzing the data. Development and refinement were frequent and ongoing throughout the first and second phases of the study. Also during the pre-evaluation phase, the investigator constructed approach fidelity, pilot-tested the utility questionnaire, met with key program administrators, identified GAE and GFE report users, and screened program materials for dissemination to the GFE team. Moreover, the pre-evaluation stage included the recruitment of evaluators, the development of the evaluators' training materials, and the training of the evaluators from the two teams.

Phase Two – Evaluation Phase

During the evaluation phase the GAE and GFE teams designed and conducted their evaluations and wrote their reports. Throughout the duration of the evaluations, the student investigator communicated with each team weekly to supervise, to serve as a liaison between the teams and the program, to answer evaluation-related questions, and to reinforce fidelity to evaluation approaches. Concluding the evaluation phase was the student-investigator's receipt of the evaluation teams' logs and reports, and the editing of the evaluation reports by the investigators and the Evaluation Center editor.

Phase Three – Utility Study Phase

The third phase, the utility phase, consisted of the dissemination and collection of the utility questionnaire, and the semi-structured telephone interviews with the evaluation users.

Phase Four – Utility Analysis and Reporting Phase

The investigator analyzed the utility questionnaire results quantitatively and qualitatively. Mostly, non-statistical methods were used to analyze and interpret the interview results, the evaluation reports, the evaluators' logs, and the fidelity to the approach. The results of the utility study were reported in the dissertation and verbally to the program administrators.

Data Collection and Recording

Data were collected via email attachment (using Microsoft Word) and by hardcopy. The raw data were maintained and analyzed in Microsoft Office applications (i.e., Word and Excel). The evaluation teams were asked to submit a hardcopy and electronic copy their evaluation reports and logs in Microsoft Word. Most evaluation users who returned their completed *Evaluation Utility Questionnaires* did so via the U.S. Postal Service; a couple evaluation users completed the survey, scanned it, and attached it and sent it via email.

During the interviews of the evaluation users, the investigator collected data via audio recording and typed note-taking. The notes were used to jot down key points that were made by the interviewee, while replaying the recordings allowed the investigator to go back and collect accurate quotations from interviewees. The transcripts of the interviews were transferred from their digital audio recorder to Microsoft Office applications.

Data Processing and Analysis

To reduce the chance of transcription errors, data transcription was conducted simultaneously by two independent transcribers, the investigator and a volunteer. The two versions of transcriptions were then compared searching for discrepancies; together the investigator and second transcriber reviewed the disputed audio recording and came to agreement as to the accurate transcription.

The objectives of this dissertation are to answer the questions: (1) From the perspective of evaluation users, is there a difference between GAE and GFE with regard

to evaluation utility? (2) What, if any, are users' perceived differences in utility between GAE and GFE and, if differences do exist, how do they differ specifically in terms of instrumental use, conceptual use, process use, and persuasive use? and (3) If differences in perceived utility exist, what explains those differences? Therefore, the objectives should be reflected in both the data and the data analyses.

Methodological Limitations

Limitations temper results. There are certainly always many limitations in analog studies of this type. Below, in no particular order, is a non-exhaustive list of several limitations of this study.

One of the most significant limitations of a small *n* posttest-only analog study is the study's external validity. This refers to the study's lack of ability to generalize to programs beyond the specific one in this study, beyond the specific GAE and GFE approaches used in this study to all other GAEs and GFEs, and beyond the particular evaluators in this study to all goal achievement and goal-free evaluators. In addition, the duration of the study contributed to issues of generalizability as contextual factors resulted in an extended timeframe for the study; moreover, there was no long-term follow-up planned as part of this study to validate previous observations and conclusions.

Being an analog study, this study controlled various aspects of the evaluations but not others. For example, this study controlled the selection of evaluation approach and goal-orientation and controlled the number of evaluation team members. However, the investigator did not control or manipulate the evaluators' research designs or methods, nor did the investigator attempt to manipulate the program's outcomes.

Another limitation is that in an analog study simulating real evaluation conditions, the outcomes for the student-evaluators might not represent real-world positive or negative evaluation outcomes for the evaluators. Therefore, a potential criticism is that the student-evaluators' motivation and incentives significantly differed from actual professional evaluation practice. For example, a few differences between actual professional evaluation practice and this analog study include the fact that student-evaluators were receiving field experience credits, were juggling regular employment with doctoral studies, and were not financially compensated. Thus, it can also be argued, the student-evaluators were not representative of real evaluators in real conditions with real consequences.

Although the student-evaluators were randomly assigned to the teams, convenience sampling was used to recruit and select the six evaluators. Therefore, there are likely differences between teams and among team members. Basic evaluator demographic data were collected to assess whether there was evidence of professional or academic differences among evaluators that could be influencing the study.

This study had a couple additional limitations. This study is susceptible to social threats to internal validity as the research is conducted in a real-world context. For example, in a study of this type, the ever-present Hawthorne effect (i.e., reactivity) on behalf of the evaluation users and the evaluators is an unavoidable potential limitation. Lastly, there are logical problems and limitations with trying to combine the quantitative data from the semantic differentials with the qualitative responses from the questionnaire's open-ended portions and from the interviews.

Chapter Summary

In Chapter III, the methods used to conduct the analog study of GAE and GFE are described. Numerous methodological consideration are addressed in this chapter such as the study's design; the selection of the study's subjects; instrumentation; the study's procedures; data collection, recording, and analysis; and the study's limitations. The next, chapter, Chapter IV, describes the study's findings.

CHAPTER IV

FINDINGS

Chapter IV describes the findings of the study. The chapter begins with the identification and description of evaluation users. However, the bulk of the chapter consists of a comparison of the evaluation reports' content especially in relation to their selected methodologies and the reports' utility per the *Evaluation Utility Questionnaires* and the interviews.

Identification of Evaluation Users

On the *Identification of Evaluation Users* questionnaire, collectively the three program administrators identified 15 total individuals, 11 individuals with authority and influence within and/or over their program, and 14 staff with responsibilities in applying evaluation findings. There were 10 program people identified by administrators as being on both lists. Below (Figure 6) are the agency affiliations and job titles of those identified by the program administrators as someone with authority or influence within or over the program and its evaluation. The job titles of those marked with an asterisk denotes an individual identified by the administrators as being a person who has authority and/or influence within/over the program *and* authority in applying the findings from an evaluation.

Agency X

Authority and/or Influence within/over the program:

- Executive Director*
- Associate Director*
- ___ Director*
- ___ Specialist

Authority and/or Influence in Applying the Findings from an Evaluation:

- Executive Director*
- Associate Director*
- ___ Director*

Agency Y

Authority and/or Influence within/over the program:

- Director*
- Program Manager*
- Supervisor*

Authority and/or Influence in Applying the Findings from an Evaluation:

- Director*
- Program Manager*
- Supervisor*
- Worker
- Worker
- Worker
- Worker

Agency Z

Authority and/or Influence within/over the program:

- Program Coordinator*
- Director of Case Management Services*
- Vice President of Human Services*
- Chief Executive Officer*

Authority and/or Influence in Applying the Findings from an Evaluation:

- Program Coordinator*
- Director of Case Management Services*
- Vice President of Human Services*
- Chief Executive Officer*

Figure 6. Identified Evaluation Users

The investigator requested that the identified program people read both of the evaluation teams' reports. Of the 15 program people asked to read the evaluations, six (38%) successfully completed both utility questionnaires and five (31%) were interviewed. Therefore, four of four Agency X staff completed both questionnaires and three of four completed the interviews, while two of five Agency Y staff completed both questionnaires and two of five completed the interviews. In summary, 67% (six of nine) of the remaining evaluation users completed both utility questionnaires and 56% (five of nine) of evaluation users completed the interview.

Comparison of the GAE and GFE Reports

Below is a comparison of the final drafts of both reports.

Length of Evaluation Reports: The number of pages in the body of the GFE report was more than quadruple the length of the GAE report.

	GAE	GFE
Number of pages in the report	14	59
Number of pages in the appendix	8	12

Time on Evaluation: The evaluation teams spent roughly the same mean number of hours on evaluation activities; however, the goal-free evaluation team did so over twice as many days.

	GAE (n = 3)	GFE (n = 2)
Average hours per evaluator spent on evaluation	34	37
Average days per evaluator spent on evaluation	21	44

Threats to Nature: Neither team reported threats to the goal-based or goal-free nature.

	GAE	GFE
Reported threats to goal-based or goal-free nature	0	0

Neither team successfully recorded or reported the *Communication Logs*.

Comparison of the GAE and GFE Reports' Contents

Table 5, Methodological Comparison of the Evaluation Reports, is an abbreviated comparison of the evaluation approaches. The portions in quotations are verbatim extractions from the GAE or GFE report. Although the purpose of this study is to determine whether there are significant differences between goal achievement and goal-free approaches in terms of their utility from the perspective of the evaluation user, an examination of the differences and similarities in the two evaluations' methodologies offers contextual information and provides possible explanations for the evaluation users' various perspectives and conclusions.

Literature Review

Both evaluation teams reported conducting a limited literature review. The GAE team focused their efforts on reviewing a 2006 evaluation report that was conducted by an independent evaluation consulting firm as well as reviewing other program documents such as grant proposals and program brochures. The program documents and other written materials were, of course, screened to prevent the GFE team from any blatant goal-related information. Two examples of screened materials that were examined by the GFE team include the 2006 evaluation report and a description of the program taken from the Agency Y website. The GFE team reported that the majority of the literature review relied on publications relating to the needs of and issues faced by homeless individuals and families.

Table 5

Methodological Comparison of the Evaluation Reports

	GAE Team's Evaluation Report	**GFE Team's Evaluation Report**
Literature Review	Prior evaluation report, program documents	Screened program documents & publications related to needs & issues of homeless families
Evaluation Approach/Type	Goal-based/achievement-based, dimensional (p. 13), outcome-based (p.10)	Goal-free, CIPP (p. 7), case study (p. 2, 8)
Criteria of Merit	1. Employment 2. Housing	1. Community vision 2. Service delivery model 3. Program supports & resources 4. Client supports & resources
Definition of Evaluand Success	The number/percent employed & housed for six months or more	"Achieving positive client outcomes... directly attributable to the program" (p.5) & the meeting of the participants' "legitimate needs" (p.6)
Data Collection Methods	Preexisting quantitative employment & housing data collected by the program staff from 2008-2009	Predominantly qualitative data (e.g., semi-structured interviews & direct observation)collected by the evaluators during Spring 2009 to examine the program participants' experiences & program processes
Research Design	Fixed	Rolling
Sampling	Population Sample: All 72 program participants who were in the program for a year & who were gainfully employed	Non-probability sampling (purposeful & modal instance): 11 participants who were "currently participating" (p. 10) & considered "typical of this program, place, and time" (p. 10)
Data Analysis	"Calculating percentages and a univariate procedure" (p.4)	Thick description & adapted event history analysis
Standards & Comparisons	A grading scale based on 50% or below of either employment or housing is considered unacceptable by program administrators	This information is neither stated nor obvious from the GFE report; but it is possible that the evaluators used & intuitive/subjective grading scale.
Synthesis of Data	Numerical weight & sum (without weighting)	This information is neither stated nor obvious from the GFE report; but it is possible that the evaluators used & intuitive/subjective method of synthesis.

Table 5—Continued

	GAE Team's Evaluation Report	**GFE Team's Evaluation Report**
Main Findings	Inconsistent: From the executive summary,: 66% maintained employment while 45% maintained housing & "[the program] is successful in the employment program but is not successful in the housing program (p. 4); in Overall Significance, it states: "Since the results of data collection indicate that success has been achieved on both dimensions of merit considered by these selected stakeholders, the overall significance of the program is that it is a worthwhile and effective program" (p.13). "___ program be acknowledged as being successful in providing both employment and housing assistance to its participants" (p. 14).	The program is successful in providing "a temporary sheltering environment and support system, helping families identify resources and move forward in constructive ways to improve the quality of their lives. However the likelihood of housing sustainability for families appeared to be very low. The participants acquired job skills, but at the time of our interviews, none had obtained a job with an income sufficient to fully support themselves and their families, and few were confident they would be able to do so in the near future... we find the program as implemented falls short of helping participants become fully self-sufficient and able to achieve sustainable housing in the time allotted for service provision" (p. 2-3)
Impact	Immediate & short term	Immediate & short term
Evaluative Conclusions	(Inconsistent) The program is successful on employment but not housing or successful on both employment & housing	1. Community vision = satisfactory/marginal 2. Service delivery model = satisfactory 3. Program supports & resources = satisfactory/excellent 4. Client supports & resources = marginal/satisfactory

Evaluation Approach/Type

As directed, both the GAE team and the GFE team maintained fidelity in implementing an evaluation methodology that employed the designated approach. In addition to the goal-based approach, the GAE team reported using an outcome-based evaluation approach, in which the team determined the "merit, worth, and/or significance of an evaluand solely based on the evaluand's performance outcomes on stated goals" (p. 10). Furthermore, analytically, the GAE team chose a dimensional evaluation which

"looks at the performance of the program on multiple dimensions of merit that pertain to the evaluand as a whole" (p. 13). The GFE team reported using, not only the goal-fee approach, but also CIPP, i.e., context, input, process, and product (Stufflebeam, 1983). According to Stufflebeam (2002), in general, these parts of an evaluation respectively ask: (1) What needs to be done? (2) How should it be done? (3) Is it being done? and (4) Did it succeed? The GFE team also used a case study approach; the cases were a selection of the program's consumers. Lastly, and although not explicitly stated in the GFE report, the goal-free team, like the GAE team, employed a dimensional evaluation.

Criteria of Merit

In general, the criteria of merit are the characteristics or qualities that an evaluand must possess to be deemed good. The goal achievement team selected the two officially stated criteria for judging the quality of the program: (1) the employment status of the program participants, and (2) the housing status of the program participants; whereas without knowledge of the program's specific intentions, the goal-free team chose to determine success according to the quality of the program's (1) community vision, (2) service delivery model, (3) program supports and resources, and (4) client supports and resources.

Definition of Evaluand Success

Each team, in its report, described the hypothetical conditions for proclaiming the program successful. For the GAE team, success is based on the number (or percent) of the program participants who have maintained employment for six months or more and the number (or percent) of program participants who have maintained housing for six

months or more. As stated in the GFE report, the GFE team defined success as "achieving positive client outcomes that could be directly attributable to client participation in the program" (p. 5) and the meeting of the participants' "legitimate needs" (p. 6).

Data Collection Methods and Research Design

Methodologically, the GAE team relied on preexisting quantitative employment and housing data that were collected by the program staff during July 2008 through July 2009. The decision to use preexisting quantitative data resulted in the GAE team having a fixed research design, whereas the GFE team used a rolling design. The GFE team's case studies emphasized qualitative data collection methods like semi-structured interviews and direct observations.

Sampling

The GAE team took a population sample as all 72 program participants who were in the program from July 2008 to July 2009 and who were gainfully employed were included in the sample. The GFE team used non-probability sampling methods like purposeful and modal instance sampling as, during spring 2009, the team interviewed 11 selected participants who were currently participating and who were considered typical of the program, place, and time.

Data Analysis

Given the quantitative nature of the data examined by the GAE team, the team's methods of analysis were also quantitative. According to the GAE report, the methods for

analyzing the participants' employment and housing data included "calculating percentages and a univariate procedure" (p. 4). The goal-free team reported analyzing its mostly qualitative thick description by adapting a statistical procedure called event history analysis (Belli, 2009; Yamaguchi, 1991) into a procedure of coding and indexing followed by thematic analysis. Additionally, the GFE team reported detailed the program's strengths, weaknesses, opportunities, and threats (Rodríguez-Campos, 2005).

Standards and Comparisons

The goal-based team determined merit absolutely, not in comparison to others. The goal achievement team developed a grading rubric whereby any performance score below 50% on either the employment or housing outcome measures was considered unacceptable and scores of 80% and above were deemed excellent. The goal-free team determined absolute merit yet used qualitative methods of data collection. However, it is not explicit from reading the GFE report how the GFE team determined the standards for deeming program performance successful or poor. It may be likely that given the qualitative nature of the data collected, the grading scale was accomplished informally somewhat intuitively or subjectively in the minds of the goal-free evaluators and then verified via team deliberation.

Synthesis of Data

The GAE team reported using numerical weight and sum to derive an overall evaluative conclusion. Because employment and housing were considered equally important to program success, the GAE team weighted them equally during the synthesis process. Similarly to *standards and comparisons* above, the specifics of how the goal-

free team combined the performances on the various criteria to come up with an evaluative conclusion is not readily apparent. Again, it is possible that the synthesis process occurred informally and intuitively.

Findings and Conclusions

The goal achievement team had several undeniable inconsistencies in the findings of its report. From the executive summary it stated that 66% of program participants maintained employment while 45% maintained housing; the report continues, "[the program] is successful in the employment program but is not successful in the housing program" (p. 4). Later in the report, it is written that "since the results of data collection indicate that success has been achieved on both dimensions of merit considered by these selected stakeholders, the overall significance of the ___ program is that it is a worthwhile and effective program" (p. 13); and "___ program be acknowledged as being successful in providing both employment and housing assistance to its participants" (p. 14). These discrepancies were not explained by the evaluation team. Through a post-study informal discussion between the investigator and goal-based team members and through examining the latter contents of the report, it appears that in general the goal-based team considered the program a success.

The goal-free team's findings are summarized in the following quotation from the GFE report.

> [The program is successful in providing] a temporary sheltering environment and support system, helping families identify resources and move forward in constructive ways to improve the quality of their lives. However the likelihood of housing sustainability for families appeared to be very low. The participants acquired job skills, but at the time of our interviews, none had obtained a job with an income sufficient to fully support themselves and their families, and few were

> confident they would be able to do so in the near future... we find the program as implemented falls short of helping participants become fully self-sufficient and able to achieve sustainable housing in the time allotted for service provision. (pp. 2-3)

Both the GAE and the GFE teams reported they felt the affects or outcomes attributable to the program are likely to impact the participant immediately and for a short duration.

Evaluative Conclusions

The overall evaluative conclusion is inconsistent in the GAE report. The report states that the program is successful on employment but not housing, yet later states that the program is successful on both employment and housing. The GFE team offered a profile of the performance on the four criteria of merit. The GFE team assigned the program a satisfactory/marginal on community vision, satisfactory on service delivery model, satisfactory/excellent on program supports and resources, and marginal/satisfactory on client supports and resources.

GAE and GFE Reports' Utility via the Questionnaire

In this study, the purpose behind the data collection is to examine evaluation utility from the perspective of the evaluation user and thus the *Evaluation Utility Questionnaire* (see Appendix E) was developed and pilot-tested. The bulk of the study's quantitative data on evaluation report utility come from the semantic differential in the utility questionnaire although the final question on the questionnaire is qualitative.

Responses to the Semantic Differential

As previously stated, the evaluation users were randomly assigned the order in which they would read, and respond to, the goal achievement and goal-free reports. Therefore, there were two rounds of questionnaire administrations. Below in Table 6 are the results from the two administrations of the semantic differential rating scales. Each column represents an evaluation user, whereas Agency X = A-X and Agency Y = A-Y. The columns with an asterisk indicate evaluation users who were responding to the goal achievement report, while the columns without represent evaluation users who were responding to the goal-free evaluation report. The numbers (from 3 to –3) represent the seven-point scale used whereby a neutral response is equivalent to zero; an example of the scale is shown in Figure 7.

Table 6 below displays a summary of evaluation users' mean scores from the semantic differential portion of the *Evaluation Utility Questionnaire*. Notice that A-Y01, A-Y02, and A-Y03 did not return the questionnaires. Also, notice that regardless of the round, the highest mean score for the GAE report was 2.92 from A-X01, while A-X02 had the lowest mean score for the GAE report at –0.24. The highest mean score for the GFE report was 2.72 by A-X01, and A-Y05 had the lowest mean score for the GFE report at –0.45. Across all six respondents, GAE has slightly more utility than GFE by 0.15 (i.e., 1.09 – 0.94 = 0.15). Also displayed in Table 7 is the difference between a respondent's GAE and GFE utility mean scores. A-X03 is the respondent who reported the greatest difference in utility between the two evaluation reports, a difference of 2.04 in favor of the goal achievement approach, whereas A-X04 found little difference in

Table 6

Evaluation Utility Questionnaire Administration Rounds 1 and 2

Round 1		*GAE A-X01	GFE A-X02	*GAE A-X03	*GAE A-X04	GFE A-Y04	*GAE A-Y05
Useful	Useless	3	2	2	1	1	1
Conclusive	Inconclusive	3	–1	2	1	0	0
Believable	Unbelievable	3	2	2	3	2	1
Trustworthy	Untrustworthy	3	1	1	x	1	1
Clear	Unclear	2	2	3	–1	3	0
Consistent	Inconsistent	3	1	2	0	1	–3
True	False	3	2	3	1	1	x
Careful	Careless	3	1	1	2	2	0
Logical	Illogical	3	1	2	2	1	1
Valid	Invalid	3	0	2	2	1	1
Meaningful	Meaningless	3	2	3	1	0	1
Worthwhile	Worthless	3	1	2	1	1	1
Complete	Incomplete	2	–1	2	2	1	–1
Correct	Incorrect	3	1	2	0	1	–3
Helpful	Unhelpful	3	1	2	0	2	1
Objective	Biased	3	2	2	1	1	–1
Specific	Vague	3	2	2	1	2	–3
Enlightening	Unenlightening	3	0	2	0	0	1
Fair	Unfair	3	1	2	1	1	1
Relevant	Irrelevant	3	1	3	1	1	0
Reasonable	Unreasonable	3	2	2	1	2	1
Informative	Uninformative	3	1	3	2	2	1
Honest	Dishonest	3	2	2	0	2	–1
Effective	Ineffective	3	0	2	1	2	–2
Balanced	Unbalanced	3	2	2	1	1	–3
	Mean	2.92	1.12	2.12	1.00	1.28	–0.21
	Standard Deviation	0.28	0.93	0.53	0.88	0.74	1.53

Table 6—Continued

Round 2		GFE A-X01	*GAE A-X02	GFE A-X03	GFE A-X04	*GAE A-Y04	GFE A-Y05
Useful	Useless	3	0	0	1	1	x
Conclusive	Inconclusive	2	0	–1	1	1	x
Believable	Unbelievable	3	0	1	1	0	x
Trustworthy	Untrustworthy	3	–1	0	2	1	–2
Clear	Unclear	3	–1	–1	1	1	2
Consistent	Inconsistent	3	–0.5	–2	–1	1	–2
True	False	2	0	1	1	1	–2
Careful	Careless	3	0	0	1	1	0
Logical	Illogical	3	0	0	1	–1	–1
Valid	Invalid	2	0	1	1	1	–1
Meaningful	Meaningless	3	–0.5	1	1	–1	–1
Worthwhile	Worthless	3	0	0	1	–1	–1
Complete	Incomplete	2	–1	–2	–2	2	0
Correct	Incorrect	2	0	0	–1	2	0
Helpful	Unhelpful	3	0	1	1	0	–3
Objective	Biased	3	0	0	2	2	0
Specific	Vague	3	–1	0	2	2	1
Enlightening	Unenlightening	3	0	–1	0	0	–3
Fair	Unfair	2	0	2	0	2	1
Relevant	Irrelevant	3	0	1	2	0	1
Reasonable	Unreasonable	3	0	0	2	2	1
Informative	Uninformative	3	0	0	1	2	1
Honest	Dishonest	3	0	0	2	2	1
Effective	Ineffective	3	0	1	1	0	–3
Balanced	Unbalanced	2	–1	0	1	2	1
	Mean	2.72	–0.24	0.08	0.88	0.92	–0.45
	Standard Deviation	0.46	0.41	0.95	1.01	1.04	1.53

Note. x = blank

Useful	3	2	1	_0_	–1_	–2_	–3_	Useless

Figure 7. The Seven-Point Scale Used in the Semantic Differential

Table 7

Summary of Means Scores from the Semantic Differential

	A - X01	A - X02	A - X03	A - X04	A - Y04	A - Y05	TOTALS (mean)
GAE Mean Score (*n*=6)	2.92	–0.24	2.12	1.00	0.92	–0.21	1.09 (*SD*=1.26)
GFE Mean Score (*n*=6)	2.72	1.12	0.08	0.88	1.28	–0.45	0.94 (*SD*=1.09)
Difference in Mean Score Favors	GAE	GFE	GAE	GAE	GFE	GAE	4 Favor GAE 2 Favor GFE
By a Difference of…	0.20	1.36	2.04	0.12	0.36	0.24	- - -

utility between the two reports represented by a difference of 0.12. The table also shows that four of six individual respondents reported scores that favored GAE over GFE.

Averaging the differences in utility mean scores from Table 7 illustrates which supporters reported the strongest favor for one of the evaluation approaches. This is displayed below in Table 8, and as can be seen, the two respondents whose utility mean scores favored GFE felt slightly more strongly about GFE (0.86) than the four who favored GAE (0.65).

Table 8

Average Difference in Utility Means Scores per Evaluation Approach

GAE	0.20 + 2.04 + 0.12 + 0.24 / 4	=	0.65
GFE	1.36 + 0.36 / 2	=	0.86
	Difference	=	0.21

The mean scores per adjective pair are displayed below in Table 9. The two columns on the left are the bipolar adjective pairs while the middle column represents the mean scores of the six evaluation users. Lastly, the range of mean scores for all adjective pairs is also included in the far right column of the table. For example, notice that the evaluation users found the GAE report informative (1.83) yet inconsistent (0.42), while the same evaluation users found the GFE report believable (1.80) yet incomplete (–0.33).

Summary of Open-Ended Responses on Utility Questionnaire

As stated, the final question on the *Evaluation Utility Questionnaire* was qualitative; it requested that the evaluation user "provide an explanation as to why the evaluation report was or was not useful." Offered below in Table 10 is a summary of the main themes mentioned by evaluation users in their responses to the open-ended question. In the table, six columns represent the six evaluation users and their responses to both reports.

GAE and GFE Reports' Utility According to the Interviews

The purpose behind the post-evaluation semi-structured telephone interviews was to triangulate on evaluation utility from the perspective of the evaluation users by supplementing the data from the semantic differentials with descriptive qualitative data. Furthermore, the interviews were used to ask specifically about instrumental, conceptual, and persuasive utility. Lastly, and as stated in the methodology section, the order in which the evaluation reports were discussed during each interview was predetermined by a random process and which, for simplicity sake, is not represented in the tables below.

Table 9

GAE and GFE Adjective Pairs Means

GAE Report Mean (n = 6)				GFE Report Mean (n = 6)			
Useful	Useless	1.33		Useful	Useless	1.40*	
Conclusive	Inconclusive	1.17		Conclusive	Inconclusive	0.20*	
Believable	Unbelievable	1.50		Believable	Unbelievable	1.80*	max.
Trustworthy	Untrustworthy	1.00*		Trustworthy	Untrustworthy	0.83	
Clear	Unclear	0.67		Clear	Unclear	1.67	
Consistent	Inconsistent	0.42	min.	Consistent	Inconsistent	0.00	
True	False	1.60*		True	False	0.83	
Careful	Careless	1.17		Careful	Careless	1.17	
Logical	Illogical	1.17		Logical	Illogical	0.83	
Valid	Invalid	1.50		Valid	Invalid	0.67	
Meaningful	Meaningless	1.08		Meaningful	Meaningless	1.00	
Worthwhile	Worthless	1.00		Worthwhile	Worthless	0.83	
Complete	Incomplete	1.00		Complete	Incomplete	-0.33	min.
Correct	Incorrect	0.67		Correct	Incorrect	0.50	
Helpful	Unhelpful	1.00		Helpful	Unhelpful	0.83	
Objective	Biased	1.17		Objective	Biased	1.33	
Specific	Vague	0.67		Specific	Vague	1.67	
Enlightening	Unenlightening	1.00		Enlightening	Unenlightening	–0.17	
Fair	Unfair	1.50		Fair	Unfair	1.17	
Relevant	Irrelevant	1.17		Relevant	Irrelevant	1.50	
Reasonable	Unreasonable	1.50		Reasonable	Unreasonable	1.67	
Informative	Uninformative	1.83	max.	Informative	Uninformative	1.33	
Honest	Dishonest	1.00		Honest	Dishonest	1.67	
Effective	Ineffective	0.67		Effective	Ineffective	0.67	
Balanced	Unbalanced	0.67		Balanced	Unbalanced	1.17	

*Denotes an adjective pair with a missing response (n = 5)

Table 10

Summary of Open-Ended Response on the Evaluation Utility Questionnaire

	GAE	GFE
A-X01	Evaluator independence valuable; report unclear in portions	Report provides insights into effectiveness & consumer needs
A-X02	Questions regarding evaluation sampling & grading methods	[Blank]
A-X03	[Blank]	Dimension of merit & capacity building helpful; recommendations & conclusions questionable; some specifics of report valuable
A-X04	Report could be more helpful if an evaluator explained it	Feedback from participants & evaluator neutrality helpful; family mentoring is good recommendation
A-Y04	[Blank]	[Blank]
A-Y05	[Blank]	[Blank]

Table 11 below is a summary of the time (in minutes) it took to complete each interview in which the evaluation users were asked questions regarding both GAE and GFE. The six columns represent the six evaluation users and the time for each interview. On the far right, is the mean time per interview (20.6 minutes) across all evaluation users interviewed (n = 5). The interview with A-X03 was the longest at 30 minutes, while the shortest was A-Y05's 10 minutes.

Table 11

Length of Time for Interviews

	A-X01	A-X02	A-X03	A-Y04	A-Y05	TOTALS (mean)
Interview Time	23	25	30	15	10	20.6 minutes (SD = 8.02)

Next in Tables 12 and 13 are the first two sets of questions that were asked of evaluation users and their responses. What did the evaluation user find to be the most useful and least useful aspects of the GAE and GFE reports?

Table 12

Question 1 – What Evaluation Users Found Most Useful About Each Approach

	What was most useful?	
	GAE	GFE
A-X01	Succinct; easy to read; quant. data is cut & dry; interpretable	Gets clients' perspective & worldview
A-X02	The numbers concrete	[Blank]
A-X03	Values section & grading system; the focus on employment & housing; the timeframes; clients disclosed to independent evaluators; reminder of multifaceted barriers to success	Clients disclosed to independent evaluators; reminder of multifaceted barriers to success
A-X04	* "I apologize I have been unable to respond to you until now. In preparation for ending my employment at Agency X, I have been extremely busy trying to meet deadlines. Unfortunately, I am unable to honor your request. My last day of employment at Agency X is today. Best of luck to you."	
A-Y04	No response because didn't read the report well	States why the clients liked/ disliked
A-Y05	Didn't read the report thoroughly; no longer working with the program	Unable to comment
Summary	Succinct; easy to read & interpret; quantitative data; values & grading system; focus on employment & housing	Clients' perspective & worldview; client disclosure; reminder of clients' barriers; explains why clients liked/disliked program

*A-X04 has been removed from the remaining tables as A-X04 terminated employment with Agency X after completing the utility questionnaire but prior to being interviewed.

Table 13

Question 2 – What Evaluation Users Found Least Useful about Each Approach

	What was least useful?	
	GAE	GFE
A-X01	Accuracy of numbers questioned; inconsistencies; no guidance for making improvements; no examination of other outcomes	Lacks representativeness; is anecdotal/idiosyncratic; information not helpful for obtaining funding
A-X02	Questions veracity of some numbers	Criticizes small sample size, representativeness, & subjectivity
A-X03	Nothing but would like more examination of impact	Difficult to read; not focused; issues with logic
A-Y04	Quantitative aspects (e.g., graphs & numbers) are not exciting & may not reflect reality	Nothing, it was relatively useful
A-Y05	Didn't read; not working with the program	Didn't read; not working with the program
Summary	Reliance on quantitative measures; accuracy & veracity of numbers; inconsistencies; no examination of alternative outcomes; no guidance for improving	Small sample size; lacks representativeness; is subjective & anecdotal; difficult to read; not focused; logic issues

Dimensions of Evaluation Utility

The next 11 tables represent the eleven measured variables associated with the three dimensions of evaluation utility employed in this study (i.e., instrumental utility, conceptual utility, and persuasive utility). The interviewer specifically solicited responses from evaluation users on these dimensions and the 11 variables associated with them. The column on the right contains summaries of the evaluation users' affirmative responses; i.e., the *no*, *none*, *not applicable*, or *do not know* responses are excluded from the summary column.

Instrumental Utility

The first five tables are associated with instrumental utility. Table 14 shows the evaluation users' responses to what they felt was information for improving the program while the second table (Table 15) asks whether there was information useful for making decisions. Table 16 displays whether there was information useful for holding the program and others accountable, while Tables 17 and 18 ask whether the evaluation report contained information useful for making generalizations regarding program performance and program effectiveness respectively.

Table 14

Question 3 – What Evaluation Users Found Useful for Improving the Program

	Have you and/or the program used information from the [GAE/GFE] report for improving the program?	
	GAE	GFE
A-X01	No	No
A-X02	No	No
A-X03	Used evaluation data to compare with self-sufficiency matrix	Possible improvements in examining clients' requests for (criminal) legal & family counseling assistance
A-Y04	No	No; not sure
A-Y05	Not sure	No
Summary	Evaluation results compared with other data	The degree to which clients request for legal & family counseling was not previously known by the program

Table 15

Question 4 – What Evaluation Users Found Useful for Making Decisions

	Have you and/or the program used information from the [GAE/GFE] report for making decisions?	
	GAE	GFE
A-X01	No	Yes, In discussions with staff regarding programming
A-X02	No	No
A-X03	No	No
A-Y04	No	No
A-Y05	No	No
Summary	No	In decisions regarding programming

Table 16

Question 5 – What Evaluation Users Found Useful for Accountability Purposes

	Have you and/or the program used information from the [GAE/GFE] report for accountability purposes?	
	GAE	GFE
A-X01	No	No
A-X02	In examining the program's processes	The report holds the program accountable for communication with clients about timeliness & expectations
A-X03	The program's partners used evaluation to justify improving employment figures	Employment supports useful especially in justifying transportation assistance for clients
A-Y04	No	Holds the program's clients accountable
A-Y05	No	No
Summary	Examines the program's processes; justifies improving employment figures	Holds clients accountable; examines communication with clients; justifies need for transportation assistance

Table 17

Question 6 – What Evaluation Users Found Useful for Making Generalizations About Program Performance

	Is there information in the [GAE/GFE] report that allows you and/or the program to make generalizations about the program's performance?	
	GAE	GFE
A-X01	No	Generalizes about some perceived negatives that are really positives
A-X02	No	No
A-X03	Employment & housing numbers	No
A-Y04	No	Generalizes about funding eligibility with clients & abruptness of funding stoppage
A-Y05	No	No
Summary	Employment & housing numbers	Generalizes about some perceived negatives; generalizes about client funding eligibility

Table 18

Question 7 – What Evaluation Users Found Useful for Making Generalizations About Program Effectiveness

	Is there information in the [GAE/GFE] report that allows you and/or the program to make generalizations about the program's effectiveness?	
	GAE	GFE
A-X01	No	Generalizes about housing supports & positive impact on clients
A-X02	No	No
A-X03	In general, employment needs to be strengthened	Insightful that clients wanted housing assistance linked to actual employment
A-Y04	No	In general, the program helps those who are motivated
A-Y05	No	No
Summary	Employment needs strengthening	Generalizes about positive effect of the program housing supports; insightful that clients want housing assistance linked w/ actual employment; the program most helpful to those who are motivated

To summarize the previous five tables concerning instrumental utility, the evaluation users reported several ways the information from each report was instrumentally useful. Beginning with the GAE report, a few common themes deemed instrumentally utile include that the GAE report provided data that could be compared with other organizations' evaluation reports, it held the program accountable for its clients' employment outcomes, and the report offered some generalizations regarding the program and its clients, whereas the instrumentally useful portions from the GFE report included information on unrecognized client needs, information useful for making decisions regarding programming, information that justifies transportation assistance for clients, and information that generalizes about housing supports and client motivation.

Conceptual Utility

The next three tables are associated with the conceptual utility dimension. Table 19 displays the responses to whether the report offers information that helps better understand the program, while Table 20 reports whether the evaluation report improved the stakeholders' understanding. The final table under conceptual utility (Table 21) shows whether the information from the evaluation report assisted the evaluation user in better comprehending her personal roles and responsibilities as it pertains to the program.

The three tables below display what the evaluation users reported was conceptually utile information from the GAE and GFE reports. Those who responded found the GAE report to be useful mostly in terms of conceptualizing the program's employment-related aspects, whereas the GFE report provided evaluation users with a broader perspective of program participants.

Table 19

Question 8 – What Evaluation Users Found Useful for Understanding What the Program Is and Does

	Is there information in the [GAE/GFE] report that allows you and/or the program to better understand what the program is and does?	
	GAE	GFE
A-X01	Some in terms of employment	With regard to clients'' perspectives
A-X02	With regard to employment successes	SWOT analysis helped
A-X03	No	No
A-Y04	No	Evaluators described the program saying things that evaluation user didn't know
A-Y05	No	No
Summary	Employment successes	Clients' perspectives; SWOT analysis; evaluators taught program staff new facts about the program

Table 20

Question 9 – What Evaluation Users Found Useful for Understanding the Program's Stakeholders

	Is there information in the [GAE/GFE] report that allows you and/or the program to better understand the program's stakeholders and what they do?	
	GAE	GFE
A-X01	With regard to the employment component	With regard to the clients' perspectives
A-X02	No	No
A-X03	No	Yes because of the external independent interviewers' ability to get honesty & openness from the program's clients; getting information from those who are failing the program useful
A-Y04	No	With regard to the report's background; interview questions & depth were useful
A-Y05	No	No
Summary	Employment info	Open & honest perspective of client useful; there are benefits of external interviewers interviewing program clients; interviews with "less"-successful clients useful; the report's "background" section useful; evaluators used good interview questions & depth

Table 21

Question 10 – What Evaluation Users Found Useful for Understanding Their Roles and Responsibilities

	Is there information in the [GAE/GFE] report that allows you to better understand *your* roles and responsibilities with regard to the program?	
	GAE	GFE
A-X01	With regard to partnership in employment supports	With regard to the need for more awareness of complexities of housing & employment success
A-X02	In comparing the program's outcomes with other organizations	With regard to the need to examine structure of client placements
A-X03	No	No
A-Y04	No	Information from the interview questions useful
A-Y05	No	No
Summary	Partnership in employment; compared evaluation report findings with other orgs	Awareness of complexities of success; info from interview questions

Persuasive Utility

The following three tables are associated with persuasive utility, the third and final dimension of evaluation utility examined in this study. Table 22 provides the responses to the question of whether the evaluation report contains information for supporting a change, while Table 23 is about opposing a change. Stakeholder ownership in the program is the topic of Table 24.

Table 22

Question 11 – What Evaluation Users Found Useful for Supporting a Change

	Is there information in the [GAE/GFE] report that allows you and/or the program to support a change within the program?	
	GAE	GFE
A-X01	No	Supports changes in housing
A-X02	No	Possible changes in partners, communication, & program delivery
A-X03	No	Supports changing communication & client feedback process
A-Y04	No	No
A-Y05	No	No
Summary	None	Changes in housing, partners, communication, program delivery, client feedback process

Table 23

Question 12 – What Evaluation Users Found Useful for Opposing a Change

	Is there information in the [GAE/GFE] report that allows you and/or the program to oppose a change within the program?	
	GAE	GFE
A-X01	No	No
A-X02	No	No
A-X03	Opposes reductions in funding as program is worthwhile	No
A-Y04	No	No
A-Y05	No	No
Summary	Oppose changes in program funding	No

Table 24

Question 13 – What Evaluation Users Found Useful for Increasing Stakeholder Ownership

	Is there information in the [GAE/GFE] report that helps increase the stakeholders' ownership of the program?	
	GAE	GFE
A-X01	No	No
A-X02	No	Probably
A-X03	In promoting program results to potential clients; showing clients the program's expectations; some statistics were reported to stakeholders	Agency Y's staff didn't complete utility surveys & interviews which shows Agency Y's lack of ownership/ commitment to the program
A-Y04	No	No
A-Y05	No	No
Summary	Promotion of program to potential clients; shows the program's expectations; some information shared with stakeholders	Probably; shows lack of Agency Y's commitment to the program

The three tables displayed above relate to persuasive utility. The GAE report contained information that may be used to persuade others not to change the program's funding as well as information that might be used promotionally and/or informatively. The GFE report can be used to persuade program stakeholders to support changes in housing, partners, communication, program delivery, and the client feedback process.

Table 25 is a summary of the GAE and GFE according to evaluation users' responses during the interviews to the questions associated with instrumental, conceptual, and persuasive utility. Evaluation users reported that the GAE report lacked information for making decisions and information for supporting any changes, while according to evaluation users, the GFE report did not have information that could be used to oppose changes.

Table 25

Summary of GAE and GFE Instrumental, Conceptual, and Persuasive Utility According to Evaluation Users' Interviews

GAE	Instrumental Utility	Information for comparing with other organizations, holding the program accountable for clients' employment outcomes, & generalizing about the program & clients
	Conceptual Utility	Information regarding the program's employment-related aspects
	Persuasive Utility	Information for persuading not to change funding & for promotional/informational purposes
GFE	Instrumental Utility	Information for recognizing client needs, making decisions regarding programming, justifying transportation assistance for clients, & generalizing about housing supports & clients
	Conceptual Utility	Information better understanding the clients' perspectives
	Persuasive Utility	Information for persuading the program stakeholders to make changes in housing, partners, communication, program delivery, & client feedback process

Table 26 displays what evaluation users attributed differences in usefulness between the two evaluation reports.

Table 26

Question 14 – What Evaluation Users Think Accounts for Difference in Usefulness between GAE and GFE

	A-X01	A-X02	A-X03	A-Y04	A-Y05
What do you think accounts for the main difference in usefulness between the two reports?	GFE is better for understanding the client, developing the program, meeting the clients' needs; while GAE is better for developing funds & partnerships	GAE is better because of numbers for comparisons but has conflicting statements & not sure who authorized official goals; GFE is too subjective & not to the point	GFE is too much like an academic exercise; GAE is better because more targeted	GFE is better because of more info & more engaging read; GAE has too many numbers	They were equal

The final three questions during the interview were not directly related to the utility dimensions. The first of the final three interview questions asks the evaluation user to provide an assessment of the two evaluation approaches (Table 27). The second to last scheduled interview question asks evaluation users to provide the evaluators with suggestions for future evaluations (Table 28), while the last question seeks users' additional comments (Table 29).

Figure 8 is a summary of the positive and negatives of both approaches per evaluation users.

Table 27

Question 15 – Evaluation Users Suggestions for Evaluators

	A-X01	A-X02	A-X03	A-Y04	A-Y05	Summary
What do you suggest that either of the evaluation teams do next time?	GAE should proofread & edit	GAE should add more detail & clarity; GFE should use GAE format	Evaluation user noticed evaluators had wide range of individual skill level with regard to working with the program's clients (some good some not so)	No	No	Proofing, formatting, detail, Clarity, range of evaluators' skills with the program clients

Table 28

Question 16 – Additional Comments by Evaluation Users Regarding GAE and GFE Utility

	A-X01	A-X02	A-X03	A-Y04	A-Y05	Summary
Is there anything else you'd like to add about either of the reports' utility?	The program will use both reports	Employment environment is very different since the evaluations began; Agency Z is no longer partnering with the program	No	No	No	Will use reports; employment environment changed

Below in Table 29 is the evaluation users' position regarding which evaluation approach is more useful as stated or implied during the semi-structured telephone interviews. After removing the one "undecided" and the one "equal," GAE had two users find it more useful while GFE had one users find it the more useful.

Table 29

Summary of Interviews

	A-X01	A-X02	A-X03	A-X04	A-Y04	A-Y05	TOTALS (mean)
Stated Favorite Approach (see Table 25)	Undecided	GAE	GAE	N/A	GFE	Equal	1 Undecided 1 Favor Equal 2 Favor GAE 1 Favor GFE

Summary of Users' Responses to Question 14
+ GAE has information for getting funding & partners, making comparisons; & is targeted
– GAE has conflicting statements, issues with goal authorization, & is too quantitative
+ GFE has information for understanding clients, program development, & meeting clients' needs; has more information in general; & is an engaging read
– GFE is too subjective, too indirect, & too much like an exercise

Figure 8. Positives and Negatives of GAE and GFE per Evaluation Users

Summary of the Individual Evaluation User

The following six tables (Tables 30-35) are summaries of the responses from each individual evaluation user. The summaries include the results from the two questionnaires as well as quotes from the interviews.

Table 30

Summary of A-X01

GAE mean score		2.92
GFE mean score		2.72
Evaluation Utility Questionnaire mean scores favors…		GAE
…by a difference of…		0.20
During the interview the evaluation user reported favoring…		Undecided
Difference btw questionnaire & interview?	Semantic Differential:	Finds GAE slightly more useful
	Interview:	Undecided which is more useful
Evaluation User's Conclusion	Interview Question 14 "From a programmatic perspective of understanding the clients, the GFE was probably the most insightful but I'd have to say the GAE is what I'd need to use for any external fundraising/fund development/partner collaboration. But the goal-free one [report] gives you, how do you develop your programs a little more, and understand the people in need, and how to best meet their needs."	

Table 31

Summary of A-X02

GAE mean score		–0.24
GFE mean score		1.12
Evaluation Utility Questionnaire mean scores favors…		GFE
…by a difference of…		1.36
During the interview the evaluation user reported favoring…		GAE
Difference btw questionnaire & interview?	Semantic Differential:	Finds GFE clearly more useful
	Interview:	Finds GAE more useful
Evaluation User's Conclusion	Interview Question 1A "I liked the actual analysis, the numbers and percentages. I like how that was delivered. It seemed a little bit more concrete than the other [GFE]." Interview Question 14 [The reason GAE was more useful is…] "I think, more or less, it's just the numbers and percentages; and having the actual numbers you can compare it to. The other one [GFE] is informative but I think it's too subjective and you have to weed through the detail to get to the summary."	

Table 32

Summary of A-X03

GAE mean score		2.12
GFE mean score		0.08
Evaluation Utility Questionnaire mean scores favors…		GAE
…by a difference of…		2.04
During the interview the evaluation user reported favoring…		N/A
Difference btw questionnaire & interview?	Semantic Differential:	Finds GAE clearly more useful
	Interview:	Finds GAE more useful
Evaluation User's Conclusion	Interview Question 14 "I felt like the first report [GFE] was more like a school lesson. I think that the targeted [GAE] was much more helpful. [With the GFE report] I felt like I was reading a bibliography of poverty reduction initiatives."	

Table 33

Summary of A-X04

GAE mean score		1.00
GFE mean score		0.88
Evaluation Utility Questionnaire mean scores favors…		GAE
…by a difference of…		0.12
During the interview the evaluation user reported favoring…		GAE
Difference btw questionnaire & interview?	Semantic Differential:	Finds GAE slightly more useful
	Interview:	N/A
Evaluation User's Conclusion	N/A	

Table 34

Summary of A-Y04

GAE mean score		0.92
GFE mean score		1.28
Evaluation Utility Questionnaire mean scores favors…		GFE
…by a difference of…		0.36
During the interview the evaluation user reported favoring…		GFE
Difference btw questionnaire & interview?	Semantic Differential:	Finds GFE slightly more useful
	Interview:	Finds GFE more useful
Evaluation User's Conclusion	Interview Question 1B "That's [GFE] the one I liked. I didn't like the other one [GAE]; it was mostly reports. GFE had a participant's statements in there, stating the reason why they did like the program or they did not like the program; and that was useful.	

Table 35

Summary of A-Y05

GAE mean score		–0.21
GFE mean score		–0.45
Evaluation Utility Questionnaire mean scores favors…		GAE
…by a difference of…		0.24
During the interview the evaluation user reported favoring…		Equal
Difference btw questionnaire & interview?	Semantic Differential:	Finds GAE slightly more useful
	Interview:	Finds approaches equally useful
Evaluation User's Conclusion	Interview Question 14 "They were both equal to me. I guess they both weren't really useful since we're really not in the program… like we were before."	

The above summarizations are abbreviated case studies of the individual evaluation user to examine internal consistency in reporting whether GAE and GFE were useful or not useful. Were the evaluation users consistent across and within the two primary methods of questioning (i.e., the questionnaires and interviews)? For instance, A-X02 might be considered the most inconsistent as this user of the evaluation report found GFE more useful according to her scores on the semantic differential rating scales; however, during the interview, she stated a clear preference for GAE, whereas A-X03 might be the most consistent in her position that GAE is more useful.

Chapter Summary

Chapter IV described the findings of the study. The chapter begins with background and demographic information on the goal achievement and goal-free evaluators, the identification and description of evaluation users, and a comparison of the evaluations' methodologies. The majority of this chapter presents the findings of the

evaluation users' responses to the *Evaluation Utility Questionnaires* and the responses from the interviews with evaluation users. The study's summary and conclusions, implications, limitations, and directions for future research are discussed next in Chapter V.

CHAPTER V

SUMMARY AND CONCLUSIONS

The previous chapter described the study's findings. This chapter provides the study's summary, conclusions, implications, and limitations as well as directions for future research.

Summary

The study consisted of a trained team of goal achievement evaluators and a trained team of goal-free evaluators who independently evaluated the same evaluand using their respective evaluation approaches. Afterward, each team produced a final evaluation report which was read by relevant evaluation users. First, the evaluation users responded to a questionnaire regarding the usefulness of the information in each report and then they were interviewed about the reports.

Below in Table 36 is a combined summary of all evaluation users who responded to either or both questionnaires and interview. An examination of the table shows that across evaluation users, GAE appears to be slightly more useful than GFE.

The bulk of the study's quantitative data was gathered using the *Evaluation Utility Questionnaire,* which consisted of a semantic differential rating scale for comparing the utility of each evaluation report according to evaluation users. For instance, as can be seen in Table 36 below, A-X03 experienced the greatest difference between the utility of the two evaluation approaches (2.04), favoring GAE, while A-X04 felt the two

approaches were nearly equally useful/useless (0.12). Additionally, the overall mean score across all evaluation users from the semantic differential was 0.15 in favor of GAE. Finally, according to the mean scores of adjective pairs in the semantic differential, the best adjectives to describe GAE and GFE are (see Table 9):

GAE report: *Informative* (1.83) yet *Inconsistent* (0.42)

GFE report: *Believable* (1.80) yet *Incomplete* (–0.33)

Table 36

Combined Summary of Evaluation Users

	A-X01	A-X02	A-X03	A-X04	A-Y04	A-Y05	TOTALS
GAE Mean Score (*n*=6)	2.92	–0.24	2.12	1.00	0.92	–0.21	1.09 (*SD*=1.26)
GFE Mean Score (*n*=6)	2.72	1.12	0.08	0.88	1.28	–0.45	0.94 (*SD*=1.09)
Difference in Mean Scores Favors	GAE	GFE	GAE	GAE	GFE	GAE	4 Favor GAE 2 Favor GFE
By a Difference of…	0.20	1.36	2.04	0.12	0.36	0.24	GAE = 0.65 GFE = 0.86
Mean of Mean Scores favor…	GAE = 1.09 GFE = 0.94						GAE
Interview Time (*n*=5)	23	25	30	N/A	15	10	20.6 minutes (*SD*=8.02)
Stated Favorite in Interview	Undecided	GAE	GAE	N/A	GFE	Equal	1 Undecided 1 Favor Equal 2 Favor GAE 1 Favor GFE

The qualitative portion of the study comes from the content analysis of the evaluation reports, the post-evaluation interviews, and an open-ended question on the utility questionnaire. The most striking feature from the content analysis of the two reports is the fact that goal achievement evaluators used a predominately quantitative

method of data collection while the goal-free evaluators employed a heavily qualitative method (elaboration of this is discussed later under *Limitations*). For instance, during the interview A-X02 epitomizes many other users' opinions with her response as to why some users found GAE more useful than GFE; A-X02 says:

> I think, more or less, it's just the numbers and percentages; and having the actual numbers you can compare it to. The other one [GFE] is informative but I think it's too subjective and you have to weed through the detail to get to the summary.

One of the purposes of this study is to determine whether there are differences in utility between GAE and GFE in terms of instrumental, conceptual, and persuasive utility. The following table (Table 37) is a brief summary of the interviews according to the three dimensions of evaluation utility for both the GAE and GFE.

Table 37

Summary of Interview Responses by Evaluation Utility Dimension

Evaluation Utility Dimension	GAE	GFE
Instrumental Utility	• For comparing with other organizations • For holding the program accountable for clients' employment outcomes • For generalizing about the program • For generalizing about clients	• For identifying client need • For making programmatic decisions • For justifying transportation assistance • For generalizing about the program's housing-related aspects • For generalizing about clients
Conceptual Utility	• For conceptualizing the program's employment aspects	• For understanding the program's clients & their perspectives
Evaluation Utility Dimension	GAE	GFE
Conceptual Utility	• For conceptualizing the program's employment aspects	• For understanding the program's clients & their perspectives
Persuasive Utility	• For persuading others not to make changes to program funding • For providing information for promotional or informational purposes	• For persuading others to change housing-related aspects, partnerships, communication & feedback with clients, & program delivery

Conclusions

Overall, there appears to be an ever so slight general trend in favor of GAE. This conclusion is primarily based on the utility mean scores which resulted in four of six evaluation users favoring GAE, and the interviews where two users found GAE more useful with only one claiming GFE more useful. However, this conclusion is far from certain. As with all studies, this study has limitations with its ability to observe true effects and there are several factors that have not yet been ruled out which may have influenced the effects or the observations. Furthermore, there are limitations based on the seemingly small effect size, i.e., the lack of observable difference between GAE and GFE. Therefore, there exists too small of a real-world, practically significant difference between GAE and GFE to state that one is definitively the more useful approach.

A conservative overall conclusion of this study is that the null hypothesis is accepted: there is no practically significant difference in evaluation utility between GAE and GFE from the perspective of the evaluation users.

$$H_0 : \text{GAE} = \text{GFE}$$

To be clear about the conclusion of this study, there are, in fact, several differences between these two evaluations and their reports. For example, the GAE report contained a blatant inconsistency in it; the GFE report consisted of more than double the number of pages than that of the GAE; and GAE team collected quantitative data while GFE was mostly qualitative. Nevertheless, it is just not conclusive as to whether or not these distinctions lead to differences in utility that can be meaningfully experienced by evaluation stakeholders or whether these differences are directly related to the particulars

of the goal achievement or goal-free approach rather than some other factor or nuisance variable.

The conclusion of this study is not necessarily surprising as the overall conclusion of Evers' (1980) study was that the evaluation reports' utility did not significantly differ. Scriven's (1974b) view puts his expectations for GFE in perspective. Scriven does not say that he anticipates GFE to replace GBE but rather that GFE's "value will be demonstrated if it sometimes picks up something significant at a cost that makes the discovery worthwhile" (p. 47).

Implications

The major implication of this study is that without more conclusive evidence, it is premature to reject GFE as it remains a legitimate approach for conducting a program evaluation. This also means that, in general terms, there is no evidence to suggest that GAE is more useful to evaluation users than GFE. Prior to this study, GFE was frequently used as a thought experiment, provided hypothetically as a polarity to the more popular goal-based evaluation. However, the conclusion of this study warrants the further use and study of GFE.

Limitations

There are two major limitations of this study. The first is regarding the fact that this study did not control a study variable: evaluator data collection methodology. The second limitation is related to evaluation user attrition.

Evaluator Data Collection Methodology

One of the more obvious limitations of this study is its ability to isolate the specific effects of the GAE and GFE approaches apart from the evaluation teams' chosen data collection methodology. This was a known limitation as the decision was made to allow the two evaluation teams independence in selecting their data collection methodology. The teams, on their own accord, used different strategies for collecting data. During the evaluation user interviews, many comments and preferences of the evaluation users may be related to the distinction between quantitative data and qualitative data collection methods, analysis, and presentation rather than the distinction between goal achievement and goal-free. Below in Figure 9 are four comments offered to illustrate the lack of clarity as to whether the evaluation users' statements actually reflect differences between quantitative and qualitative methods and data, or differences between the evaluation approaches.

Comments like those above might be better attributed to the quantitative-qualitative distinction as opposed to any real difference in usefulness between goal achievement and goal-based evaluations, something for future research.

Attrition of Evaluation Users

The second major limitation of this study was evaluation user attrition. Of the 15 program people identified as evaluation users and asked to read both evaluation reports, six (38%) actually completed the two administrations of the *Evaluation Utility Questionnaire* and five (31%) were interviewed. Attrition of the identified evaluation

users occurred in two forms: (1) attrition of one of the partnering organization, and (2) attrition of individual program staff members.

A-Y04's Response to Interview Question 14:

> "…But when I opened up the GAE, I read the first two pages, and then I always go see what is next, and then I saw all those graphs and I said 'oh, I'll look at this later'."

A-X02's Response to the GAE Report and Interview Question 1:

> "I liked the actual analysis, the numbers and percentages. I like how that was delivered. It seemed a little bit more concrete than the other [GFE]."

A-X01's Response to the GFE Report and Interview Question 2:

> "It [GFE report] gives us a range which is good… if you thought that every situation was going to be the same, this gives you a true sense for a caseworker perspective, in particular. Like, wow, I've got to view it from all kinds of directions. So I think it's very eye opening for developing the range or menu of services you might have to deliver to a client to get a successful outcome. But from a statistical perspective of proving if the program is a success or not, I don't think the study helps us do that. It gives us anecdotal information that we could pop into a funding request; that is very helpful versus just the cold hard facts."

A-X01's Response to the GFE Report and Interview Question 2:

> A-X01 alluded that the program needs to combine the findings and conclusion of both reports to "make it work" for the program, meaning to have a useful evaluation report.

Figure 9. Statements Possibly Referring to Data Collection Methodology Rather than Evaluation Approach

At the onset of this study, three organizations were identified as contributing partners to the program: Agency X, Agency Y, and Agency Z. For reasons unrelated to this study, Agency Z left the partnership midway through the evaluations. Consequently, data on evaluation utility was not collected from the four identified evaluation users who worked for Agency Z.

The second form of evaluation user attrition was the losing of individual staff from the remaining two partnering agencies: Agency X and Agency Y. Two of Agency Y's employees terminated their employment and were not replaced thus leaving nine remaining program staff between Agencies X and Y. Furthermore, one of the Agency X staff terminated after completing both questionnaires but prior to being interviewed. Thus, four of four Agency X staff completed both questionnaires and three of four completed the interviews. While two of five Agency Y staff completed both questionnaires, and two of five from Agency Y completed the interviews. Therefore, 67% (six of nine) of the available evaluation users from Agency X and Agency Y completed both questionnaires, while 56% of (five of nine) evaluation users completed the interviews. Further adding to potential forms of attrition, both A-Y04 and A-Y05 reported that they did not afford significant time to examining the evaluation reports because A-Y had been inactive with regard to the program for several months; therefore, A-Y's inactivity should too be taken into consideration when weighting the importance of this user's responses and opinions. In conclusion, it may be argued that the three (or four) Agency X staff working with the program are the "real" program evaluation users and if so, that means that only approximately three of the original 15 identified evaluation users (20%) ended up being actual users of the evaluations.

The number of program people who read the evaluation reports, and completed the surveys and interviews is considered somewhat realistic given the real-world setting of the study. Pre-identified program people who did not participate in the surveys and/or interviews likely represent what might be expected from a "real" evaluation—not under highly controlled conditions. The evaluation users who were identified yet did not

participate or participated very limitedly likely reflect these users' lack of initiative with regard to the program and/or possibly it shows their agencies' shifting priorities. Consequently, the lack of engagement with the program likely means that these staff persons were either incorrectly identified as evaluation users or were evaluation users but at some point during the study, became non-evaluation users.

In summary, there are two major limitations of this study. The first limitation is related to a potential nuisance variable and its influence on the study's conclusions. Attrition of evaluation users is the second significant limitation of this study.

Recommendations for Further Study

With each published study comes the potential for related studies. These studies can confirm or fail to confirm previous study findings, investigate other aspects in a similar fashion, or build upon existing findings, for example. Bulleted below, in no particular order, are examples of potential studies that may further the study presented in this dissertation.

- An examination of GAE and GFE utility from the perspective of the program's consumers or other downstream impactees, rather than upstream stakeholders.
- An examination of GFE as compared with another well-articulated goal-based evaluation model such as theory-driven evaluation.
- A longitudinal investigation of the two approaches' utility using repeated measures. For example, a re-administration of the same questionnaire

consisting of the semantic differential to the evaluation users at various intervals.

- A post-evaluation examination of changes in the program and program processes that seem attributable to the evaluation(s) without relying on the users' reporting. Instead, the investigator searches for other sources of evidence that the evaluations produced changes in efficiency, effectiveness, and consumer outcomes that can be attributed to information found in the evaluations.
- An examination into pre-evaluation conditions which influence the optimality of GFE.

Chapter Summary

Chapter V summarizes and concludes the study as well as describes some of the study's implications and limitations. The last section of the chapter offers suggestions for continued study of GFE. The findings from this study suggest that GFE deserves further consideration and that additional empirical inquiry into GFE's utility is needed.

Appendix A

Introduction to the Study: A Handout

Dear [__] program administrators and staff:

The Evaluation Center at Western Michigan University has been asked to provide an independent evaluation of the [__] program. Some of you may remember that [the program] was evaluated by Evaluation Center affiliated evaluators in previous years as well.

This year, the evaluators will be conducting two distinct approaches, thus two evaluation teams will be used. The first team's approach is the traditional evaluation approach where the evaluator examines outcomes as they relate to [the program]'s intentions via the stated goals and objectives. This team, called the Goal Achievement Evaluation team, judges the program according to its performance in achieving these goals and objectives. The second evaluation team, called the Goal-Free Evaluation team, specifically avoids learning any information related to the program's stated goals and objectives; instead with this second approach, the evaluators examine all relevant outcomes and judge the program based on the effects that it has on the consumers. This means that all documents and communiqués between program staff and the Goal-Free Evaluation team are screened to prevent the team from learning the stated goals and objectives of the program. The Goal-Achievement Evaluation team consists of three evaluators (A, B, & C) and the Goal-Free Evaluation team also has three evaluators (X, Y, & Z). The evaluation teams will be conducting their evaluations simultaneously.

Each evaluation team will collect information to:

- Establish relevant criteria for judging the program's merit
- Determine relevant standards describing performance at various levels
- Measure or observe outcomes and compare the outcomes to the standards
- Make a conclusion regarding the program's merit

This is not only an evaluation of [the program] but it is also a study of the two program evaluation approaches. Your assessment of each evaluation approach's usefulness is important because you are the users of the evaluations' findings. Specifically, you will be asked to: (i) assist evaluators with their data collection, (ii) read each team's evaluation report, and (iii) complete two one-page questionnaires asking for your opinions regarding the utility of the evaluation reports' findings.

Following the completion of the 24 week evaluations, you will receive a phone call, email, and/or interoffice memo explaining further details regarding the dissemination of the final evaluation reports and questionnaires. You will be given one week to read the first report and complete the corresponding questionnaire. Once you've returned that questionnaire, you will be given the second evaluation report and questionnaire and again you will be given one week to complete and return the second questionnaire. Again, these questionnaires seek your opinions and perceptions regarding the utility of the evaluation reports' findings. You will be allowed to keep the evaluation reports.

During the evaluations, both teams will respect your work and your consumers; additionally, evaluators will attempt to be as discrete as possible when conducting measurements, observations, or interviews. To gather a sufficient amount of data within the limited timeframe of the evaluations and the study, the evaluators may spend a significant amount of time on-site observing activities, administering questionnaires, and/or speaking with program staff and consumers. It is crucial that all correspondence and documentation related to [the program]'s official goals and objectives is screened by the experimenter to ensure that the goal-oriented information is eliminated prior to distributing it to the Goal-Free Evaluation team. Please keep this in mind when speaking or corresponding with any program evaluator.

All of your responses will remain anonymous in the evaluation reports and the study report. No names or identifying information of program staff or consumers will be included in drafts or final evaluation reports or in this study on program evaluation. [the program]'s personnel/staff and the consumers' cooperation with this evaluation and study are completely voluntary. This means that you may refuse to participate or quit at any time during the study without prejudice or penalty. If you have any questions or complaints please (a) speak to the evaluation team; (b) contact the experimenter at: ( ) ; (c) contact the Interdisciplinary Ph.D. in Evaluation program director at: ( ) ; or contact the [program job title] mailto:thomaz.chianca@wmich.eduat: ( ) .

There are four purposes for today's meeting: (i) to introduce you to this study on program evaluation using this handout; (ii) to ask you to identify the people with authority or influence within or over [the program] using the *Identification of Evaluation Users* questionnaire; (iii) to gather demographic information from you; and (iv) to request your assistance in collecting pre-existing documents and archival records from [the program].

Please examine the following list of program documents and archival records. If the document and archival record (or something like it) exists, please provide two copies of each document to the experimenter. The documents and archival records will be used for providing background and contextual information to the evaluators while they prepare for the evaluation.

[Program] **Documents and Archival Records:**
Documents:

- — Program descriptions
- — Program brochures and promotional materials
- — Employee/staff roster (e.g., administrators, supervisors, managers, employees, staff, practitioners, customer services, coordinators, educators, trainers, etc.)
- — Board of Directors roster
- — List of program funders
- — List of partnering organizations or programs
- — Program staff training materials (e.g., curricular-texts, study guides, tests, etc.)
- — Client training materials (e.g., curricular-texts, study guides, tests, etc.)

— Program policy manuals
— Organizational flowcharts
— Program administration and staff job descriptions and responsibilities
— Client flowcharts
— Client eligibility program requirements
— Contracts or agreements between the program and its consumers

Archival Records:

— Prior internal program evaluation reports
— Prior external program evaluation reports
— Program monitoring records (e.g., progress reports, meeting minutes, raw data)
— Program financial records (e.g., annual financial reports, budget status reports)
— Prior grant proposals
— Official correspondence between program and funding agent(s)
— Client intake data (e.g., demographics), tracking data, and demographic data

Please give the experimenter copies of anything else that you think is relevant.

In a week or so, you may receive a reminder email and/or phone call to request: (i) a copy of a position/job description; and (ii) assistance in obtaining copies of job descriptions for other positions.

In approximately 24 weeks, you will be requested to assist me in contacting all of you who were specified in the *Identification of Evaluation Users* questionnaire. You all will be invited to read the evaluation reports, and complete questionnaires on each of the evaluation approaches.

Thank you for your help and cooperation with the evaluations and with this study. If you have further interest in evaluation please see the Evaluation Center website at:

http://www.wmich.edu/evalctr/

Sincerely,

Appendix B

Evaluand Informed Consent Form

Western Michigan University
Department of Interdisciplinary Ph.D. in Evaluation
Principal Investigator:
Student Investigator:

You have been invited to participate in a research project entitle "An Analog Study Comparing Goal-Free Evaluation and Goal Achievement Evaluation Utility." This research is intended to study how you perceive the utility of goal achievement evaluation and goal-free evaluation. This project is __'s dissertation project.

You will also be asked via email, memo, or telephone to provide general information about your job, such as your job title and roles and responsibilities. Following the completion of the program evaluation, you will be asked to read two evaluation reports and respond to a survey questionnaire regarding each. The surveys will ask you to rate and discuss your perceptions on the usefulness of the findings of each evaluation report.

As in all research, there may be unforeseen risks to the participant. If an accidental injury occurs, you should take appropriate emergency measures; however, no compensation or treatment will be made available to you except as otherwise specified in this consent form. Potential risks of participation in this project are that you may be upset by the content of the evaluation reports; and if the evaluators were to report on a limitation that falls within your responsibilities, you may be at risk for psychological and social discomfort. The student investigator is prepared to provide consultation should you become significantly upset and he is prepared to make a referral if you need further consultation about these topics.

One way in which you may benefit from this activity is learning about your program, its operations, its outcomes, and it stakeholders at no financial cost to the program. Additionally, by completing the two questionnaires, you will be contributing to the body of knowledge regarding the two types of evaluation approaches, thus the program as well as other programs may benefit from the knowledge that is gained through this research.

All of the information collected from you is confidential. That means that your name will not appear on any papers on which this information is recorded. The forms will all be coded, and investigator will keep a separate master list with the names of participants and the corresponding code numbers. Once the data are collected and analyzed, the master list will be destroyed. All other forms will be retained for at least three years in a locked file in the principal investigator's office.

You may refuse to participate or quit at any time during the study without prejudice or penalty. If you have any questions or concerns about this study, you may contact either the student investigator at ( ) or the principal investigator at ( ) .

Your signature below indicates that you have read and/or had explained to you the purpose and requirements of the study and that you agree to participate.

_____

Print Name

_____ _____

Signature Date

Consent obtained by: _____ _____

Initials of researcher Date

Appendix C

Goal Achievement Evaluation Evaluator Training Handbook

GOAL ACHIEVEMENT EVALUATION
EVALUATOR TRAINING HANDBOOK

An Introduction to the Handbook

You have agreed to participate with this study as a program evaluator and have been randomly assigned to be on the Goal Achievement Evaluation (GAE) team. In accepting this work assignment, you are agreeing to adhere to certain methodological procedures for collecting information and reporting it back. This handbook accompanies today's four-hour training and provides the following sections to assist you with the evaluation.

I. SETTING OF THE EVALUTION

An independent evaluation firm affiliated with the Evaluation Center is currently contracted to the program which is a cooperative among three organizations operating in the County. The evaluation firm affiliated with the Evaluation Center began its contract to study [the program] in 2004 and is expected to continue its evaluation services.

Previously, the student and principal investigators held a meeting with a key [program] administrator to hear the program plans and evaluation information needs, as well as to allow the administrator to ask questions about the study.

It should be noted that there may be limits to which your team will be given certain information on the program and the study; the rationale for doing so will become increasingly apparent throughout the training.

II. A CONCEPTUAL OVERVIEW OF THE GOAL ACHIEVEMENT EVALUATOR'S ROLE

You will be conducting an outcomes-based summative evaluation assessing absolute merit(s) on various dimensions (or criteria) of the program. The evaluation question your evaluation team seeks to answer is: What is the absolute merit of [the program]?

The evaluator's objectives are as follows:

To collect both descriptive and judgmental information on the evaluand based on the evaluation approach described in the next section.

To summarize the raw data collected and to report it in the format described in a later section.

Your team's evaluation product is a full-length evaluation report.

The following three principles should guide the evaluators and the evaluation:

- Conduct a safe and ethical evaluation
- Maintain fidelity to GAE
- Conduct a sound evaluation and report

Throughout the evaluation, error on the side of behaving ethically first; second, maintain the goal-based nature of the evaluation; and third, ensure that you conduct a quality evaluation and report. If anything is potentially a significant conflict with the nature of GAE, record the conflict and contact the student investigator.

Evaluation Timeline

- Training of student-evaluators: Friday, February 7, 2009
- Student-evaluators are eligible to begin goal achievement and goal-free evaluations: Monday, February 9, 2009
- Student-evaluators bi-weekly debriefings with the student and principal investigators begin after the evaluation training.
- Student-evaluators submit final evaluation report (and logs) approximately July 2009
- Student-evaluators submit time logs approximately July 2009

III. AN INTRODUCTION TO GOAL ACHIEVEMENT EVALUATION

Goal achievement evaluation (GAE) is the process of determining the merit, worth, and/or significance of an evaluand solely according to the evaluand's performance outcomes on stated (or documented) goals and objectives. GAE is goal-based evaluation (GBE) in its most rudimentary form as it is a monitoring system with the sole task of determining whether the evaluand met or is meeting its goals and objectives (Scriven, 1991).[23]

A goal is a "general description of an intended outcome;" whereas an objective is the operationalization of a goal, thus more specific (Scriven, 1991, p. 178). In a GAE, the external evaluator adopts the program's goals and objectives as stated by the program and/or program people and accepts them as criteria of merit (or adapts them only when necessary into criteria). Therefore, GAE is an outcome evaluation where the only outcomes of concern to the evaluator are those directly related to the program's goals or objectives; all other effects and impacts are disregarded as beyond the scope of the evaluation. According to Hezel (in Frechtling, 1995), GBE refers to:

> Cases where programmatic goals have been clearly established during the program's formation, the goals and subsequent concrete and precise objectives become the criteria for measuring the "success" of the program. The goals-based approach is particularly useful for evaluating those aspects of the program that are circumscribed by goals established for the program. In this case, the goals established for the program articulate in a general way the outcomes expected from the program. In turn, the expected outcomes form the basis for the measurement of actual outcomes. (p. 47)

OBJECTIVES-ORIENTED & MANAGEMENT-ORIENTED APPROACH TO EVALUATION

Conceptually, GAE is probably both an objectives-oriented evaluation approach and a management-oriented evaluation approach (Fitzpatrick, Sanders, & Worthen, 2004). It is objectives-oriented in that "the distinguishing feature… is that the purposes of some activity are specified, and then evaluation focuses on the extent to which those purposes are achieved" (Fitzpatrick, Sanders, & Worthen, p. 71). According to the definition of objectives-based evaluation (Christie & Alkin in Mathison, 2005), GBE and GAE are distinctive in their emphasis on the attainment of preordinate objectives; Christie and Alkin state that an objectives-based evaluation "refers to a class of evaluation approaches that centers on the specification of objectives and the measurement of outcomes" (p. 281). Historically, credit for the development of the objectives-oriented evaluation approach has been given to Ralph Tyler (1942); it has been refined over the

---

[23] Throughout this study, GAE is assumed subsumed within GBE.

years by Bloom, et al. (1956), Chen (1990), Cronbach (1963, 1982) Metfessel and Michael (1967), Provus (1971, 1973), Tyler (1949, 1974), Weiss (1972, 1997), and many others.

In addition to the straightforward procedures of the objectives-oriented approach, there are other reasons for using this evaluation approach. According to Fitzpatrick, Sanders, and Worthen (2004):

> The objectives-oriented evaluation approach has caused program directors to reflect about their intentions and to clarify formerly ambiguous generalities about intended outcomes. Discussions of appropriate objectives with the community being served have given objectives-oriented evaluation the appeal of face validity—the program is, after all, merely being held accountable for what its designers said it was going to accomplish, and that is obviously legitimate. The objectives-oriented evaluation approach is one that directly addresses Standard U4, Values Identification, in *The Program Evaluation Standards* (Joint Committee, 1994). Its emphasis on clearly defining outcomes as the basis for judging the program helps evaluators and others to see the value basis for judging the program. (p. 82)

As previously stated, GAE is arguably a management-oriented evaluation approach as its primary emphasis is serving the decision making of evaluation users. Management-oriented approaches have been furthered in publications by Alkin (1969, 1991), Stufflebeam (1968, 1971, 2000), and Wholey (1983), among others. According to Fitzpatrick, Sanders, and Worthen (2004):

> The management-oriented evaluation approach is probably the preferred choice in the eyes of most managers and boards… given the emphasis this approach places on information for decision makers. By attending directly to the informational needs of people who are to use the evaluation, this approach addressed one of the biggest criticisms of evaluation in the 1960s: that it did not provide useful information. (p.95)

The primary argument in favor of GAE is that a program is designed to do certain things in a certain way; hence, a program should be judged according what it is designed to do in comparison with its performance outcomes and to a degree, the outcomes of its consumers (see Scriven (2005) "The Problem of Free Will in Program Evaluation"). In other words, goals "…are not haphazard wishes or incidental desires" (Vedung, 1997, p. 61). Usually, goals and objectives were designed, with careful reflection, according to meeting some relevant need, or needs, of target consumers. They represent the intervention effects desired by the key and most influential involved with the program. Additionally, program managers and staff must monitor their efforts but sometimes they have a limited ability to collect relevant data from relevant sources and issues of credibility are always present with internal evaluations. Therefore, an external goal achievement evaluator offers an independent analysis as to whether or not these

objectives are being met via various data collection methods from multiple sources. The goal achievement evaluator judges the congruence between actual performance outcomes in relation to the satisfaction of the program's goals and objectives.

The specific principles of GAE evaluation are:

1. Identify the evaluand's goals and objectives
2. Operationalize the goals and objectives
3. Measure performance on the goals and objectives
4. Compare the performance to the achievement of the goals and objectives

Thus, the ultimate question for the goal-based evaluator is: Did this intervention achieve these goals?

IV. THE LOGIC OF GOAL ACHIEVEMENT EVALUATION

All evaluations operate under a general logic of evaluation; and GAE is no exception. According to Fournier (1995), there are four basic operations in an evaluation. The first stage, or operation, dictates that if something is to be judged, one must determine the criteria by which to judge it. Second, justifiable standards defining success-failure are constructed. Third, performance is measured on each identified criterion and compared with the standards. Fourth, the data are synthesized to draw evaluative conclusions. This section of the handbook briefly describes the general logic of evaluation in relation to GAE.

Criteria

The criteria of merit are based on the program's current conscious and stated goals and objectives. The goal achievement evaluator accepts the program's goals and objectives as stated. "The strength of goals is that they direct programs by focusing actions on specific outcomes (Friedman, Rothman, & Withers, 2006, p. 202). Quality goals and objectives drive the program and represent the program's intentions thus the evaluator measures/observes the program on these intentions. Therefore, the goals and objectives are the only criteria investigated and the only criteria by which the program is judged (e.g., Scriven's (2007) KEC value (vii) personal, group, and organizational goals/desires).

Clearly articulated goals and objectives require minimal to no adaptation by the goal achievement evaluator. The goal achievement evaluator affords even less attention to the alignment of goals than does the goal-based evaluator. Only when it is necessary does the goal achievement evaluator conduct any type of goal alignment.

Program Goal Alignment

One of the tools of the goal achievement evaluator is the goal alignment. In a GAE, the program's goals and objectives are adapted into criteria and the program's performance outcomes are judged relative to the achievement of these goal-related criteria. The goal achievement evaluator accepts the stated goals as is (whenever possible) and provides minimal emphasis on the quality or appropriateness of the goals and objectives beyond what the program people accept as legitimate. The goal achievement evaluator investigates and judges the program's performance in achieving these stated goals; all other outcomes, effects, and impacts are considered irrelevant. In other words, if the program successfully attains its goals and objectives, it is deemed of merit.

For programs with goals and objectives that are too vague, redundant, incomplete, outdated, or incorrect, it may be necessary for the goal achievement evaluators to work with the program people to clearly articulate the current official program goals and objectives. Typically, the evaluator works with the key program staff and stakeholders until an acceptable workable set of goals and objectives is agreed upon. It is through this process that the evaluators may assist the program by aligning its goal and objectives.

During a goal alignment, the evaluator distinguishes between: (i) original versus current goals and objectives; and (ii) conscious, stated, documented, official, and announced goals versus unconscious, unstated, and implied goals. In a GAE, the usual method for identifying, articulating, or verifying goals and objectives is for the evaluator and program administrators and possibly other key stakeholders to meet and decide which goals and objectives will be included as official. Other sources for information on goals and objectives include program proposals, websites, progress reports, staff training materials, promotional materials, and evaluation reports, among others.

GAE differs from GBE in that the goal-based evaluator likely spends significantly more effort in investigating and aligning these goals and objectives with what the program actually intends to do and does; the goal-based evaluator offers some verification of the relevance of the goals and objectives possibly through interviews, focus groups, and surveys of upstream stakeholders (e.g., program administrators, staff, funders) and program consumers. In a GAE, the purpose of the goal alignment is to adapt goals and objectives to restate them in evaluation-friendly manner, and to use them as criteria. This is done only when necessary. Once a setting of goals and objectives is determined, the evaluator adopts these goals and objectives as the criteria of merit. The evaluator observes the actual level of consumer functioning in areas related to each of these criteria (i.e., goals), and judges the program according to its performance in meeting these goals and objectives.

Standards

Standards describe program performance or quality at various levels on all criteria and subcriteria. In a GAE, the existing standards are adopted by the evaluator; these performance standards are based on the program's performance on the stated goals. If the standards that exist are vague or outdated, they are examined by working with program people to create them and/or update them. Other than consulting with program people, the evaluator may investigate former program benchmarks and standards, the scholarly and professional literature, legal and legislative documents, and so on in determining relevant performance standards.

Outcome Measurement & Comparison to the Standards

The evaluation team collects factual data regarding the program's performances on each identified criterion and subcriterion (i.e., goal, sub-goal, and objective) using qualitative, quantitative, or mixed-methods; the results of the observations are compared to the performance standards. In this study, the evaluation team is required to collect information that is both descriptive (i.e., describing what is/was) and judgmental (i.e., pertaining to merit, worth, and significance determination).

Synthesis

The evaluator combines the program's performances on all identified criteria and subcriteria into an evaluative conclusion or multiple conclusions; and/or the evaluator combines performances on subcriteria into a conclusion on one criterion.

Below are the dos and don'ts of GAE:

Goal Achievement Evaluation

Dos

- ☐ Review program plans and meet with program staff to determine goals and objectives/identify the program's stated goals and objectives. If the program's goals are vague, translate them into measurable objectives.
- ☐ Determine that the goals and objectives are reasonably accurate, current, feasible, and specific.
- ☐ Identify or create standards/benchmarks based on the goals and objectives
- ☐ Identify and select justifiable tools to measure performance (i.e., tools that are reasonable with adequate grounds for use)
- ☐ Measure performance related to goals and objectives

☐	Compare factual information with performance standards/benchmarks on the set goals and objectives and determine the extent to which the program achieved its goals and objectives
☐	Report on the program's performance in relation to its goals and objectives
	Don'ts
☐	Search for, measure, or report on side effects
☐	Conduct a needs assessment

Goal Achievement Evaluation is the process of determining merit by analyzing whether the evaluand met or is meeting its goals and objectives.

A **goal** is a broad or general statement of a program's or intervention's purposes usually constituting longer-term expectations.

An **objective** is a specific, concrete, measurable statement of a program's or intervention's purpose usually constituting shorter-term expectations; it is the operationalization of a goal.

V. EVALUATION REPORTING AND STUDY REQUIREMENTS

Evaluator Supervision

All evaluators are supervised by the IDPE program director (i.e., the principal investigator) and the student investigator during bi-weekly debriefings. Furthermore, the student investigator will make site visits to monitor fidelity to the evaluation approach.

The GAE team leader is responsible for overall direction of the evaluation, including guiding the data collection and analysis, and report writing. The team leader also serves the role of liaison. S/he is responsible for direct communication between his/her evaluation team and evaluand's stakeholders as well as direct communication with the study's investigators.

The student and principal investigators met a key evaluation stakeholder prior to today's training; each evaluation team leader should introduce him/herself to this person, one of the [program] key administrators. Below is her name, title, and contact information.

Name, ___ Director
xxx
Michigan

Telephone: ( )
Fax: ( )

Email:

All attempts should be made to schedule an initial meeting among the goal achievement team, ___ Director, and the student and principal investigators.

Report Format

In order to provide a relatively consistent evaluation and reporting format for the evaluators, the following guidelines on headings and number of (single spaced) pages should be approximated in the evaluation report:

*Executive Summary — 2 pages
Introduction — 3-5 pages
Methodology — 5-10 pages
Findings — 5-10 pages
Conclusion and Recommendations — 3-5 pages
Appendices — No limits on page numbers or content

Your report should reflect a technical evaluation report not an academic paper, thus APA is not required.

Evaluation Report Documentation

Your team must produce a full-length evaluation report in the format described above. In particular, each evaluation team is required to document the evaluation processes and decisions by reporting on the following information as it applies:

— The tools, instruments, and/or procedures used for determining criteria of merit especially during the goal alignment (e.g., interviews, checklists, questionnaires, other measurement instruments)

— The tools, instruments, and/or procedures used for determining standards which describe performance at various levels (e.g., grading rubrics)

— The tools, instruments, and/or procedures used for determining/ weighting importance (e.g., questionnaires, focus groups, interviews, checklists, etc.)

— The tools, instruments, and/or procedures used for measuring and/or observing the evaluand's and/or consumers' performance outcomes

— The tools, instruments, and/or procedures used during synthesis and in determining merit

To document your evaluation activities, the student investigator created three forms for use by the evaluators. (To assist you in understanding how to complete these forms, examples are included):

- Evaluation Team's Log of Potential Threats to the Goal Achievement Nature: A record of any potential breach of the goal achievement nature of the evaluation—anything said, read, or requested—that might lead the evaluation team toward observing a non-stated goal or objective.

- Evaluator's Communication Log: A record between evaluators and program people, and between evaluators and consumers/impactees. Each evaluator must maintain his/her own communication log.

- Evaluator's Time Log: A record describing time spent on evaluation-related activities by name, date, and time. Each evaluator must maintain his/her own time log.

Evaluator's Log of Potential Threats to the Goal Achievement Nature

This log is to be completed whenever a goal achievement evaluator has any indication that s/he received information that might direct them toward considering a non-stated goal, objective, or outcome. One row on the form should be completed whenever an evaluator encounters a possible threat. If multiple evaluators simultaneously experience the same threat, only one form should be completed. If the evaluator is unsure whether something was a threat to the goal achievement nature of the evaluation, the evaluator should complete this log. The evaluator completes this form by indicating the date and time that the potential threat occurred; the evaluators (including himself/herself) who were potentially jeopardized; and where this threat occurred. Next, the evaluator describes the nature of the threat by recording what was said or read that is considered potentially related to a non-goal or objective. Then the evaluator should record the source of the threat and its context. In the next column, the evaluator writes his/her

response to the threat; and lastly, the evaluation team leader determines whether the threat warrants contacting the student or principal investigator and indicates whether it was done. A couple examples of potential threats might include a situation where a program staff member requests a relevant yet unstated goal of which the evaluation team should observe; or on one occasion, several program clients mentioned an outcome worthy of investigation yet is unrelated to a program goal or objective.

Evaluator's Communication Log

This log should be completed by each team member goal achievement team when ever the team member communicates directly with program stakeholders such as program personnel and program consumers. This may include but is not limited to communication via phone, fax, email, face-to-face, mail, memo, text message, and so on. Each evaluator should record the date the communication occurred; and as applicable, the beginning and end time of communication and the total amount of time spent in communication. Next, the evaluator should record who communicated with whom; obviously in a face-to-face conversations and phone calls both are communicating with each other however the evaluator should record who initiated the conversation as applicable. The evaluator should record the mode of communication and, as applicable, where the communication occurred. Lastly, the evaluator should describe the nature of the communication offering a brief summary of what was communicated by both parties.

Evaluator's Time Log

This log should be completed by each team member from the goal achievement team when the team member conducts any activity related to the evaluation. Some of these interactions will overlap with the Evaluator's Communication Log which is expected. Each evaluator should record the date and total time spent on evaluation-related activities for that date. Next, the evaluator should record a summary of any and all evaluation-related activities which includes but is not limited to reading background information, instrument development, communication with program staff and consumers, meetings with team members, data collection and analysis, report writing, and so on.

- The evaluation team's final evaluation report and logs are to be submitted to the student investigator by hardcopy and electronically in Microsoft Word 1997-2003.

- The evaluation team must submit its evaluation report (and logs) approximately July 2009.

- Each evaluator must submit his/her time logs approximately July 2009.

Evaluation Requirements

In agreeing to accept this assignment you are asked to sign an **evaluator contract** stating that you will maintain integrity to the study by adhering to the principles of GAE and will not discuss this evaluation with evaluators from the opposite team; additionally you will be asked to sign a **letter of consent** stating that you understand the study, your role and participation in the study, the potential benefits and risks, and confidentiality. Lastly, you will be asked to complete a short **evaluator demographic questionnaire**. Below is a sample of the contract, letter of consent, and questionnaire.

SAMPLE - Evaluator Contract

An Analog Study Comparing Goal-Free Evaluation
and Goal Achievement Evaluation Utility

I _____ agree to adhere to all of the requirements set forth in this study. Specifically, I will uphold the goal-free or goal achievement approach to which I am assigned. I will not discuss or share information regarding either evaluation approach or the program with anyone who is not affiliated with this study and I will not discuss the study or the evaluation with members of the opposite team.

Failure to abide by these stated restrictions will not only jeopardize the study's fidelity but also may result in academic consequences as deemed appropriate by the IDPE program director.

Print Name: _____

Sign Name: _____

Date: _____

SAMPLE – Informed Consent

Western Michigan University
Department of Interdisciplinary Ph. D in Evaluation
Principal Investigator:
Student Investigator:

You have been invited to participate in a research project entitle "An Analog Study Comparing Goal-Free Evaluation and Goal Achievement Evaluation Utility." This research is intended to study goal achievement evaluation and goal-free evaluation. This project is ___'s dissertation project.

You will be asked to attend a 4-hour training on the assigned evaluation approach with the student investigator

and the principal investigator; additionally, you will also be asked to meet both investigators at a monthly briefing throughout the duration of the evaluation and reporting phases. These meetings will take place at the Evaluation Center.

The first session will consist of the training and will involve completing a questionnaire to gather background information on you. You will also be asked to provide general information about yourself, such as your age, gender, years of research experience, years of evaluation experience, and a rating of your evaluation experience. Your primary task is to conduct an evaluation of a local program and write an evaluation report on it.

As in all research, there may be unforeseen risks to the participant. If an accidental injury occurs, you should take appropriate emergency measures; however, no compensation or treatment will be made available to you except as otherwise specified in this consent form. The main potential risk of participation in this project is based on opportunity costs as you will be asked to dedicate significant amounts of time to this evaluation; additionally, there are social pressures inherent in working within a team including the potential for revealing possible limitations in your evaluation skills. The investigator is prepared to provide consultation should you become significantly upset and he is prepared to make a referral if you need further consultation about these topics.

One way in which you may benefit from this activity is the experience of conducting a program evaluation as well as receiving field experience credit. Additionally, you will be seeing a dissertation being conducted which may serve you in designing and conducting your own dissertation. Lastly, by completing the evaluation report, you will be contributing to the body of knowledge regarding the program being evaluated as well as the two types of evaluation approaches being conducted. Thus the program itself and evaluation scholars may benefit from the knowledge that is gained through this research.

All of the information collected from you is confidential. That means that your name will not appear on any papers on which this information is recorded. All of the forms will be coded, and the investigator will keep a separate master list with the names of participants and the corresponding code numbers. Once the data are collected and analyzed, the master list will be destroyed. All other forms will be retained for at least three years in a locked file in the principal investigator's office.

If you have any questions or concerns about this study, you may contact either the student investigator at ( ) or the principal investigator at ( ) .

Your signature below indicates that you have read and/or had explained to you the purpose and requirements of the study and that you agree to participate.

_____
Print Name

_____ _____
Signature Date

Consent obtained by: _____ _____
Initials of researcher Date

SAMPLE - Evaluator Demographic Questionnaire

1. What is your full name?	
_____	
2. What is your date of birth (month/day/year)?	**3. What is your gender?**
_____/_____/_____	☐ Male ☐ Female
4. How many years of research-specific experience do you have?	**5. How many years of evaluation-specific experience do you have?**
_____	_____
Check the box below, that best completes the following statement:	
6. I would describe my evaluation experience as…	
☐ Extensive ☐ Moderate ☐ Minimal	

VII. AN EXAMPLE OF GAE

The following is a description of a school district's summer school program and an example of how one might develop a GAE based on this program.

Program Description

The Middle School Summer Enrichment Program (MSSEP) serves students who:

(1) Attended 7th or 8th grade the previous school year,

(2) Satisfied At Risk 31a and/or Title 1 criteria in reading and/or math and/or received a final grade of "D" or "F" in a core subject, and

(3) Attended A Middle School, B Middle School, or C Middle School.

MSSEP is funded with Section 31a At Risk and Title I funds. Thus, students attended at no financial cost to their families. The program was short in duration; classes met four hours a day (8:15 through 12:15), four days a week (Monday through Thursday) for 5 weeks (June 14, 2004, through July 15, 2004). The program did not meet on Monday, July 5, 2004, due to the Independence Day holiday.

A program who serves students of the following type according to Title I: Section 31a "at risk" students:

(i) performed poorly academically particularly in English and Math
(ii) score less than Moderate in reading or math and less than Novice in science on the MEAP
(iii) demonstrate atypical behavior or attendance
(iv) have a family history of school failure, incarceration, or substance abuse
(v) are/were victims of abuse or neglect
(vi) are pregnant or teen parents
(vii) come from families that are economically disadvantaged (eligible for free or reduced lunch); historically, the MSSEP student population consists of more males than females and more African-American students than Caucasian

The core activity of the program was an eight-step *Issue Investigation* process. Students completed the process twice, once as an entire class and then as individuals or in small groups. The first process was guided step-by-step by the instructor. The second time students worked independently on the Issue Investigation. Students used various resources, including daily newspapers, to identify and research their chosen issues.

Students also worked in small teams on *Brain Hurricane Creativity Kits*. These activities sought to develop students' teamwork skills and challenge them to think in new ways. All the kit activities incorporated the Michigan Curriculum Benchmarks and Standards.

MSSEP pursued nine goals based on three values:

Values	Goals
Giving students opportunities to explore	● Improve student writing, thinking, and problem-solving skills. ● Provide students with new experiences. ● Help students make connections between core content areas.
Recognizing individual learning patterns in students	● Improve student writing, thinking, and problem-solving skills. ● Provide students with new experiences.

	●Help students make connections between core content areas.
Maintaining a climate that enables students to pursue program goals	● Improve student writing, thinking, and problem-solving skills. ● Provide students with new experiences. ● Help students make connections between core content areas.

MSSEP Evaluation Design

Purpose and Clients

The purpose of the evaluation was to assist in:

(1) Data collection by providing the school district with appropriate instruments

(2) Developing an ongoing evaluation plan

(3) Designing follow-up processes to track intermediate and long-term effects of the program on students

(4) Formulating recommendations for program improvement, future evaluations, and data collection and utilization

Audiences for this evaluation included the Title I staff at the district level; program staff; students in the program and their families; the larger district administration including the school board; and the community at large. This evaluation report wais presented to the Title I staff and was presented at a televised school board meeting, aiding dissemination of the report's findings within the community.

EXERCISE (15 minutes): Working with your team members and given the above description, how might you begin designing the evaluation and data collection? Specifically, how would you measure/observe program and consumer performance on the above goals and objectives?

MSSEP Evaluators' *Evaluation Questions*

This evaluation was guided by certain questions, which were based on the goals and values of the program identified by district staff. A comprehensive evaluation of the program would address its merit, worth, and significance while looking at the needs of the program recipients (the students). For our purposes we relied on district staff's determination of these students' needs as presented in the goals of the program. Our evaluation questions were based on goals identified from program materials. The questions provided the basis for the survey instruments developed by the evaluation team.

The basic question addressed by the evaluation is: What are the students gaining from the summer enrichment program? This question can be broken down into different categories by looking at the program's goals and the values underlying those goals (see above). Three main goals for the program were identified based on underlying values which derived from the characteristics of a middle school as defined by the Association for Supervision and Curriculum Development (ASCD).

MSSEP Evaluators' Data Collection & Design

After, the program was introduced to the evaluation team in a meeting with Title I staff, during which past program activities and achievements and former forms of assessment were discussed, an evaluation plan was drafted and accepted by the Title I staff at the district.

- Pre- and post-survey/interview instruments were developed to assess the program's immediate effects on students' attitudes toward school, their future, and their behaviors.

- Pre- and post-surveys to program staff were developed to assess their attitudes about and experience with aspects of the program as well as observed changes in student performance.

- A single survey was administered to parents at the end of the program to provide insight into their perceptions about their children's improvements, interest in school, and their behaviors in communicating about school.

- Immediate academic effects were measured by standardized pre- and post- tests in math and reading using the EdPerformance – Standards-Based Adaptive Measurement (SAM) – test.

- The evaluation team was provided prior year student report cards. Due to the program's short duration this evaluation report includes recommendations for

creating an ongoing evaluation framework linking data from the regular academic year to data from the MSSEP. By doing so, program effects can be tracked and the program can be evaluated, over a longer time period.

District staff collected data from students and provided staff with the survey instruments. Students gave surveys to their parents to be returned for extra credit. Survey results were entered into a data file by district staff and analyzed by the evaluation team. This arrangement helped lower the cost of the evaluation while still providing accurate and timely data. A more extensive evaluation would entail observations by evaluators of program activities and a comparison of program costs with the cost of similar programs. Due to the relatively low cost and short duration of the MSSEP the size of the evaluation budget and activities was appropriate.

VIII. PROGRAM DOCUMENTS AND ARCHIVAL RECORDS (Attached)

The next part of this training involves the examination of the program's documents and archives to determine the criteria of merit and develop an evaluation strategy or plan.

EXERCISE (15-20 minutes)
Begin by read the program documents and archives. While you read, jot down answers to the following questions:

- What are potential criteria of merit (i.e., goals and objectives)?
- What are potential performance standards?
- What or who are potential for sources of information?
- What ethical issues are of particular concern in this evaluation?
- What questions need to be answered before finalizing an evaluation plan?
- How can we immediately delegate tasks among the three evaluators on our team?

A housing and employment retention program collaboratively sponsored by Agencies X, Y, and Z.

The program attempts to dissolve barriers between the stand-alone housing and employment "silos." Given, an isolated service delivery system can never garner the duplicate mainstream resources required to alleviate poverty and its debilitating symptoms such as homelessness. The program is a wrap-around service delivery model clearly demonstrating the interrelatedness of stable housing to stable employment, and vice versa.

The program focuses on bridging gaps in mainstream programming contributing to chronic unemployment and homelessness. Many programs and services regularly operate in isolation from one another creating layers of conflicting requirements. Often unwittingly penalizing persons in need as they strive to navigate multiple systems thereby limiting positive outcomes.

This pioneering program design has established the viability of housing assistance as a support for persons moving from welfare to work. It also proves the efficiency of a multi-collaborative wrap around service delivery model. We are pleased to report many chronically unemployed and homeless families sustained or increased their earning potential while in the program. Allowing a greater chance of maintaining both employment and housing upon program exit.

Purpose: The program provides housing stabilization, employment retention and job development services. Many participants are able to open bank accounts and plan for their financial futures for the first time.

The Program Team is a collaboration of 18 people between Agencies, X, Y, and Z. The team developed its program to assist the people they served who were trying to leave welfare but were continually failing due to instability in either housing and employment or both.

The goals of the program include: preventing homelessness, preventing unemployment, encouraging career development or earnings, promoting self-sufficiency and maximizing access to community resources.

An evaluation of the program's impact was completed in 2006. After reviewing data on 70 clients who were subjects in the study that lasted almost a year, 87 percent retained their housing and 83 percent retained employment. The program does make a difference!

Below is a list of the members:

[The Program]: Supports program participants in their efforts to prevent future episodes of unemployment and homelessness. Participants receive help finding employment and stable housing. Additional services such as rental assistance, financial management, childcare, transportation, and 24-hour problem solving assistance. Cosponsored by Agency X. and the county Family Independence agency. Since the program's inception in 2001, 120 participants and their families have received wraparound supportive services with 63% of those who have been followed for a full year retaining stable housing and income at program exit. This is generally twice the rate of successful employment retention compared to the area's traditional Work First program model.

EVALUATOR'S LOG OF POTENTIAL THREATS TO THE GOAL ACHIEVEMENT NATURE

Date & Time of Incident		
Who were the evaluator(s) involved?		
Where did the incident occur?		
What was said or read that is potentially threatening to the goal achievement nature of the evaluation?		
Who was the source and what was the context of the statement or writing?		
What was the evaluators' response to the threat?		
Was it reported to one of the investigators?	No ☐	Yes ☐ Date: Investigator:

EVALUATOR'S COMMUNICATION LOG

Date	
Start time	
End time	
Total time (hrs & mins)	
Who communicated with whom?	
How did you communicate?	
Where did the communication occur?	
Describe the nature of the communication.	

EVALUATOR'S TIME LOG

Evaluator's Name:

_____

Date		
Total Time Spent on the Evaluation (hrs & min)		
Describe your daily evaluation-related activities.		

Appendix D

Goal-Free Evaluation Evaluator Training Handbook

GOAL-FREE EVALUATION
EVALUATOR TRAINING HANDBOOK

An Introduction to the Handbook

You have agreed to participate with this study as a program evaluator and have been randomly assigned to be on the Goal-Free Evaluation (GFE) team. In accepting this work assignment, you are agreeing to adhere to certain methodological procedures for collecting information and reporting it back. This handbook accompanies today's four-hour training and provides the following sections to assist you with the evaluation.

I. SETTING OF THE EVALUTION

An independent evaluation firm affiliated with the Evaluation Center is currently contracted to the program which is a cooperative among three organizations operating in the County. The evaluation firm affiliated with the Evaluation Center began its contract to study [the program] in 2004 and is expected to continue its evaluation services.

Previously, the student and principal investigators held a meeting with a key [program] administrator to hear the program plans and evaluation information needs, as well as to allow the administrator to ask questions about the study.

It should be noted that there may be limits to which your team will be given certain information on the program and the study; the rationale for doing so will become increasingly apparent throughout the training.

II. A CONCEPTUAL OVERVIEW OF THE GOAL ACHIEVEMENT EVALUATOR'S ROLE

You will be conducting an outcomes-based summative evaluation assessing absolute merit(s) on various dimensions (or criteria) of the program. The evaluation question your evaluation team seeks to answer is: What is the absolute merit of [the program]?

The evaluator's objectives are as follows:

> To collect both descriptive and judgmental information on the evaluand based on the evaluation approach described in the next section.
>
> To summarize the raw data collected and to report it in the format described in a later section.

Your team's evaluation product is a full-length evaluation report.

The following three principles should guide the evaluators and the evaluation:

- Conduct a safe and ethical evaluation
- Maintain fidelity to GFE
- Conduct a sound evaluation and report

Throughout the evaluation, error on the side of behaving ethically first; second, maintain the goal-based nature of the evaluation; and third, ensure that you conduct a quality evaluation and report. If anything is potentially a significant conflict with the nature of GFE, record the conflict and contact the student investigator.

Evaluation Timeline

- Training of student-evaluators: Friday, February 7, 2009
- Student-evaluators are eligible to begin goal achievement and goal-free evaluations: Monday, February 9, 2009
- Student-evaluators bi-weekly debriefings with the student and principal investigators begin after the evaluation training.
- Student-evaluators submit final evaluation report (and logs) approximately July 2009
- Student-evaluators submit time logs approximately July 2009

III. AN INTRODUCTION TO GOAL-FREE EVALUATION

Goal-free evaluation is the process of determining the merit, worth, and/or significance of an evaluand conducted partially or fully independent of the stated (or implied) goals and objectives of the evaluand. According to the Program Evaluation Standards (PES), a GFE is an "evaluation of outcomes in which the evaluator functions without knowledge of the purposes or goals" (Joint Committee, 1994, p. 206). In a GFE, the evaluator intentionally avoids learning the official or stated goals and objectives of the evaluation client and stakeholders; rather, the evaluator observes and measures the actual outcomes founded in logical and definitional premises and on the program's performance in meeting the consumers' needs.

CONSUMER-ORIENTED APPROACH TO EVALUATION

Conceptually, GFE is a consumer-oriented evaluation approach (Fitzpatrick, Sanders, & Worthen, 2004). It is consumer-oriented in that its emphasis is on the consumers' needs rather than the program or staff needs. The justification for this is that the primary needs of the consumer are the *raison d'être*, or the "rationale for the existence" of the service deliverers and delivery systems (Altschuld & Witkin, 2000, p. 10). Scriven (1967, 1991), and his recognition of the Consumers Union's consumer-oriented product evaluation, has been the main contributor to the consumer-oriented evaluation approach (Fitzpatrick, Sanders, & Worthen, 2004). According to Fitzpatrick, Sanders and Worthen:

> The rationale for goal-free evaluation can be summarized as follows: First, goals should not be taken as given; like anything else, they should be evaluated. Further, goals are generally little more than rhetoric and seldom reveal the real objectives of the project or changes in intent. In addition many important program outcomes are not included in the list of original program goals or objectives. (p. 84)

The argument in favor of GFE is the prevention of tunnel-vision, a perceptual blindness that biases the evaluator and contaminates his/her judgment of the evaluand's "true" outcomes and "true" merit. This blindness is typically present during the establishment of the criteria of merit and during the measurement and observation of the evaluand's performance. Tunnel-vision toward goal-oriented effects also influences program administrators and staff as well (Evers, 1980). Scriven (1972) says that the tunnel-vision is not a matter of honesty but rather failing to see the forest for the trees. Therefore, the independent evaluator's ignorance to specific goals is considered a strength and the GFE design is developed to maximize this independence.

Stated goals and objectives are unnecessary noise for the external evaluator; yet, are essential for the internal evaluator and program managers in monitoring the program's efforts (Scriven, 1972). If one accepts the definition of evaluation as the

systematic determination of merit and since the program was designed to meet some relevant needs of a target consumer, the evaluator sees that the program's intentions are not required in determining what makes the program good or bad. In fact, goals and objectives often prevent the recognition of relevant unintended positive and negative side effects and side impacts. Thus, the goal-free evaluator attempts to observe/measure all possible areas for relevant actual outcomes while being screened from stated goal-oriented information.

In theory, if the program is doing what it intends then many of the criteria identified by the evaluator should match the program's goals and the outcomes of which program is attempting to produce. Patton (1997) recommends using GFE as a method of program goal alignment as he states: "[a] result of goal-free evaluation is a statement of goals… a statement of operating goals becomes it outcome" (p. 182). However Scriven discourages the determination of the "true" program goals as an outcome of GFE because it takes the focus away from the needs of the consumers and back to the goals of management. Rather, according to Patton (1997), Scriven says that GFE's outcome is the determination of merit with an emphasis on the satisfaction of consumers' needs. Therefore, attempting to extrapolate the program's actual goals and objectives is considered beyond the scope of GFE in this study.

Since you have been assigned to the GFE team, you will be prohibited from learning information that the program and staff pose as intentions, goals, or objectives. This goal-and objective-oriented information is often found in program websites, promotional material, program proposals, progress reports, staff training materials, evaluation reports, and by communicating with program administrators, managers, staff, funders, and clients. Thus, action is taken to prevent you from learning this information. It should be noted that simply learning the names of the cooperating organizations, may lead one to infer the program's general aims; however, identifying the program's specifically stated objectives is not so obvious (for this reason, Scriven (1991) points out that GFE might better be called objective-free evaluation). Furthermore, even if someone accidentally tells you a goal or objective, it does not mean that s/he accurately stated it.

The specific principles of GFE evaluation are:

1. Identify relevant effects of which to examine without referencing goals and objectives

2. Identify what occurred without the prompting of evaluand goals and objectives

3. Determine if what occurred can logically be attributed to the intervention

4. Determine the degree to which the effect(s) are positive or negative

Thus, the ultimate question for the goal-free evaluator is: What occurred that can be attributed to the intervention?

IV. THE LOGIC OF GOAL-FREE EVALUATION

All evaluations operate under a general logic of evaluation; and GFE is no exception. According to Fournier (1995), there are four basic operations in an evaluation. The first stage, or operation, dictates that if something is to be judged, one must determine the criteria by which to judge it. Second, justifiable standards defining success-failure are constructed. Third, performance is measured on each identified criterion and compared with the standards. Fourth, the data is synthesized to draw evaluative conclusions. This section of the handbook briefly describes the general logic of evaluation in relation to GFE.

Criteria

1. The criteria of merit are primarily based on three values: (i) logical premises, (ii) definitional premises, and (iii) the consumers' needs. Logical premises are those founded on reason or rational thought; examples may include: safety, ethics, law, professionalism, etc. Definitional premises are those based on what it means to be a good one of its type. For example, a good cordless electric razor must, by definition, be able to effectively cut whiskers and have a satisfactory re-chargeable battery. Criteria based on logical and definitional premises are identified via numerous means such as scholarly and professional literature, expert judgment, legal and legislative documents, certain program documents and archival records, critical competitors' program documents and reports, and various checklists (e.g., KEC, PES), among others. The third primary value in a GFE is founded on the program's meeting of the consumers' relevant needs. To determine which needs are the relevant needs, the goal-free evaluator may conduct a needs assessment. For example, via a needs assessment, the evaluation team determines that a program designed to teach wheelchair tennis to children with disabilities have consumers who "need" transportation to and from the tennis facility regardless whether or not it is a goal or objective of the program.

   Consumers' Needs Assessment:

   A tool of the goal-free evaluator is the needs assessment. According to Davidson (2005) and Scriven (2007), meeting the relevant needs of the consumer represents one of several criteria of merit or values. Each relevant consumer need, identified via the needs assessment, may represent a potential evaluation criterion or subcriterion.

   During a needs assessment, the evaluator focuses on the consumers' needs (however secondarily, the evaluator may examine the program or upstream stakeholder needs). The evaluator distinguishes between consumers': (i) needs versus wants, (ii) met versus unmet needs, (iii) treatment versus performance needs, and (iv) conscious versus unconscious needs (See Davidson, 2005). Next, the evaluator determines

which of these needs (if any) are particularly relevant or critical by questioning the consumers (and possibly other stakeholders), and measuring/observing the consumers' performance to determine what the program is actually doing or did. Further justification for the inclusion of a particular need is usually based on examining logical and definitional premises, consulting the literature, examining similar programs' reports, and asking downstream impactees, among others.

Substance abuse treatment group by Program X as a simplified illustration:

Since the goal-free evaluator is aware that the evaluand is a substance abuse treatment group they are able to begin postulating relevant criteria of merit by knowing a bit about what it means to be a substance abuse group, by reviewing certain Alcoholics Anonymous documents and archival records, by reviewing relevant checklists (KEC, PES, AA), by conducting a needs assessment to determine the consumers' needs, by examining the program's actual outcomes to extract possible criteria, and so on. The goal-free evaluator avoids learning the stated goals and objectives of the program as stated by the program people or in program documents; rather, s/he offers a perspective of what the program is actually doing via the evaluator's observations of the program. Merit is determined according to the program's performance in meeting or satisfying relevant criteria.

The data from the needs assessment fit into the overall evaluation picture in that observations on the criterion "meeting the relevant needs of consumers" and all its subcriteria (i.e., needs) are used to judge the program in relation to the program's performance in meeting these relevant needs and producing satisfactory consumer functioning. The evaluator determines absolute merit by observing actual consumer functioning on each of the needs (now deemed subcriteria) and uses a logical method of synthesizing the data on each subcriterion into an overall judgment on the criterion "meeting relevant needs of consumers." The needs-based criterion is logically synthesized with the other criteria (those identified by other means such as a literature review, document analysis, etc.) to make judgments of the program across all identified criteria. It should be noted that some of these needs are logically deemed criteria and not subcriteria under "meets consumers' needs"; for example, in a juvenile detention facility, the juveniles' need for "safety" might be considered a criterion rather than a subcriterion falling under "meets consumers' needs."

Standards

Standards describe program performance or quality at various levels on all criteria and subcriteria. In a GFE, relevant standards can be identified via scholarly and professional literature, certain program documents and reports, legal and legislative documents, standards established by other similar programs or interventions, various checklists, among others. Standards are set by comparing the program's actual performance outcomes in meeting the consumers' relevant needs against what is required to reach and exceed satisfactory functioning. Rather than consulting with program administrators, the goal-free evaluator determines performance standards by examining the congruence between actual outcomes and satisfactorily meeting the consumers' relevant needs while considering the program's contextual and resource constraints.

Outcome Measurement & Comparison to the Standards

The evaluation team collects factual data regarding the program's performances on each identified criterion and subcriterion as well as is open to other effects that may appear while observing evaluand outcomes. The team uses qualitative, quantitative, or mixed-methods in collecting the data and determines whether the effects are positive or negative. The goal-free evaluation team then determines whether effects can be reasonably attributed to the program and compares them to the performance standards. In this study, the evaluation team is required to collect information that is both descriptive (i.e., describing what is/was) and judgmental (i.e., pertaining to merit, worth, and significance determination).

Synthesis

The evaluator combines the program's performances on all identified criteria and subcriteria into an evaluative conclusion or multiple conclusions; and/or the evaluator combines performances on subcriteria into a conclusion on one criterion.

Below are the dos and don'ts of GFE:

Goal-Free Evaluation	
Dos	
☐	Identify and use a screener (i.e., an intermediary who ensures that no goal- or objective- based information is communicated to the goal-free evaluators)
☐	Refer all communiqués to screener and involve the screener throughout the evaluation to protect from potential contamination
☐	Have all written material screened for references to program goals or objectives prior to evaluator receipt
☐	Advise all program people of goal-free nature and the parameters of goal-free evaluation. Ensure that they understand they are not to relay goal/objective-related information.

☐	Stop program staff if they begin talking about goal-oriented information
☐	Identify potential areas in which to search for effects (in part through a needs assessment) and use these as the basis for criteria to be measured
☐	Identify and select justifiable tools to measure performance and actual effects (i.e., tools that are reasonable with adequate grounds for use)
☐	Measure performance and actual effects/experience (observe the program as is)
☐	Compare factual information about the program effects/experiences with pre-identified needs to assess the program's impact on consumer needs
☐	Offer a profile of the positive and negative effects
	Don'ts
☐	Communicate with program staff regarding goals or objectives
☐	Attempt to find stated goals and objectives

Goal-Free Evaluation is the process determining merit with the evaluator maintaining partial or full independence from the stated (or implied) goals and objectives of those who design, produce, or implement the evaluand.

A **goal** is a broad or general statement of a program's or intervention's purposes usually constituting longer-term expectations.

An **objective** is a specific, concrete, measurable statement of a program's or intervention's purpose usually constituting shorter-term expectations; it is the operationalization of a goal.

V. EVALUATION REPORTING AND STUDY REQUIREMENTS

Evaluator Supervision

All evaluators are supervised by the IDPE program director (i.e., the principal investigator) and the student investigator during bi-weekly debriefings. Furthermore, the student investigator will make site visits to monitor fidelity to the evaluation approach.

The GFE team leader is responsible for overall direction of the evaluation, including guiding the data collection and analysis, and report writing. The team leader also serves the role of liaison. S/he is responsible for direct communication between his/her

evaluation team and evaluand's stakeholders as well as direct communication with the study's investigators.

The student and principal investigators met a key evaluation stakeholder prior to today's training; each evaluation team leader should introduce him/herself to this person, one of the program's key administrators. Below is her name, title, and contact information.

Name, ___ Director
xxx
Michigan

Telephone: ( )
Fax: ( )

Email:

All attempts should be made to schedule an initial meeting among the goal achievement team, ___ Director, and the student and principal investigators.

Report Format

In order to provide a relatively consistent evaluation and reporting format for the evaluators, the following guidelines on headings and number of (single spaced) pages should be approximated in the evaluation report:

*Executive Summary — 2 pages
Introduction — 3-5 pages
Methodology — 5-10 pages
Findings — 5-10 pages
Conclusion and Recommendations — 3-5 pages
Appendices — No limits on page numbers or content

Your report should reflect a technical evaluation report not an academic paper, thus APA is not required.

Evaluation Report Documentation

Your team must produce a full-length evaluation report in the format described above. In particular, each evaluation team is required to document the evaluation processes and decisions by reporting on the following information as it applies:

— The tools, instruments, and/or procedures used for determining criteria of merit especially during the needs assessment (e.g., interviews, checklists, questionnaires, other measurement instruments)

— The tools, instruments, and/or procedures used for determining standards which describe performance at various levels (e.g., grading rubrics)

— The tools, instruments, and/or procedures used for determining/weighting importance (e.g., questionnaires, focus groups, interviews, checklists, etc.)

— The tools, instruments, and/or procedures used for measuring and/or observing the evaluand's and/or consumers' performance outcomes

— The tools, instruments, and/or procedures used during synthesis and in determining merit

To document your evaluation activities, the student investigator created three forms for use by the evaluators. (To assist you in understanding how to complete these forms, examples are included):

- Evaluation Team's Log of Potential Threats to the Goal-Free Nature: A record of any potential breach of the goal-free nature of the evaluation—anything said or read—that might be a stated goal or objective.

- Evaluator's Communication Log: A record between evaluators and program people, and between evaluators and consumers/impactees

- Evaluator's Time Log: A record describing time spent on evaluation-related activities by name, date, and time.

Evaluator's Log of Potential Threats to the Goal-Free Nature

This log is to be completed whenever a goal-free evaluator has any indication that s/he received information that is possibly considered related to a program goal or objective. One row on the form should be completed whenever an evaluator encounters a possible threat. If multiple evaluators simultaneously experience the same threat, only one form should be completed. If the evaluator is unsure whether something was a threat to the goal-free nature of the evaluation, the evaluator should complete this log. The evaluator completes this form by indicating the date and time that the potential threat occurred;

the evaluators (including himself/herself) who were potentially jeopardized; and where this threat occurred. Next, the evaluator describes the nature of the threat by recording what was said or read that is considered potentially related to a goal or objective. Then the evaluator should record the source of the threat and its context. In the next column, the evaluator writes his/her response to the threat; and lastly, the evaluation team leader determines whether the threat warrants contacting the student or principal investigator and indicates whether it was done. A couple examples of potential threats might include a situation where the evaluator documents that s/he overheard program staff members naming and discussing a specific objective or the evaluator began reading a document given to him/her by a program consumer during an interview.

Evaluator's Communication Log

This log should be completed by each team member from the goal-free team when ever the team member communicates directly with program stakeholders such as program personnel and program consumers. This may include but is not limited to communication via phone, fax, email, face-to-face, mail, memo, text message, and so on. Each evaluator should record the date the communication occurred; and as applicable, the beginning and end time of communication and the total amount of time spent in communication. Next, the evaluator should record who communicated with whom; obviously in a face-to-face conversations and phone calls both are communicating with each other however the evaluator should record who initiated the conversation as applicable. The evaluator should record the mode of communication and as applicable where the communication occurred. Lastly, the evaluator should describe the nature of the communication offering a brief summary of what was communicated by both parties.

Evaluator's Time Log

This log should be completed by each team member from the goal-free team when the team member conducts any activity related to the evaluation. Some of these interactions will overlap with the Evaluator's Communication Log which is expected. Each evaluator should record the date and total time spent on evaluation-related activities for that date. Next, the evaluator should record a summary of any and all evaluation-related activities which includes but is not limited to reading background information, instrument development, communication with program staff and consumers, meetings with team members, data collection and analysis, report writing, and so on.

- The evaluation team's final evaluation report and logs are to be submitted to the student investigator by hardcopy and electronically in Microsoft Word 1997-2003.

- The evaluation team must submit its evaluation report (and logs) approximately July 2009.

- Each evaluator must submit his/her time logs approximately July 2009.

Evaluation Requirements

In agreeing to accept this assignment you are asked to sign an **evaluator contract** stating that you will maintain integrity to the study by adhering to the principles of GFE and will not discuss this evaluation with evaluators from the opposite team; additionally you will be asked to sign a **letter of consent** stating that you understand the study, your role and participation in the study, the potential benefits and risks, and confidentiality. Lastly, you will be asked to complete a short **evaluator demographic questionnaire**. Below is a sample of the contract, letter of consent, and questionnaire.

SAMPLE - Evaluator Contract

An Analog Study Comparing Goal-Free Evaluation
and Goal Achievement Evaluation Utility

I _____ agree to adhere to all of the requirements set forth in this study. Specifically, I will uphold the goal-free or goal achievement approach to which I am assigned. I will not discuss or share information regarding either evaluation approach or the program with anyone who is not affiliated with this study and I will not discuss the study or the evaluation with members of the opposite team.

Failure to abide by these stated restrictions will not only jeopardize the study's fidelity but also may result in academic consequences as deemed appropriate by the IDPE program director.

Print Name: _____

Sign Name: _____

Date: _____

SAMPLE – Informed Consent

Western Michigan University
Department of Interdisciplinary Ph.D. in Evaluation
Principal Investigator:
Student Investigator:

You have been invited to participate in a research project entitle "An Analog Study Comparing Goal-Free Evaluation and Goal Achievement Evaluation Utility." This research is intended to study goal achievement evaluation and goal-free evaluation. This project is ___'s dissertation project.

You will be asked to attend a 4-hour training on the assigned evaluation approach with the student investigator and the principal investigator; additionally, you will also be asked to meet both investigators at a monthly briefing throughout the duration of the evaluation and reporting phases. These meetings will take place at the Evaluation Center.

The first session will consist of the training and will involve completing a questionnaire to gather background information on you. You will also be asked to provide general information about yourself, such as your age, gender, years of research experience, years of evaluation experience, and a rating of your evaluation experience. Your primary task is to conduct an evaluation of a local program and write an evaluation report on it.

As in all research, there may be unforeseen risks to the participant. If an accidental injury occurs, you should take appropriate emergency measures; however, no compensation or treatment will be made available to you except as otherwise specified in this consent form. The main potential risk of participation in this project is based on opportunity costs as you will be asked to dedicate significant amounts of time to this evaluation; additionally, there are social pressures inherent in working within a team including the potential for revealing possible limitations in your evaluation skills. The investigator is prepared to provide consultation should you become significantly upset and he is prepared to make a referral if you need further consultation about these topics.

One way in which you may benefit from this activity is the experience of conducting a program evaluation as well as receiving field experience credit. Additionally, you will be seeing a dissertation being conducted which may serve you in designing and conducting your own dissertation. Lastly, by completing the evaluation report, you will be contributing to the body of knowledge regarding the program being evaluated as well as the two types of evaluation approaches being conducted. Thus the program itself and evaluation scholars may benefit from the knowledge that is gained through this research.

All of the information collected from you is confidential. That means that your name will not appear on any papers on which this information is recorded. All of the forms will be coded, and the investigator will keep a separate master list with the names of participants and the corresponding code numbers. Once the data are collected and analyzed, the master list will be destroyed. All other forms will be retained for at least three years in a locked file in the principal investigator's office.

If you have any questions or concerns about this study, you may contact either the student investigator at ( ) or the principal investigator at ( ) .

Your signature below indicates that you have read and/or had explained to you the purpose and requirements of the study and that you agree to participate.

_____
Print Name

_____ _____
Signature Date

Consent obtained by: _____ _____
Initials of researcher Date

SAMPLE - Evaluator Demographic Questionnaire

1. What is your full name?	
_____	
2. What is your date of birth (month/day/year)?	**3. What is your gender?**
_____/_____/_____	☐ Male ☐ Female
4. How many years of research-specific experience do you have?	**5. How many years of evaluation-specific experience do you have?**
_____	_____
Check the box below, that best completes the following statement:	
6. I would describe my evaluation experience as…	
☐ Extensive ☐ Moderate ☐ Minimal	

VII. AN EXAMPLE OF GFE

The following is a description of a school district's summer school program and an example of how one might develop a GAE based on this program.

The Middle School Summer Enrichment Program (MSSEP) served students who:

(1) Attended 7th or 8th grade the previous school year,

(2) Satisfied At Risk 31a and/or Title 1 criteria in reading and/or math and/or received a final grade of "D" or "F" in a core subject, and

(3) Attended A Middle School, B Middle School, or C Middle School.

MSSEP is funded with Section 31a At Risk and Title I funds. Thus, students attended at no financial cost to their families. The program was short in duration; classes met four hours a day (8:15 through 12:15), four days a week (Monday through Thursday) for 5 weeks (June 14, 2004, through July 15, 2004). The program did not meet on Monday, July 5, 2004, due to the Independence Day holiday.

A program who serves students of the following type according to Title I: Section 31a "at risk" students:

(i) performed poorly academically particularly in English and Math
(ii) score less than Moderate in reading or math and less than Novice in science on the MEAP
(iii) demonstrate atypical behavior or attendance
(iv) have a family history of school failure, incarceration, or substance abuse
(v) are/were victims of abuse or neglect
(vi) are pregnant or teen parents
(vii) come from families that are economically disadvantaged (eligible for free or reduced lunch); historically, the MSSEP student population consists of more males than females and more African-American students than Caucasian

MSSEP Evaluation Design

Purpose and Clients

The purpose of the evaluation was to assist in:

(1) Data collection by providing the school district with appropriate instruments

(2) Developing an ongoing evaluation plan

(3) Designing follow-up processes to track intermediate and long-term effects of the program on students

(4) Formulating recommendations for program improvement, future evaluations, and data collection and utilization

Audiences for this evaluation included the Title I staff at the district level; program staff; students in the program and their families; the larger district administration including the school board; and the community at large. This evaluation report wais presented to the Title I staff and was presented at a televised school board meeting, aiding dissemination of the report's findings within the community.

EXERCISE (15 minutes): Working with your team members and given the above description, how might you begin designing the evaluation and data collection? What are potential criteria of merit of which you might observe program effects and outcomes? Specifically, how would you measure/observe program and consumer performance?

MSSEP Evaluators' *Evaluation Questions*

This evaluation was guided by certain questions, which were based on logic and the needs of the consumer as identified by the evaluation team. A comprehensive evaluation of the program would address its merit, worth, and significance while looking at the needs of the program recipients (the students). For our purposes we relied on students' needs as presented by students and staff during interviews and classroom observations. The interviews provided justification for several of the survey questions developed by the evaluation team.

The GFE focused on three essential questions: What are the impacts of the program and how are they evidenced? What are the implications of this learning process? What are suggestions for future efforts? These questions can be broken down into different categories or criteria which will be discussed next.

MSSEP Evaluator's Data Collection

In a GFE, the evaluator specifically avoids learning any information regarding the program's stated goals, or what the program intends to achieve. Program stakeholders (i.e., MSSEP administrators, teachers, paraprofessionals and Parent Corps participants, and other staff) were instructed not to allow the goal-free evaluator to hear or see information regarding the program's stated goals or explicitly refer to the intended goals. The evaluator is told the stated goals after program completion. The evaluator assessed the impacts of the program as they occurred by conducting interviews with students and observing the classrooms. There are several benefits of using a GFE. Scriven (1991) offered some methodological strengths including its ability to assess what the program is actually doing, discover unintended side-effects, provide minimal program disruption, offer a less threatening evaluation for participants and program implementers, and discover if the effects are large enough to notice without the bias of cuing.

This design was non-experimental as there was neither random assignment nor a control group. Rather, the GFE is a "snapshot" of what occurred in the 19 days of MSSEP. Causal claims of program outcomes and/or impacts should be considered with caution (due to the evaluation design, and the short program duration).

The investigative framework was based on classroom observations, structured open student interviews, and standard instruments. The evaluator also interviewed the MSSEP principal, home school interventionist, counseling intern, and program secretary to discuss their roles within the program. Since the GFE was conducted independently yet simultaneously with a GBE, the screening of written material was conducted by evaluators from the GBE team, and both the goal-free evaluator and the GBE team frequently reminded stakeholders not to provide goal-oriented information to the goal-free evaluator.

The evaluator was on-site for 14 days of the 19 days that the students attended (56 hours of the total 76 hours, or 74 percent of the program). Prior to the first day of the program, the goal-free evaluator agreed with the Coordinator of Title I and School Improvement to: (i) learn the intended goals (without prompting), determine student needs, and observe actual program outcomes; (ii) write a goal-free evaluation report; (iii) receive feedback from the program coordinator; and (iv) write a GFE report, including a section after the evaluator becomes aware of the program's stated goals.

The evaluator visited each classroom unannounced and remained in that classroom for the majority of the program day at least once per classroom. Typically, the evaluator entered the classroom with or shortly following the students and observed from the back of the room. All interactions, with exception of the student interviews, were not initiated by the evaluator.

- When students were breaking for lunch the evaluator approached the teacher and asked for permission to address the class and then bring the students to the hall to interview them during lunch. The estimated time per student interview was between 3-7 minutes. A total of 70 student interviews were conducted. The students interviewed were selected by purposeful sampling and conducted during the students' 20 minute to half hour lunch. The students who were interviewed were selected from the same classroom in which the evaluator was observing. Prior to interviewing students, the evaluator introduced himself; announced that the evaluation is of the program and not of teachers or students; explained confidentiality between the students/teachers and the evaluator; and discussed confidentiality in the evaluation report. Some students volunteered to be interviewed, while other students were prompted by the teacher or asked by the evaluator. No student refused or was uncooperative with the evaluator during the interviews. Interviews were administered in the hall to ensure privacy.

- The standardized instrument was the Ratings of Key Indicators (RKI) which included the Classroom Observation Protocol (COP) (Horizon Research, 2004[24]). It was selected because it was general enough to potentially describe the effects of the program without prior knowledge of the intended program goals. With the RKI the evaluator rates the teacher's role in delivering the curriculum by rating the teacher on seven dimensions on a Likert scale from 1 = "Not at All" to 5 = "To a Great Extent." The evaluator rated the teacher and classroom by

[24] Adapted from: Horizon Research, Inc. (2004). "Classroom Observation Protocol." Retrieved April 25, 2005 from http://www.horizon- research.com/pdconvocation/20021001 /abbreviated.pdf

comparing them with other classrooms within the MSSEP. The final item on the RKI is the COP which is an overall assessment of the quality and likely impact of the lesson scored from one to five (1=Ineffective Instruction, 2=Elements of Effective Instruction, 3=Beginning stages of effective instruction, 4=Accomplished Effective Instruction, and 5=Exemplary Instruction). The evaluator used the RKI after spending at least half of a program day in a particular classroom. Additionally, the evaluator examined the classrooms' RKI scores and compared them with other observations, outcomes, and impacts of the program.

- To obtain an overall impression of the program processes, the evaluator also interviewed the counseling intern, the program secretary, the home school interventionist, and the program principal.

Data was recorded in handwritten form and then transferred and synthesized using Microsoft Office applications. Raw data was analyzed by the goal-free evaluator. To ensure evaluation quality, the final GFE was compared to the "Qualitative Evaluation Checklist" by Michael Quinn Patton (2003) and the "Key Evaluation Checklist" by Michael Scriven (1991).

Validity was maintained by observing multiple classrooms at multiple times, by maximizing objectivity through expert consultation, by interviewing key program staff, and by developing a plan for dealing with possible contamination of the goal-free evaluator by learning the program's intended goals. Additionally, the goal-free evaluator had prior training with interview and survey techniques; and experience in observing classrooms, student learning, child and adolescent behavior, and group dynamics and processes.

In obtaining support for the MSSEP staff and students, the evaluator distributed a letter at a meeting of administrators, teachers and teacher aides. The letter contained an introduction, summarized the GFE methodology, discussed confidentiality, and outlined a complaint process for staff. To the evaluator's knowledge, no staff made an informal or formal complaint regarding the GFE or evaluator. Prior to the student interviews, each teacher permitted the evaluator to introduce himself and explain to the students the nature of the evaluation and confidentiality with the interviews and the evaluation report. As a result, the district supported the evaluator and provided full access to student and program data under the agreement that all access to raw data, student histories, and other material from the district was restricted solely to the evaluator.

VIII. PROGRAM DOCUMENTS AND ARCHIVAL RECORDS (Attached)

The next part of this training involves the examination of the program's documents and archives to determine the criteria of merit and develop an evaluation strategy or plan. *EXERCISE (15-20 minutes)*

Begin by read the program documents and archives. While you read, jot down answers to the following questions:

- What are potential criteria of merit or relevant outcomes?
- What are potential performance standards?
- What or who are potential for sources of information?
- What ethical issues are of particular concern in this evaluation?
- What questions need to be answered before finalizing an evaluation plan?
- How can we immediately delegate tasks among the three evaluators on our team?

The Program Team is a collaboration of 18 people between Agencies X, Y, and Z.

An evaluation of the program's impact was completed in 2006.

Below is a list of the team members:

EVALUATOR'S LOG OF POTENTIAL THREATS TO THE GOAL-FREE NATURE

Date & Time of Incident		
Who were the evaluator(s) involved?		
Where did the incident occur?		
What was said or read that is potentially threatening to the goal-free nature of the evaluation?		
Who was the source and what was the context of the statement or writing?		
What was the evaluators' response to the threat?		
Was it reported to one of the investigators?	No ☐	Yes ☐ Date: Investigator:

EVALUATOR'S COMMUNICATION LOG

Date	
Start time	
End time	
Total time (hrs & mins)	
Who communicated with whom?	
How did you communicate?	
Where did the communication occur?	
Describe the nature of the communication.	

EVALUATOR'S TIME LOG

Evaluator's Name:

_____

Date		
Total Time Spent on the Evaluation (hrs & min)		
Describe your daily evaluation-related activities.		

Appendix E

Three Versions of the Evaluation Utility Questionnaire

Evaluation Utility Questionnaire Instructions

In completing this questionnaire, please make your judgments based on the information found in the evaluation report provided to you. You will be presented a set of rating scales on evaluation utility followed by an open-ended question.

Instructions for Completing the Ratings on the Scales:

Do not be concerned if on occasion you feel as though you have seen the same or a similar item on the scales. Make each item a separate and independent judgment. Work at a fairly high speed. Do not ponder individual items. Do not refer back to the evaluation report. It is your first impressions; your immediate "feelings" about the items in relation to the evaluation report that is sought. On the other hand, please do not be careless, because we want your true impressions.

Using the Scales

If you feel that one dimension of the evaluation report is very closely related to either end of the descriptive scale, you should place your mark as follows:

Ex.	Useful	_X_ | __ | __ | __ | __ | __ | __	Useless

OR

Ex.	Useful	__ | __ | __ | __ | __ | __ | _X_	Useless

If you feel that the dimension is quite closely related or usually related to either end of the scale (but not extremely), you should place your mark as follows:

Ex.	Useful	__ | _X_ | __ | __ | __ | __ | __	Useless

OR

Ex.	Useful	__ | __ | __ | __ | __ | _X_ | __	Useless

If you feel that the dimension is slightly related to either end of the scale (but is not neutral), you should place your mark as follows:

Ex.	Useful	__ | __ | _X_ | __ | __ | __ | __	Useless

OR

Ex.	Useful	__ | __ | __ | __ | _X_ | __ | __	Useless

The direction toward which you check depends on which of the two ends of the scale seem most characteristic of the dimension that you are judging.

If you feel that a dimension is neutral on the scale (i.e., both sides of the scale are equally associated with the concept) then you should place your mark in the middle space as follows:

Ex.	Useful	__	__	__	_X_	__	__	__	Useless

Evaluation Utility Questionnaire (VERSION 1)

A. What is your full name?

B. In regard to the usefulness of the evaluation, the evaluation report is…

B-1	Useful	___ | ___ | ___ | ___ | ___ | ___ | ___	Useless
B-2	Conclusive	___ | ___ | ___ | ___ | ___ | ___ | ___	Inconclusive
B-3	Believable	___ | ___ | ___ | ___ | ___ | ___ | ___	Unbelievable
B-4	Trustworthy	___ | ___ | ___ | ___ | ___ | ___ | ___	Untrustworthy
B-5	Clear	___ | ___ | ___ | ___ | ___ | ___ | ___	Unclear
B-6	Consistent	___ | ___ | ___ | ___ | ___ | ___ | ___	Inconsistent
B-7	True	___ | ___ | ___ | ___ | ___ | ___ | ___	False
B-8	Careful	___ | ___ | ___ | ___ | ___ | ___ | ___	Careless
B-9	Logical	___ | ___ | ___ | ___ | ___ | ___ | ___	Illogical
B-10	Valid	___ | ___ | ___ | ___ | ___ | ___ | ___	Invalid
B-11	Meaningful	___ | ___ | ___ | ___ | ___ | ___ | ___	Meaningless
B-12	Worthwhile	___ | ___ | ___ | ___ | ___ | ___ | ___	Worthless
B-13	Complete	___ | ___ | ___ | ___ | ___ | ___ | ___	Incomplete
B-14	Correct	___ | ___ | ___ | ___ | ___ | ___ | ___	Incorrect
B-15	Helpful	___ | ___ | ___ | ___ | ___ | ___ | ___	Unhelpful
B-16	Objective	___ | ___ | ___ | ___ | ___ | ___ | ___	Biased
B-17	Specific	___ | ___ | ___ | ___ | ___ | ___ | ___	Vague
B-18	Enlightening	___ | ___ | ___ | ___ | ___ | ___ | ___	Unenlightening
B-19	Fair	___ | ___ | ___ | ___ | ___ | ___ | ___	Unfair
B-20	Relevant	___ | ___ | ___ | ___ | ___ | ___ | ___	Irrelevant

B-21	Reasonable	___ | ___ | ___ | ___ | ___ | ___ | ___	Unreasonable
B-22	Informative	___ | ___ | ___ | ___ | ___ | ___ | ___	Uninformative
B-23	Honest	___ | ___ | ___ | ___ | ___ | ___ | ___	Dishonest
B-24	Effective	___ | ___ | ___ | ___ | ___ | ___ | ___	Ineffective
B-25	Balanced	___ | ___ | ___ | ___ | ___ | ___ | ___	Unbalanced

B-26. In the space below, please provide an explanation as to why the evaluation report was or was not useful:

Thank you for completing the first of two *Evaluation Utility Questionnaires*.

REMINDERS:

- Please review your questionnaire to ensure that you've answered every item and that all of your responses are clearly marked.

- Once you've returned the questionnaire, you should expect to receive the second evaluation report within a week.

- Please return this questionnaire to THE INVESTIGATOR by January X, 2010.

Address:

Email:
Phone: ( )

Again, your participation in this important study of program evaluation is greatly appreciated.

Evaluation Utility Questionnaire (VERSION 2)

A. What is your full name?

B. In regard to the usefulness of the evaluation, the evaluation report is…

B-1	Enlightening	___ I ___ I ___ I ___ I ___ I ___ I ___	Unenlightening
B-2	Clear	___ I ___ I ___ I ___ I ___ I ___ I ___	Unclear
B-3	Trustworthy	___ I ___ I ___ I ___ I ___ I ___ I ___	Untrustworthy
B-4	Helpful	___ I ___ I ___ I ___ I ___ I ___ I ___	Unhelpful
B-5	Consistent	___ I ___ I ___ I ___ I ___ I ___ I ___	Inconsistent
B-6	Complete	___ I ___ I ___ I ___ I ___ I ___ I ___	Incomplete
B-7	Effective	___ I ___ I ___ I ___ I ___ I ___ I ___	Ineffective
B-8	Logical	___ I ___ I ___ I ___ I ___ I ___ I ___	Illogical
B-9	Careful	___ I ___ I ___ I ___ I ___ I ___ I ___	Careless
B-10	Specific	___ I ___ I ___ I ___ I ___ I ___ I ___	Vague
B-11	Informative	___ I ___ I ___ I ___ I ___ I ___ I ___	Uninformative
B-12	Objective	___ I ___ I ___ I ___ I ___ I ___ I ___	Biased
B-13	Useful	___ I ___ I ___ I ___ I ___ I ___ I ___	Useless
B-14	Fair	___ I ___ I ___ I ___ I ___ I ___ I ___	Unfair
B-15	Reasonable	___ I ___ I ___ I ___ I ___ I ___ I ___	Unreasonable
B-16	Correct	___ I ___ I ___ I ___ I ___ I ___ I ___	Incorrect
B-17	True	___ I ___ I ___ I ___ I ___ I ___ I ___	False
B-18	Conclusive	___ I ___ I ___ I ___ I ___ I ___ I ___	Inconclusive
B-19	Valid	___ I ___ I ___ I ___ I ___ I ___ I ___	Invalid
B-20	Balanced	___ I ___ I ___ I ___ I ___ I ___ I ___	Unbalanced
B-21	Meaningful	___ I ___ I ___ I ___ I ___ I ___ I ___	Meaningless

B-22	Believable	___ I ___ I ___ I ___ I ___ I ___ I ___	Unbelievable
B-23	Relevant	___ I ___ I ___ I ___ I ___ I ___ I ___	Irrelevant
B-24	Honest	___ I ___ I ___ I ___ I ___ I ___ I ___	Dishonest
B-25	Worthwhile	___ I ___ I ___ I ___ I ___ I ___ I ___	Worthless

B-26. In the space below, please provide an explanation as to why the evaluation report was or was not useful:

Thank you for completing the first of two *Evaluation Utility Questionnaires*.

REMINDERS:

- Please review your questionnaire to ensure that you've answered every item and that all of your responses are clearly marked.

- Once you've returned the questionnaire, you should expect to receive the second evaluation report within a week.

- Please return this questionnaire to THE INVESTIGATOR by January X, 2010.

Address:

Email:
Phone: ( )

Again, your participation in this important study of program evaluation is greatly appreciated.

Evaluation Utility Questionnaire (VERSION 3)

A. What is your full name?	

B. In regard to the usefulness of the evaluation, the evaluation report is…			
B-1	Objective	___ | ___ | ___ | ___ | ___ | ___ | ___	Biased
B-2	Honest	___ | ___ | ___ | ___ | ___ | ___ | ___	Dishonest
B-3	Complete	___ | ___ | ___ | ___ | ___ | ___ | ___	Incomplete
B-4	Fair	___ | ___ | ___ | ___ | ___ | ___ | ___	Unfair
B-5	Meaningful	___ | ___ | ___ | ___ | ___ | ___ | ___	Meaningless
B-6	True	___ | ___ | ___ | ___ | ___ | ___ | ___	False
B-7	Useful	___ | ___ | ___ | ___ | ___ | ___ | ___	Useless
B-8	Consistent	___ | ___ | ___ | ___ | ___ | ___ | ___	Inconsistent
B-9	Enlightening	___ | ___ | ___ | ___ | ___ | ___ | ___	Unenlightening
B-10	Trustworthy	___ | ___ | ___ | ___ | ___ | ___ | ___	Untrustworthy
B-11	Correct	___ | ___ | ___ | ___ | ___ | ___ | ___	Incorrect
B-12	Conclusive	___ | ___ | ___ | ___ | ___ | ___ | ___	Inconclusive
B-13	Careful	___ | ___ | ___ | ___ | ___ | ___ | ___	Careless
B-14	Relevant	___ | ___ | ___ | ___ | ___ | ___ | ___	Irrelevant
B-15	Clear	___ | ___ | ___ | ___ | ___ | ___ | ___	Unclear
B-16	Balanced	___ | ___ | ___ | ___ | ___ | ___ | ___	Unbalanced
B-17	Helpful	___ | ___ | ___ | ___ | ___ | ___ | ___	Unhelpful
B-18	Specific	___ | ___ | ___ | ___ | ___ | ___ | ___	Vague
B-19	Reasonable	___ | ___ | ___ | ___ | ___ | ___ | ___	Unreasonable
B-20	Effective	___ | ___ | ___ | ___ | ___ | ___ | ___	Ineffective

B-21	Believable	___	|	___	|	___	|	___	|	___	|	___	|	___	Unbelievable
B-22	Worthwhile	___	|	___	|	___	|	___	|	___	|	___	|	___	Worthless
B-23	Logical	___	|	___	|	___	|	___	|	___	|	___	|	___	Illogical
B-24	Valid	___	|	___	|	___	|	___	|	___	|	___	|	___	Invalid
B-25	Informative	___	|	___	|	___	|	___	|	___	|	___	|	___	Uninformative

B-26. In the space below, please provide an explanation as to why the evaluation report was or was not useful:

Thank you for completing the first of two *Evaluation Utility Questionnaires*.

REMINDERS:

- Please review your questionnaire to ensure that you've answered every item and that all of your responses are clearly marked.

- Once you've returned the questionnaire, you should expect to receive the second evaluation report within a week.

- Please return this questionnaire to THE INVESTIGATOR by January X, 2010.

Address:

Email:
Phone: ( )

Again, your participation in this important study of program evaluation is greatly appreciated.

Appendix F

Responses to the Open-Ended Questionnaire Question

The final question on the *Evaluation Utility Questionnaire* asked the evaluation user to "please provide an explanation as to why the evaluation report was or was not useful." In the following transcriptions of the evaluation users' open-ended responses, the grammar and punctuation reflect that of the original handwritten response as best as possible; for example if a respondent wrote a fragment, abbreviated a word, or used an ellipsis, it is also so in the transcriptions below.

Open-ended Responses to the GAE Report

A-X01: [Program] funding is based on successful outcomes and the evaluation offers third party confirmation projected outcomes are indeed achieved. - The only unclear portion was whether retaining housing/employment under 6 mos. Applied to only those who exited prior to 6 mos. or also those who had not been in the program beyond 6 mos.

A-X02: Questions arose around the criteria of merit and the actual numbers disclosed, i.e. population sample. - The grade assigned to housing seems generous. - Also, questions regarding the selection of participants, based upon the judgments of case managers involved. How is this done and does that impact the results?

A-X03: [Blank]

A-X04: The report will most likely be helpful if an evaluator explained the report more completely. The [Program] partners meet monthly (typically the 3rd Monday of each month).

A-Y04: [Blank]

A-Y05: [Blank]

Open-ended responses to the GFE report:

A-X01: The evaluation report was very useful. I acquired more insight into the effectiveness of the program under the "goal free" evaluation method. I found value in identifying consumer needs as outcome criteria for measuring the success of the program. - On the other hand, the ___ program primary goal is employment as a means to housing stability and it appears from the consumer perspective the reverse is true. Clearly, the basic need for shelter trumps other goal areas and we need to explore program adaptations given the negative economic climate and consumer perspective.

A-X02: [Blank]

A-X03: Dimensions of merit helpful. Capacity building helpful. - 4, 2, 2 Evaluator Rec's seemed limited in value and scope and conclusions did not seem well supported. Overall, I did find value in some specifics of report-client recommendations, staff client interactions, etc.

A-X04: Receiving the feedback comments from the participants involved was helpful in that the information was given to a neutral person; therefore more likely to be honest info. (they had nothing to gain or lose by giving their input). I appreciated knowing their thoughts of what was helpful/not helpful, suggestions for improvements... I also really like the suggestion of family mentoring family since I STRONGLY believe that relationships produce success far more than programs.

A-Y04: [Blank]

A-Y05: [Blank]

Appendix G

Letter from the Human Subjects
Institutional Review Board

WESTERN MICHIGAN UNIVERSITY
Human Subjects Institutional Review Board

Date: October 14, 2008

To: Chris Coryn, Principal Investigator
Brandon Youker, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: Approval not needed

This letter will serve as confirmation that your project "An Analog Study Comparing Goal-Free Evaluation and Goal Achievement Evaluation Utility" has been reviewed by the Human Subjects Institutional Review Board (HSIRB). Based on that review, the HSIRB has determined that approval is not required for you to conduct this project because you are studying program evaluation approaches and not collecting information about individuals. Thank you for your concerns about protecting the rights and welfare of human subjects.

A copy of your protocol and a copy of this letter will be maintained in the HSIRB files.

Walwood Hall, Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276

BIBLIOGRAPHY

Academic Analytics. (2008, May 16). *FSP index top performing individual programs*. Retrieved May 19, 2005, from http://www.academicanalytics.com/TopSchools /TopPrograms.aspx#4

Alkin, M. C. (1969). Evaluation theory development. *Evaluation Comment, 2,* 2-7.

Alkin, M. C. (1972). Wider context goals and goal-based evaluators. *Evaluation Comment: The Journal of Educational Evaluation, 3*(4), 10-11.

Alkin, M. C. (1991). Evaluation theory development: II. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century.* Ninetieth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press.

Alkin, M. C. (2004a). Comparing evaluation points of view. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influence* (pp. 3-11). Thousand Oaks, CA: Sage.

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

Altschuld, J. W., & Witkin, B. R. (2000). From needs assessments to action: *Transforming needs into solution strategies.* Thousand Oaks, CA: Sage.

Amo, C., & Cousins, J. B. (2007). Going through the process: An examination of the operationalization of process use in empirical research on evaluation. *New Directions for Evaluation, 116*, 5-26. San Francisco: Jossey-Bass.

Argyris, C., & Schon, D. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison-Wesley.

Belli, Robert. (2009). *Calendar and time diary: Methods in life course research.* Los Angeles: Sage.

Bernstein, R. J. (1983). *Beyond objectivism and relativism: Science, hermeneutics, and praxis*. Philadelphia: University of Pennsylvania Press.

Bickman, L. (2005). Evaluation research. In S. Mathison (Ed.), *Encyclopedia of evaluation* (p. 141). Thousand Oaks, CA: Sage.

Bickman, L., & Rog, D. (1998). Introduction: Why a handbook of applied social research methods? In L. Bickman & D. Rog (Eds.), *Handbook of applied social science research methods* (pp. xiii-xiv). Thousand Oaks, CA: Sage.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I. Cognitive domain.* New York: David McKay.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. W. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research, 2*(2), 75-87.

Bourque, L. B., & Fielder, E. P. (2003). *How to conduct self-administered and mail surveys* (2nd ed.). Thousand Oaks, CA: Sage.

Brinkerhoff, R. O. (2003). *The success case method*. San Francisco: Berrett Koehler.

Bullock, A., & Trombley, B. (Eds.). (1999). *The Fontana dictionary of modern thought*. London: Harper-Collins.

Campbell, S., & Stanley, J. C. (1963/1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Chen, H. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.

Chen, H., & Rossi, P. (1983). Evaluating with sense: The theory driven approach. *Evaluation Review, 7,* 283-302.

Chen, Q., Fisher, D. T., Clancy K. A., Gauguet, J. M., Wang, W.C., Unger, E., et al. (2006, June). Fever-range thermal stress promotes lymphocyte trafficking across high endothelial venules via an interleukin 6 trans-signaling mechanism. *Nature Immunology*, *7,* 1299-1308.

Christie, C. A., & Alkin, M. C. (2005). Objectives-based evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 281-284). Thousand Oaks, CA: Sage.

Committee on Institutional Cooperation. (2008, October). CIC summer study in Mexico program evaluation report. Retrieved February 14, 2009, from http://www.cic.net/Libraries/ProgramEvals/Guanajuato2008.sflb.ashx

Consumers Union of U.S., Inc. (2000). *Our history.* Retrieved October 10, 2006, from http://www.zillions.org/z_history.html

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Coryn, C. L. S. (2005, October). Book review: "Revisiting realistic evaluation." *Journal of MultiDisciplinary Evaluation, 3.* Retrieved September 19, 2006, from

http://evaluation.wmich.edu/jmde/content/JMDE003content/8_Revisiting_Realistic_Evaluation.htm

Coryn, C. L. S. (2007). *Evaluation of researchers and their research: Toward making the implicit explicit.* Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Coryn, C. L. S., Noakes, L. A., Westine, C. D., & Schröter, D. C., (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation*, *32*(2), 199-226.

Cottingham, J., Stoothoff, R., Murdoch, D., & Kenny, A. (Eds.). (1991). *Descartes: The philosophical writings of Descartes* (J. Cottingham, R. Stoothoff, D. Murdoch, & A. Kenny, Trans.). Cambridge, UK: Cambridge University Press. (Original work published 1641)

Cousins, J. B. (2004). Commentary: Minimizing evaluation misuse as principled practice. *American Journal of Evaluation*, *25*(3), 391-397.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record, 64,* 672-683.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: John Wiley & Sons.

Cronholm, S., & Goldkuhl, G. (2003). *Six generic types of information systems evaluation.* Paper presented at the meeting of the 10th European Conference on Information Technology Evaluation (ECITE-2003), Madrid, Spain.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

DePanfilis, D., & Salus, M. K. (1992). *A coordinated response to child abuse and neglect: A basic manual*. Washington, DC: U.S. Department of Health and Human Services National Center on Child Abuse and Neglect.

Dennett, D. C. (1987). *The intentional stance*. Boston: The MIT Press.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life.* New York: Simon & Schuster.

Editors of Chambers (Eds.). (2004). *Chambers concise dictionary*. Edinburgh, United Kingdom: Chambers Harrap Publishers.

Eisenhower, D. D. (1960). *Public papers of the presidents: Military-industrial complex speech, Dwight D. Eisenhower, 1961*, (pp. 1035-1040). Retrieved May 14, 2008 from http://coursesa.matrix.msu.edu/~hst306/documents/indust.html

Eisner, E. W. (1979a). *The educational imagination: On the design and evaluation of school programs.* New York: Macmillan.

Eisner, E. W. (1979b). The use of qualitative forms of evaluation for improving educational practice. *Educational Evaluation and Policy Analysis, 1,* 11-19.

Eisner, E. W. (1985). *The art of educational evaluation: A personal view*. Philadelphia: The Falmer Press.

Eisner, E. W. (1990). The meaning of alternative paradigms for practice. In E. Guba (Ed.), *The paradigm dialog* (pp. 88-102). Newbury Park, CA: Sage.

Eisner, E. W. (1991). Taking a second look: Educational connoisseurship revisited. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century. Ninetieth Yearbook of the National Society for the Study of Education, Part II.* Chicago: University of Chicago Press.

Evers, J. W. (1980). *A field study of goal-based and goal-free evaluation techniques*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Fetterman, D. M., & Wandersman, A. (Eds.). (2005). *Empowerment evaluation principles in practice.* New York: The Guilford Press.

Firestone, W. A. (1990). Accommodation: Toward a paradigm-praxis dialectic. In E. Guba (Ed.), *The paradigm dialog* (pp. 105-124). Newbury Park, CA: Sage.

Fischer, F. (1985). Critical evaluation of public policy: A methodological case study. In J. Forester (Ed.)*, Critical theory and public life* (pp. 231-257). Cambridge, MA: MIT Press.

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

Fournier, D. M. (Ed.). (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation, 68*, 15-32. San Francisco: Jossey-Bass.

Fournier, D. M. (2005). Logic of evaluation: Working logic. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 238-242). Thousand Oaks, CA: Sage.

Friedman, V. J., Rothman, J., & Withers, B. (2006). The power of why: Engaging the goal paradox in program evaluation. *American Journal of Evaluation, 27*(2), 201-218.

Garrison, F. (1960). *An introduction to the history of medicine: With medical chronology, suggestions for study and biographical data* (4th ed.). Philadelphia: W. B. Saunders.

Goodlad, J. (1979). *What schools are for.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Gordon, E. W. (1974). Toward a qualitative approach to assessment. In R. Tyler & R. Wolf (Eds.), *Crucial issues in testing* (pp. 58-62). Berkeley, CA: McCutchan.

Gould, J. A. (Ed.). (1994). *Classical philosophical questions* (8th ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Greene, J. C. (2005). Mixed methods. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 255-256). Thousand Oaks, CA: Sage.

Grinnell, R. M., & Unrau, Y. A. (Eds.). (2008). *Social work research and evaluation: Foundations of evidence-based practice* (8th ed.). New York: Oxford University Press.

Grinnell, R. M., Unrau, Y. A., & Gabor, P. A. (2008). Program evaluation. In R. Grinnell & Y. Unrau (Eds.), *Social work research and evaluation: Foundations of evidence-based practice* (8th ed.) (pp. 527-539). New York: Oxford University Press.

Grove, P. B. (Ed.). (1986). *Webster's third new international dictionary of the English language* (unabridged). Springfield, MA: Merriam-Webster.

Guba, E. G. (1969). The failure of educational evaluation. *Educational Technology, 9,* 29-38.

Guba, E. G. (1990a). The alternative paradigm dialog. In E. Guba (Ed.), *The paradigm dialog* (pp. 17-27). Newbury Park, CA: Sage.

Guba, E. G. (Ed.). (1990b). *The paradigm dialog.* Newbury Park, CA: Sage.

Guba, E. G. (2005). Paradigm. In S. Mathison (Ed.), *Encyclopedia of evaluation* (p. 289). Thousand Oaks, CA: Sage.

Hellström, T., & Jacob, M. (2003). Knowledge without goals? Evaluation of knowledge management programmes. *Evaluation, 9*(1), 55-72.

Henry, G. T. (1998). Practical sampling. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 101-126). Thousand Oaks, CA: Sage.

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

Hezel, R. T. (1995). Considerations for the evaluation of the National Science Foundation programs. In J. Frechtling (Ed.), *Footprints: Strategies for non-traditional program evaluation* (pp. 47-52). Sponsored by the National Science Foundation Directorate for Education & Human Resources. Retrieved November 11, 2007, from http://www.nsf.gov/pubs/1995/nsf9541/nsf9541.pdf

Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *Psychology of attitudes* (pp. 23-88). Thomson/Wadsworth.

Hollis, M. (1995). *The philosophy of social science: An introduction* (Rev. ed.). Cambridge, England: Cambridge University Press.

Horkheimer, M. (1972). *Critical theory*. New York: Herder and Herder.

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

House, E. R. (1983). Assumptions underlying evaluation models. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models*. Boston: Kluwer-Nijhoff.

Institute of Medicine. (2001). *Improving the quality of long-term care.* Washington, DC: National Academy Press.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.

Kettner, P. K., Moroney, R. K., & Martin, L. L. (in press). *Designing and managing programs: An effectiveness-based approach* (3rd ed.). Thousand Oaks, CA: Sage.

Kidder, L., & Judd, C. M. (1986). *Research methods in social relations* (5th ed.). New York: Hold, Rinehart & Wilson.

Krosnick, J. A., & Alwin, D. F. (1987). *Satisficing: A strategy for dealing with the demands of survey questions.* Paper presented at the American Association for Public Opinion Research Annual Meeting, Hershey, PA.

Kuhn, T. S. (1962, 1970, 1996). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kushner, S. (2005). Independence. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 198-199). Thousand Oaks, CA: Sage.

Lee, J., & Ahn, K. (2004). *Categorization and effectiveness of Korean housing policy: An analysis of effectiveness and price impact.* Paper presented at the Asian Real Estate Society 9th Annual International Conference, Seoul, South Korea. Retrieved September 9, 2008 from www.asres.org/2004Conference/papers/Lee%20&%20Ahn.doc

Madaus, G., & Stufflebeam, D. (1989). *Educational evaluation: Classic works of Ralph W. Tyler*. Boston: Kluwer.

Madaus, G. F., Stufflebeam, D. L., & Scriven, M. (1983). Program evaluation: A historical overview. In G. Madaus, D. Stufflebeam, & M. Scriven (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 3-22). Boston: Kluwer-Nijhoff.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs.* San Francisco: Jossey-Bass.

Martí-Ibáñez, F. (1961). *A prelude to medical history*. New York: MD Publications.

Masterman, M. (1970). The nature of a paradigm. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 59-89). Cambridge: Cambridge, UK: University Press.

Mathison, S. (2005a). Bias. In S. Mathison (Ed.), *Encyclopedia of evaluation* (p. 33). Thousand Oaks, CA: Sage.

Mathison, S. (2005b). Synthesis. In S. Mathison (Ed.), *Encyclopedia of evaluation* (p. 405). Thousand Oaks, CA: Sage.

Mauk, K. L., & Schmidt, N. K. (Eds.). (2004). *Spiritual care in nursing practice*. Philadelphia: Lippincott Williams and Wilkins.

Mautner, T. E. (Ed.) (2005). *The Penguin dictionary of philosophy* (2nd ed.). London: Penguin Books.

Metfessel, N. S., & Michael, W. B. (1967). A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measurement, 27,* 931-943.

Mohan, R., Tikoo, M., Capela, S., & Bernstein, D. J. (2006). Increasing evaluation use among policymakers through performance measurement. *New Directions for Evaluation, 112*, 89-97. San Francisco: Jossey-Bass.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcomes studies. *Clinical Psychology Review, 11,* 247-266.

Mowbray, C. T, Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315-340.

Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (2004). Meeting the need for practical evaluation approaches: An introduction. In J. S. Wholey, H. P. Hatry, & K.

E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed.) (pp. xxxiii-xliv). San Francisco: Jossey-Bass.

Nuland, S. (1988). *Doctors: The biography of medicine*. New York: Knopf.

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Parlett, M., & Hamilton, D. (1976). Evaluation as illumination: A new approach to the study of innovatory programs. In G. V. Glass (Vol. Ed.), *Evaluation studies review annual: Vol. 1.* (pp. 140-157). Beverly Hills, CA: Sage.

Patton, M. Q. (1988). The evaluator's responsibility for utilization. *Evaluation Practice, 9*(2), 5-24.

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

Patton, M. Q. (2002a). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2002b). *Utilization-focused evaluation (U-FE) checklist.* Retrieved August 10, 2008, from Western Michigan University, Evaluation Center website: http://www.wmich.edu/evalctr/archive_checklists/ufe.pdf

Patton, M. Q. (2007). Process use as a usefulism. *New Directions for Evaluation, 116*, 99-112. San Francisco: Jossey-Bass.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Pickett, J. P. (Ed.). (2000). *The American heritage dictionary of the English language* (4th ed.). Boston: Houghton Mifflin.

Popham, W. J., Eisner, E. W., Sullivan, H. J., & Tyler, L. L. (1969). *Instructional objectives.* American Educational Research Association Monograph Series on Curriculum Evaluation No. 3. Chicago: Rand McNally.

Posavac, E. J., & Carey, R. G. (1997). *Program evaluation: Methods and case studies* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Preskill, H., Zuckerman, B., & Matthews, B. (2003). An exploratory study of process use: Findings and implications for future research. *American Journal of Evaluation, 24*(4), 423-442.

Provus, M. M. (1971). *Discrepancy evaluation*. Berkeley, CA: McCutchan.

Provus, M. M. (1973). Evaluation of ongoing programs in the public school system. In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice.* Belmont, CA: Wadsworth.

Random House (Eds.). (2001). *The Random House Webster's unabridged dictionary* (2nd ed.). New York: Random House.

Reichardt, C. S., & Mark, M. M. (1998). Quasi-experimentation. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 193-228). Thousand Oaks, CA: Sage.

Rodríguez-Campos, L. (2005). *Collaborative evaluations: A step-by-step model for the evaluator*. Tamarac, FL: Llumina Press.

Rogers, P. J. (2000). Program theory: Not whether programs work but how they work. In D. Stufflebeam, G. Madaus, & T. Kellaghan (Eds.), *Evaluation models*. Boston: Kluwer Academic.

Rogers, P. J. (2005). Accountability. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 2-4). Thousand Oaks, CA: Sage.

Rossi, P. H. (2004). My views of evaluation and their origins. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 122-131). Thousand Oaks, CA: Sage.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

Royse, D., Thyer, B. A., Padgett, D. K., & Logan, T. K. (2006). *Program evaluation: An introduction* (4th ed.). Belmont, CA: Thomson Brooks/Cole.

Salasin, S. (1974) Exploring goal-free evaluation: An interview with Michael Scriven. *Evaluation, 2*(1), 9-16.

Sanders, J. (2006, November). Ten things evaluation needs: An evaluation needs assessment. *Journal of MultiDisciplinary Evaluation*, *3*(6), 58-59. Retrieved March 10, 2008, from http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view /40/49

Schalock, R. L., & Thornton, C. V. D. (1988). *Program evaluation: A field guide for administrators.* New York: Plenum.

Schuftan, C. (1988). Multidisciplinary, paradigms and ideology in development work. *Scandinavian Journal of Development Alternatives, 8*, (2). Retrieved January 6, 2007, from http://www.humaninfo.org/aviva/ch10.htm

Schwandt, T. A. (1990). Paths to inquiry in the social disciplines: Scientific, constructivist, and critical theory methodologies. In E. G. Guba (Ed.), *The paradigm dialog* (pp. 258-276). Newbury Park, CA: Sage.

Schwandt, T. A. (2001). *Dictionary of qualitative inquiry* (2nd rev. ed.). Thousand Oaks, CA: Sage.

Schwandt, T. A. (2002). *Evaluation practice reconsidered.* New York: Peter Lang.

Schwandt, T. A. (2005). Positivism. In S. Mathison (Ed.), *Encyclopedia of evaluation* (p. 324). Thousand Oaks, CA: Sage.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Curriculum evaluation.* American Educational Research Association Monograph Series on Evaluation No. 1. Chicago: Rand McNally.

Scriven, M. (1972). Pros and cons about goal-free evaluation. *The Journal of Educational Evaluation, 3*(4), 1-7.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process* (pp. 319-328). Berkeley, CA: McCutchan.

Scriven, M. (1974a). Evaluation perspectives and procedures. In W. J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 1-93). Berkeley, CA: McCutchan.

Scriven, M. (1974b). Prose and cons about goal-free evaluation. In W. J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 34-67). Berkeley, CA: McCutchan.

Scriven, M. (1976). Evaluation bias and its control. In G. V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1). Newbury Park, CA: Sage.

Scriven, M. (1983). Evaluation ideologies. In G. Madaus, D. Stufflebeam, & M. Scriven (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 229-260). Boston: Kluwer-Nijhoff.

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

Scriven, M. (1993). Hard-won lessons in program evaluation. *New Directions for Evaluation, 58*. San Francisco: Jossey-Bass.

Scriven, M. (2004, October). The fiefdom problem. *Journal of MultiDisciplinary Evaluation, 1*, 1-8. Retrieved August 22, 2006, from http://www.evaluation .wmich.edu/jmde/content/JMDE1content/2_Editorial.htm

Scriven, M. (2005a). Logic of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 235-238). Thousand Oaks, CA: Sage.

Scriven, M. (2005b). The problem of free will in program evaluation. *Journal of MultiDisciplinary Evaluation*, *2*(2), 102-104. Retrieved May 8, 2006, from http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/124/139

Scriven, M. (2007, February) *Key evaluation checklist.* Retrieved May 14, 2007, from Western Michigan University, Evaluation Center website http://www.wmich.edu /evalctr/checklists/kec_feb07.pdf

Scriven, M., & Davidson, E. J. (2000, November). *The synthesis problem: Issues and methods in the combination of evaluation results into overall evaluative conclusions.* Symposium conducted at the meeting of the American Evaluation Association, Honolulu, HI. Retrieved January 10, 2008, from http://davidsonconsulting.co.nz/index_files/pres/synthHNL.pdf

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15-35). Thousand Oaks, CA: Sage.

Shadish, W. R. (1995). Philosophy of science and the quantitative-qualitative debates: Thirteen common errors. *Evaluation and Program Planning, 18*(1), 63-75.

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

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

Siegel, R. (Interviewer) & Dunne, B. (Respondent). (2007, February 12). ESP research lab closes after 28 years. On *All things considered* (Online Recording http://www.npr.org/templates/story/story.php?storyId=7371765). Washington, DC: National Public Radio.

Simpson, J., & Weiner, E. (Eds.). (1989). *The Oxford English dictionary* (2nd ed., Vols. 5 and 6). Oxford, United Kingdom: Oxford University Press.

Sinnett, M. W. (1987). Method versus methodology: A note on the ultimate resource. *The Review of Austrian Economics, 1*(1): 207-223.

Sirotnik, K. A., & Oakes, J. (1990, Spring). Evaluation as critical inquiry: School improvement as a case in point. *New Directions for Program Evaluation,* 45, 37-59. San Francisco: Jossey-Bass.

Skrtic, T. M. (1990). Social accommodation: Toward a dialogical discourse in educational inquiry. In E. Guba (Ed.), *The paradigm dialog* (pp. 125-135). Newbury Park, CA: Sage.

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *American Journal of Evaluation, 14*(2), 237-242.

Smith, N. L. (1994). Evaluation models and approaches. In T. Husen & T. N. Postlethwaite (Eds.), *The International encyclopedia of education* (2nd ed.). Oxford: Pergamon Press.

Stake, R. E. (1967). The countenance of educational evaluation. *Teacher College Record, 68,* 523-540.

Stake, R. E. (1974). *Nine approaches to educational evaluation.* Unpublished chart. Urbana: University of Illinois, Center for Instructional Research and Curriculum Evaluation.

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Stake, R. E. (2004). *Standards-based and responsive evaluation.* Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (1968). *Evaluation as enlightenment for decision making.* Columbus: Ohio State University Evaluation Center.

Stufflebeam, D. L. (1971). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education, 5,* 19-25.

Stufflebeam, D. L. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation*. Boston: Kluwer Nijhof.

Stufflebeam, D. L. (2000). The CIPP model for evaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kelleghan (Eds.)*, Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 274-317). Boston: Kluwer.

Stufflebeam, D. L. (2001). Evaluation models*. New Directions for Evaluation,* 89, 8-98. San Francisco: Jossey-Bass.

Stufflebeam, D. L. (2002). *Utilization-focused evaluation (U-FE) checklist.* Retrieved August 10, 2008, from Western Michigan University, Evaluation Center website http://www.wmich.edu/evalctr/checklists/ufe.pdf

Stufflebeam, D. L. (2005). Personnel evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 308-313). Thousand Oaks, CA: Sage.

Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2000). *Evaluation models*, (Rev. ed.). Boston: Kluwer.

Stufflebeam, D. L., & Shinkfield, A. J. (Eds.). (1985). *Systematic evaluation*. Boston: Kluwer-Nijhoff.

Stufflebeam, D. L., & Webster, W. J. (1983). An analysis of alternative approaches to evaluation. In G. Madaus, M. Scriven, & D. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 23-43). Boston: Kluwer-Nijhoff.

Suchman, E. (1967). *Evaluative research: Principles and practice in public service and social action programs.* New York: Russell Sage.

Suchman, E. (1969). Evaluating educational programs. *Urban Review, 3*(4), 15-17.

Susman, G. I., & Evered, R. D. (1978, December). An assessment of the scientific merits of action research. *Administrative Science Quarterly, 23,* 582-603.

Teresi, D. (2002). *Lost discoveries: The ancient roots of modern science—From the Babylonians to the Maya*. Riverside, NJ: Simon & Schuster.

Thiagarajan, S. (1975). Goal-free evaluation of media. *Educational Technology, 15*(5), 38-40.

Thoreau, H. D. (1849). Resistance to civil government. In E. Peabody (Ed.), *Æsthetic papers*. Boston: Elizabeth Palmer Peabody.

Tourmen, C. (2009). Evaluators' decision making: The relationship between theory, practice, and experience. *American Journal of Evaluation, 30*(1), 7-30.

Travers, R. (1977, October). Presentation in a seminar at the Western Michigan University Evaluation Center, Kalamazoo, MI.

Tucker, J. G. (2005). Goal. In S. Mathison (Ed.), *Encyclopedia of evaluation* (p. 171). Thousand Oaks, CA: Sage.

Tyler, R. W. (1942). General statement on evaluation. *Journal of Educational Research*, *35*, 492-501.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.

Tyler, R. W. (1974). The use of tests in measuring the effectiveness of educational programs, methods, and instructional materials. In R. W. Tyler & R. M. Wolf (Eds.), *Crucial issues in testing* (pp. 143-155). Berkeley, CA: McCutchan.

Vedung, E. (1997). *Public policy and program evaluation.* New Brunswick, NJ: Transaction.

Venturi, L. (1964). *History of art criticism* (Rev. ed.). (C. Marriott, Trans.). New York: E. P. Dutton & Co. (Original work published 1936)

Wall, B. E. (1975). Anatomy of a precursor: The historiography of Aristarchus of Samos. *Studies in History and Philosophy of Science, 6*(3), 201-228.

Walter, E. (Ed.). (2005). *Cambridge advanced learner's dictionary* (2nd ed.). Cambridge, United Kingdom: Cambridge University Press.

Webster's dictionary and thesaurus (Encyclopedic ed.). (2000). Printed in Canada: Trident Press International.

Weiss, C. H. (1967, April). Utilization of evaluation: Toward comparative study. In House of Representatives committee on government operations, *The use of social research in federal domestic programs¸ Part III* (pp. 426-432). Washington, DC: Government Printing Office.

Weiss, C. H. (1972). *Evaluation research: Methods for assessing program effectiveness.* Englewood Cliffs, NJ: Prentice-Hall.

Weiss, C. H. (1988). Evaluation for decisions: Is anybody there? Does anybody care? *Evaluation Practice, 9*(1), 5-19.

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. In D. Rog and D. Fournier (Eds.), *Progress and future directions in evaluation: Perspectives on theory, practice, and methods. New Directions for Evaluation*, 76. San Francisco: Jossey-Bass.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation*, *19*(1), 21-33.

Weiss, H. B., & Jacobs, F. H. (1988). *Evaluating family programs.* Hawthorne, NY: Aldine de Gruyter.

Whitehead, A. N. (1942). *Adventure of ideas*. London: Pelican Books.

Wholey, J. S. (1983). *Evaluation and effective public management*. Boston: Little, Brown.

Winter, R. (1989). *Learning from experience: Principles and practice in action-research.* Philadelphia: The Falmer Press.

Woolfolk, A. (2001). *Educational psychology* (8th ed.). Boston: Allyn and Bacon.

Worthen, B. R., & Sanders, J. R. (1973). *Educational evaluation: Theory and practice.* Worthington, OH: Charles A. Jones.

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

Yamaguchi, K. (1991). Event history analysis. *Applied social research methods* (Vol. 28). Newbury Park, CA: Sage.

Yarbrough, D. B., Shulha, L. M., & Caruthers, F. (2004). Background and history of the Joint Committee's program evaluation standards. *New Directions for Evaluation, 104*, 15-30. San Francisco: Jossey-Boss.

Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed.). Thousand Oaks, CA: Sage.

Yin, R. K. (1998). The abridged version of case study research: Design and method. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 229-259). Thousand Oaks, CA: Sage.

Youker, B. W. (2005a, October). Ethnography and evaluation: Their relationship and three anthropological models of evaluation. *Journal of MultiDisciplinary Evaluation, 3*, 113-132. Retrieved May 14, 2008, from http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/102/117

Youker, B. W. (2005b, November). *Goal-free evaluation*. Paper presented at the meeting the Evaluation Center's Evaluation Café, Kalamazoo, MI. Retrieved September 20, 2007, from http://www.wmich.edu/evalctr/evalcafe/goal-free.pdf

Youker, B. W. (2006, November). *What are values and standards in evaluation?* Paper presented at the meeting of the American Evaluation Association, Portland, Oregon.

Youker, B. W. (2010). The logic of evaluation and not-for-profit arts organizations: The perspective of an evaluation consultant. *International Journal of Arts Management*, *12*(3), 4-12.

Young, R. E. (1990). *A critical theory of education: Habermas and our children's future.* New York: Teachers College Press.

Zink, M. R. (2001, May) Case management is critical in PPS. *Home Healthcare Nurse, 19*(5), 283-288.