Journal of MultiDisciplinary Evaluation
Volume 12, Issue 27, 2016

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Who Needs Goals? A Case Study of Goal-Free Evaluation

Brandon W. Youker
Grand Valley State University

Alayna Zielinski
Hope Network

Ouen C. Hunter
Grand Valley State University

Nicholas Bayer
Wings of Hope Hospice

Background: Goal-free evaluation (GFE) is any evaluation in which the evaluator conducts the evaluation without knowledge of or reference to predetermined goals and objectives whereas the goal-based evaluator determines merit according to the evaluand's goal achievement.

Purpose: To examine a GFE in actual practice focusing on its operationalization as well as paying particular attention to the evaluation users' perspective of its utility.

Setting: The evaluand was a day long training of summer camp counselors on occupational therapy (OT) related skills such as feeding, dressing, and bathing.

Intervention: GFE was the intervention however for comparison purposes an independent and simultaneous goal-based evaluation (GBE) also evaluated the evaluand.

Research Design: Case study.

Data Collection and Analysis: After the evaluation users read both the GBE and GFE reports, data collection consisted of a semantic differential questionnaires followed by a focus group. Additionally, the research team analyzed both GBE and GFE reports for relevant themes.

Findings: The evaluation users reported a slightly more positive attitude toward the GFE report on the semantic differential yet many focus group respondents stated that they found the GBE report more useful or perceived no difference between the two. Evaluation users reported the benefits of GFE to include its potential for developing or aligning goals, expanding the pool of potential outcomes, and supplementing GBE strategies.

Keywords: *goal-free evaluation; goal-based evaluation; case study; evaluation utility; comparison.*

Introduction and Background

Goal-free evaluation is any evaluation in which the evaluator lacks the knowledge of or simply disregards the evaluand's stated goals and objectives. Rather the goal-free evaluator investigates the evaluand's actual outcomes—past and present—not its stated intentions. There is a modest body of literature on GFE (Scriven, 1973, 1974, 1991, Stufflebeam, 2001; Worthen, 1990); yet there remain several questions particularly as to how an evaluator operationalizes GFE.

To grasp the underlying tenets of GFE, it is important to understand goal-based evaluation (GBE)—sometimes referred to as objectives-oriented evaluation. With GBE the evaluator judges the program mostly according to the degree to which the program achieves its goals and objectives. Since the 1940s, evaluation has been inextricably tied to GBE and GBE continues to dominate evaluation practice (Alkin, 2004; Fitzpatrick, Sanders, & Worthen, 2004; Madaus & Stufflebeam, 1989; Patton, 1997; Scriven, 1991). Therefore, there is a plethora published on goal-based approaches and methods (e.g. 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). Friedman, Rothman, and Withers (2006) assert that "as evaluation emerged as an independent field within the social sciences, it became closely identified with the measurement of goal attainment" (p. 201). Mark, Henry, and Julnes (2000) agree that historically GBE was the dominant evaluation approach as they state:

> 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)

In contrast, during the late 1960s there were a handful of evaluation scholars such as Cronbach (1963), Scriven (1967), and Stake (1967) who started promoting evaluative inquiry beyond simple goal attainment. In their publications, they highlighted limitations associated with pre-specified goals and objectives. They argued that the assessment of goal achievement is only one component of the evaluation process as the evaluator also has a responsibility to explore side effects (Stake, 1967). As a consequence, Scriven (1972) introduced a radical concept that urged evaluators to intentionally avoid program goals and objectives. He called this approach "goal-free". Several of Scriven's (1972, 1973, 1976, 1991) subsequent publications proclaimed GFE's theoretical underpinnings and methodological strengths.

Following its introduction, there was mild interest in GFE amongst evaluation scholars yet most of the literature on the subject consisted of philosophical debates regarding its logic, strengths, weaknesses, and feasibility (e.g. Alkin, 1972; House, 1980; Irving, 1979; Salasin, 1974; Scriven, 1972, 1973, 1974, 1991; Welch, 1978). 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., 2012; Grinnell, Unrau, & Gabor, 2011; Patton, 2002). That said, the articulation of specific methods for conducting GFE remains scant; and nearly half a century since its introduction, GFE has remained conceptually abstract and highly theoretical in the minds of most evaluators with very few practitioners and even fewer who have written about it (Youker & Ingraham, 2013). Moreover, there still are only two known research studies on GFE, both doctoral dissertations (Evers, 1980; Youker, 2011). This led Shadish, Cook, and Leviton (1991) to claim that "goal-free evaluation has been widely criticized for lack of operations by which to conduct it" (p. 61). Thus, it can be reasonably concluded that those who support or oppose GFE do so primarily based on ideology rather than actual evidence, leading to the question, why without evidence, "would they [evaluators], for example, prefer one method to another?" (Tourmen, 2009, p. 7).

In determining whether an evaluation approach is worthy of consideration, evaluation scholars tend to agree that evaluators are ethically obliged to consider the programs' use of their evaluations (Joint Committee on Standards for Educational Evaluation, 1994; Patton, 1988, 1997; Scriven, 1991, 2005; Shulha & Cousins, 1997; Weiss, 1998). Therefore, an emphasis on evaluation utility 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, p. 23). The consumers of evaluation are its users, or those who fund the evaluations as well as those who are responsible for the program and for applying the evaluation findings. Davidson (2005) states the users "have invested time, effort, money, and/or egos in the design, development, and/or implementation of an evaluand" (p. 249). Furthermore, the traditional meaning of evaluation utility, according to Weiss (1998), refers to the evaluation's instrumental use or its utility for program decision making, accountability, and improvement. Examples of instrumental use include decisions to "end a program, extend it, modify its activities, [and] change the training of staff" (p. 23).

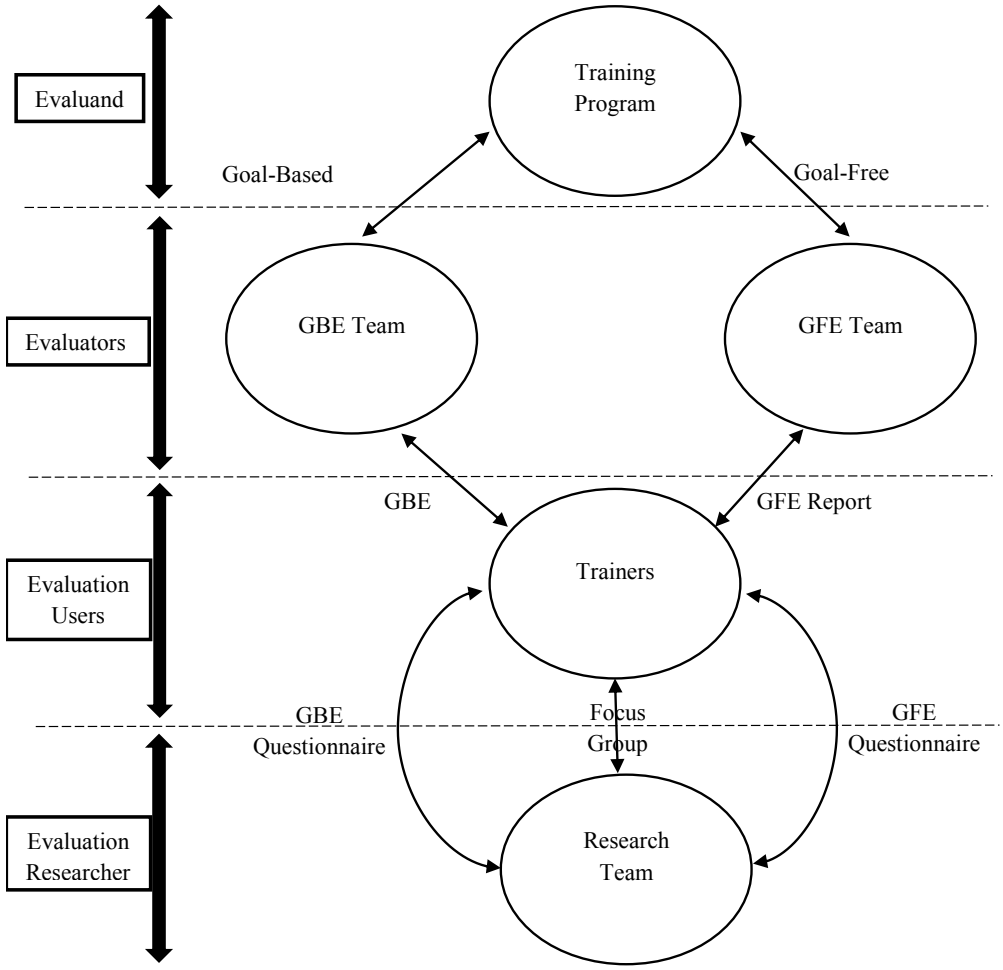

Figure 1. Relationships among the evaluand, evaluators, evaluation users, and researchers

Inadequate guidance for the evaluator in how to conduct a GFE persists as there are no published guidebooks, handbooks, or instruction manuals. There is one publication examining GFE case studies (e.g., Youker, Ingraham, & Bayer, 2014) however the authors used secondary research methods. Furthermore, there are no examinations that directly compare GBE and GFE according to their utility. A few former goal-free evaluators reported benefits realized in practice (Manfredi, 2003; Stufflebeam, 2001; Stufflebeam & Shinkfield, 2007; Stufflebeam & Coryn, 2014; Thiagarajan, 1975); however, there are no reports of these benefits from the actual users of the GFE. Therefore, there are three specific objectives investigated in this case study; they are as follows:

- What are the methods and procedures employed by this GFE team?
- From the perspective of evaluation users, is there a difference between GBE and GFE with regard to evaluation utility?

- What, if any, do the users perceive as benefits of GFE?

Methodology

This IRB-approved study employs a case study methodology providing a reflective narrative offering insight into practice specifically regarding GFE techniques and methods in attempt to extrapolate lessons learned as well as principles for conducting GFE. This case study also included a simultaneous and independent GBE for comparing with the GFE. Two evaluation teams consisted of graduate students enrolled in a program evaluation course. One team was trained in GBE and the other in GFE. The teams worked separately and simultaneously yet evaluated the same evaluand: a training program. The teams independently produced their own evaluation reports to present their findings and conclusions. After the evaluation users read a report, they were asked for their perceptions regarding the usefulness of the report and its findings; this process was repeated with the second evaluation report.

Figure 1 illustrates the relationships among the training program, the two evaluation teams, the evaluation users, and the researchers of the evaluations' utility.

There were three methods used to investigate the primary research questions:

1. Questionnaire: Evaluation users completed structured questionnaires; the first questionnaire asked evaluation users' for their attitudes toward the GBE report and the identical second questionnaire asked for attitudes toward the GFE report.
2. Focus Group: A follow up focus group with evaluation users sought further elaboration on their perceived differences between the GBE and GFE reports.
3. Content Analysis: An analysis of the evaluation reports was conducted to compare their designs and methods.

Both teams were required to submit an evaluation report thus permitting an analysis of the similarities and differences between the two evaluation approaches as is discernable from the reports. A type of attitude survey called a semantic differential was administered to the evaluation users to assess general attitudes toward each evaluation report. Semantic differential rating scales were chosen as they are relatively easy to design, administer, and interpret and have been proven reliable (Himmelfarb, 1993; Osgood, Suci, & Tannenbaum, 1957; Powell, 1982) as well as have been demonstrated as applicable for judging evaluation reports (Evers, 1980). The semantic differentials consisted of bipolar adjective pairs used to describe an evaluation report. For example, evaluation users indicated their attitudes on a continuum by rating adjective pairs such as useless-useful, worthless-worthwhile, and biased-objective; an example adjective pair as it appeared on the questionnaire is displayed in Figure 2. For analysis purposes, numeric values were assigned to the response options (-3 to +3); an adjective pair with the values added is presented in Figure 3. The instrument was then pilot-tested with 15 students in an evaluation doctoral program as well as with a masters-level social work research course; consequently the original list of over 80 adjective pairs was reduced to 25 pairs. To account for the order effect, the research team created and used three different questionnaire versions with the sole difference among the versions being the order in which the adjective pairs were presented. The final item on the instrument was an open-ended question asking the respondents to describe their opinions as to whether the report was or was not useful. This final question employed an emic perspective allowing the evaluation users to define 'useful' per their discretion. Lastly, three expert evaluators, all of whom were former American Evaluation Association presidents, reviewed the questionnaire and gave their suggestions and ultimate approval of the instrument.

Figure 2. Example bipolar adjective pair from the semantic differential questionnaire

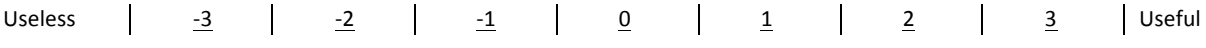

Figure 3. Example adjective pair with numeric values assigned for analysis

The raw data were transposed to further enable proper analysis of the semantic differentials. After data were transformed, they were imported into SAS 9.3 Enterprise for further analysis. PROC NPAR1WAY using the Empirical Distribution Function (EDF) was used to analyze the transposed dataset.

A week after the semantic differentials were administered and analyzed, evaluation users participated in a follow-up video recorded focus group. The purpose of the focus group was to elicit a more descriptive account of the evaluation users' attitudes by asking for explanations as to their reactions to the adjective pairs with the highest and lowest mean scores as well as those with the greatest standard deviations. The hour long focus group was transcribed in its entirety using denaturalized transcription and then coded and indexed to draw emergent themes and categories.

The content analysis began upon receipt of the evaluation reports. The research team first individually analyzed each report and then discussed and combined their analyses. The purpose of the content analysis was to look for notable similarities and differences in terms of the evaluation designs employed by the evaluators as well as distinctions in terms of the evaluations' criteria, standards, measurement, and synthesis processes. These four principles, described in Fournier's (1995, 2005) logic of evaluation, served as a justifiable and neutral rubric for comparing the evaluation reports as this general logic has been deemed one of the primary ways evaluators conceptualize evaluation (Julnes, 2012; Kundin, 2010).

Subject Selection and Characteristics

Three sets of subjects were necessary to conduct this case study: (1) the evaluand, (2) the evaluation users, and (3) the evaluators. The evaluand was a training program provided by occupational therapy (OT) graduate students where the student-trainers instructed counselors at a summer camp for individuals with ability needs on several OT-related techniques such as feeding, dressing, and transferring. The program was selected via convenience sampling. Two historical characteristics of this training program lent themselves to a field-based investigation of evaluation: (1) the program's maturity and (2) the program's relationships with the principal investigator. First, the OT department provided the training at the camp for the previous seven years thus the program and relationships were well established. Previous informal internal monitoring efforts investigated the training and assessed some of its outcomes; and therefore, the training program and camp administrators were willing to examine a potentially broader range of criteria and outcomes for this evaluation initiative. Second, not only had the camp been working with the OT department but it was also separately working with this study's principal investigator and prior semesters of evaluation students. Camp administrators reported positive experiences working with both; thus, to a degree, working relationships and rapport were already established. Lastly, the faculty members who instructed the OT student-trainers and the study's principal investigator had a connection as they hail from the same Master's Large University in the Midwest. These relationships along with the maturity of the OT training program likely influenced the camp administrators and OT faculty members in their evaluation preparedness, their willingness to be involved with this case study, and their willingness to incorporate the two evaluation approaches.

The evaluation users consisted of a cohort of 30 third-semester OT master's students. Per the requirements of their Adult Practice OT course, the students were required to not only design, implement, and reflect on the training of over 40 camp staff, but they were responsible for using the evaluations' findings to restructure future trainings. The OT students had a mean cumulative GPA of 3.89 (see Table 1) and only two of the students were male.

Table 1
Occupational Therapy Student-Trainer Demographics

	Graduate Credit Hours Completed	Cumulative GPA
M	30.40	3.89
SD	1.52	0.08

The goal-based and goal-free evaluators were responsible for designing, conducting, and reporting on the program evaluations. The 13 student-evaluators were enrolled in a graduate social work course on program evaluation. Random assignment designated students to either the GBE team or the GFE team. Tables 2 and 3 compare the student-evaluators according to gender, number of graduate credit hours completed, cumulative GPA, and GPA in their graduate-level research courses.

Table 2
Goal-Based Evaluator Demographics

Gender	Graduate Credit Hours Completed	Cumulative GPA	Research I GPA	Research II GPA
F	0	-	*	-
M	18	3.67	4.00	-
F	27	3.12	2.66	-
F	0	-	*	-
F	51	3.92	3.66	4.00
F	0	-	*	-
F	54	3.79	4.00	4.00
M	21.43	3.62	3.58	4.00
SD	23.65	0.35	0.63	-

Note: *Indicates an advanced standing student for whom the Research I requirement was waived.

Table 3
Goal-Free Evaluator Demographics

Gender	Graduate Credit Hours Completed	Cumulative GPA	Research I GPA	Research II GPA
F	0	-	*	-
F	0	-	*	-
F	64	3.37	4.00	-
F	21	3.61	3.33	-
F	0	-	*	-
F	45	3.81	3.66	4.00
M	21.67	3.60	3.66	4.00
SD	27.37	0.22	0.34	-

Note: *Indicates an advanced standing student for whom the Research I requirement was waived.

Fidelity to the Approach

Fidelity is defined as the extent to which delivery of an intervention abides by the protocol or program model that was originally developed (Mowbray, Holter, Teague, & Bybee, 2003); and in this case, the particular evaluation approaches (i.e. GBE and GFE) are the interventions. There were several strategies used to ensure fidelity to the goal-based and goal-free approaches.

Training handbooks were created and distributed to each team. The GBE handbook described the assigned evaluation approach which included a dos and don'ts checklist (Youker, 2011); a log for recording threats to evaluator independence as well as threats to the goal-based nature of the evaluation; and a goal-based approach fidelity checklist. In attempts to outline an idealized version of GBE, the investigator provided four principles to guide the evaluation team. The goal-based evaluators were to adhere to the general principles articulated by Youker which are:

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

The GFE handbook contained a log for recording threats to evaluator independence as well as threats to the goal-free nature of the evaluation. Youker (2013) published a GFE fidelity checklist, a dos and don'ts checklist, a list of materials and documents to be screened due to their likelihood of containing goal-related information, and the following principles for conducting GFE:

1. Identify relevant effects to examine without referencing goals and objectives
2. Identify what occurred without the prompting of goals and objectives
3. Determine if what occurred can logically be attributed to the program or intervention
4. Determine the degree to which the effects are positive, negative, or neutral. (p. 434)

A formal pilot-testing of the approach fidelity checklists was not feasible because of resource limitations and the lack of known GFE practitioners. Instead, the principal investigator generated criteria for approach fidelity by reviewing the literature on both approaches and sought expert opinion on the initial list of items for inclusion and exclusion. After the initial list of dos and don'ts for each approach was established, the investigator requested that over a dozen selected evaluation experts assess the importance of each ingredient to determine which are in fact essential for differentiating between a GBE and a GFE.

The course instructor trained and supervised both evaluation teams to ensure evaluation quality and to maintain fidelity to the assigned evaluation approach. During class, the students received combined lectures on the general logic of evaluation while separated into their respective evaluation teams for various worksheets, exercises, discussions, and additional readings. In fact, the instructor dedicated two course periods for learning about, discussing, practicing, and planning just for the GBE evaluation approach while another two periods focused on GFE instruction only. Lastly, all student-evaluators signed a contract stating that they would reasonably and ethically attempt to maintain fidelity to the assigned evaluation approach and that they would avoid commingling and communicating with evaluators from the other team regarding the evaluations.

During the evaluation phase, the GBE and GFE teams designed and conducted their evaluations and wrote their reports. Throughout the duration of the evaluations, the course instructor communicated with each team weekly to supervise, to serve as a liaison between the teams and the training program, to answer evaluation-related questions, to screen the goals from the GFE team, and to ensure fidelity to evaluation approaches. Both the students and the course instructor assessed the evaluators' adherence to the approach fidelity checklist regularly throughout the semester. Teams submitted their logs and reports and then the investigator, a graduate assistant, and an undergraduate student researcher edited the reports eliminating blatant errors, ambiguities, and stylistic differences.

Findings

The level of measurement for the study is ordinal thus a non-parametric analysis was chosen to determine if there were any differences in the distribution of GBE versus GFE. The non-parametric statistics of EDF using the two-sample Kolmogorov-Smirnov (K-S) test investigated whether there were any distribution differences between GBE and GFE in terms of the cumulative ratings of 30 evaluation users who rated each evaluation approach. Researchers did not assign identifiers to evaluation users thus pairings were not possible; therefore the two groups are independent of each other.

There are a total of 750 observations for each evaluation approach because there are 25 adjective pairs and 30 participants who responded to both evaluation reports. There are four missing scores from GBE and one from GFE. Medians for both evaluation approaches are at 2 but the higher mean for GFE indicates that evaluation users who rated the GFE report tended to rate the adjective pairs in a more positive manner. The minimum score further demonstrates this whereas the minimum for GFE is -1 compared to the minimum for GBE which is -3. Maximum scores for both evaluations are 3 (see Table 4).

Table 4
Descriptive Statistics for Goal-Based Evaluation and Goal-Free Evaluation Teams

Type	*N* Obs	*N*	*N* Miss	*M*	*Med*	Mode	Minimum	Maximum
GBE	750	746	4	1.91	2.00	2.00	-3.00	3.00
GFE	750	749	1	2.00	2.00	3.00	-1.00	3.00

Utilizing the K-S statistics, there is a marginally significance difference in the distribution between GBE and GFE with a Pr>KSa of 0.0736 at α = 0.05 (Table 5). The marginal significance indicates that the evaluation users who reviewed the GFE were more likely to rate the adjective pairs in a more positive manner where the evaluation users who rated the GBE report tended to rate them more negatively. A diagram (Figure 4) of the findings suggests that although both evaluation approaches have high positive ratings, the differences occurred in the negative ratings where more evaluation users rated GBE negatively as compared to GFE.

Table 5
Kolmogorov-Smirnov Two-Sample Test (Asymptotic)

KS	0.0332	D	0.0664
KSa	1.2851	Pr>KSa	0.0736

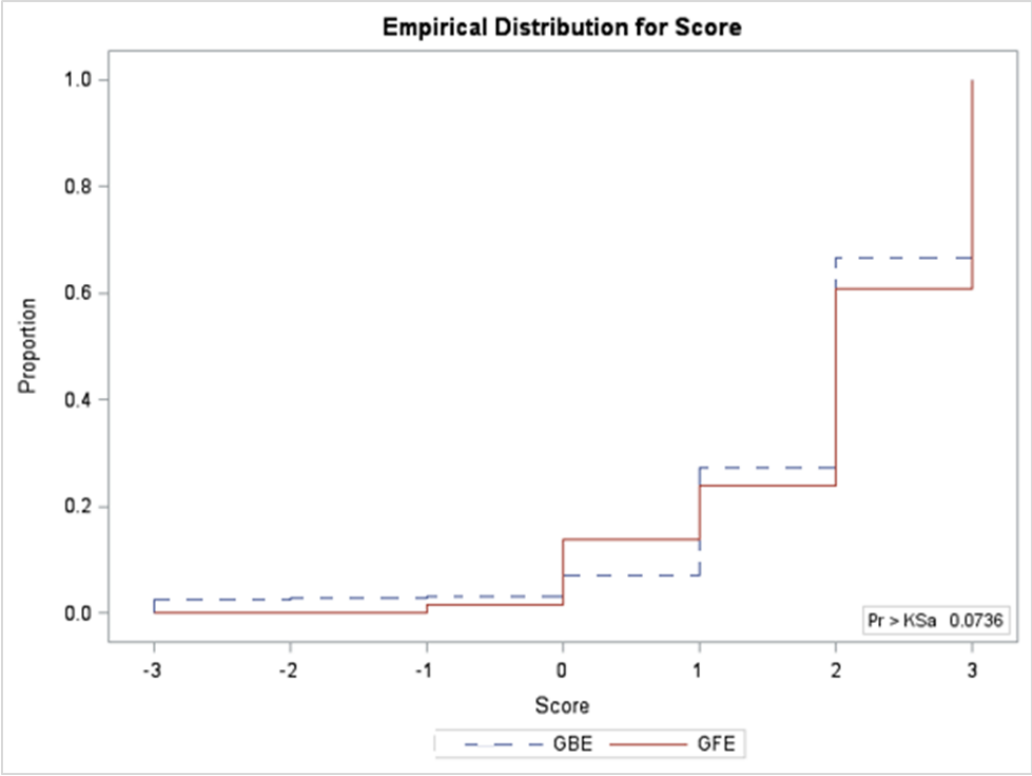

Figure 4. Empirical distribution function (EDF) plot on ratings/scores

Figure 5 presents the distributions and the differences in ratings between GBE and GFE. The 30 evaluation users chose the most positive adjectives 291 times when rating the GFE report while the same 30 evaluation users only marked the most positive adjectives in 248 instances when assessing the GBE report. With marginal significance, respondents assessing the GFE report rated the adjectives more positively compared to the GBE report. Table 6 displays the fact that although there are not many combined negative ratings, GBE (n = 24) received more negative ratings than GFE (n = 13).

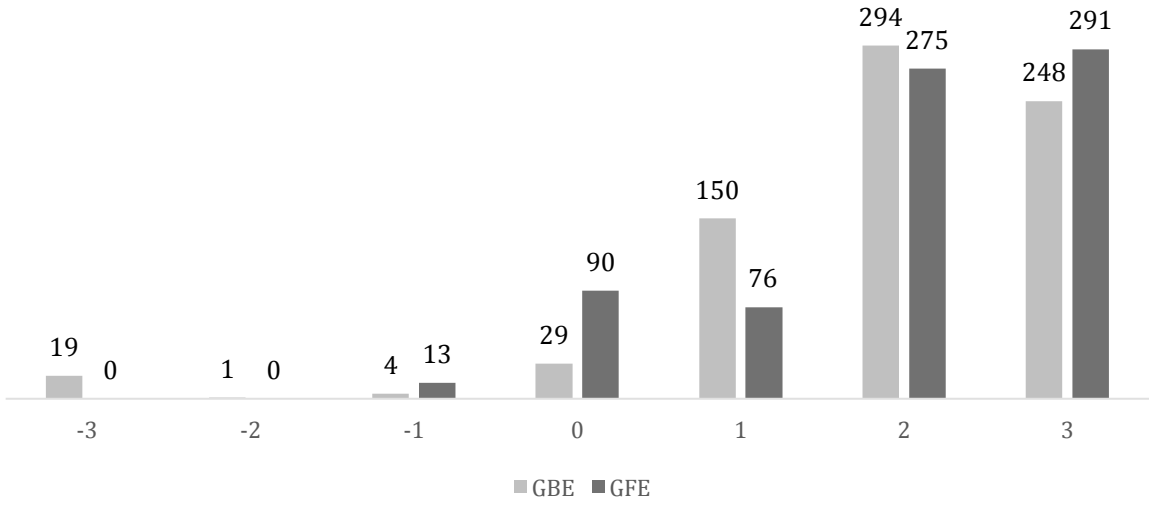

Figure 5. Distribution of ratings on goal-based evaluation and goal-free evaluation

Table 6
Count Distribution of Ratings for Goal-Based Evaluation and Goal-Free Evaluation Reports

Adjective	Ratings	GBE	GFE
	3	248	291
Positive	2	294	275
↕	1	150	76
	0	29	90
	-1	4	13
	-2	1	0
Negative	-3	19	0
	Total	745	745

Figure 6 displays the evaluation users' mean scores per adjective pair for both evaluation reports. The mean across all adjectives pairs for the GBE report is 1.91 with an average standard deviation 0.66 while the mean for the GFE report is 2.0 (*SD* = 0.42). The evaluation users found the biggest difference between the two reports on the adjective pair unfair-fair and careless-careful where respondents felt the GFE report was noticeably more fair and more careful than the goal-based report. Although a few evaluation users assigned a negative rating to some of the adjective pairs, no individual adjective pair from either evaluation report garnered a negative mean score.

In assessing the GBE report, the evaluation users assigned the highest rating on the adjective pairs unclear-clear (*M* = 2.07, *SD* = 1.17), irrelevant-relevant (*M* = 2.07, *SD* = 1.26), unhelpful-helpful (*M* = 2.03, *SD* = 1.19), and incomplete-complete (*M* = 2.03, *SD* = 1.27). The lowest mean scores were assigned to unfair-fair (*M* = 1.60, *SD* = 1.28), inconclusive-conclusive (*M* = 1.69, *SD* = 1.28, ineffective-effective (*M* = 1.70, *SD* = 1.21), and careless-careful (*M* = 1.70, *SD* = 1.12). The highest rated adjective pairs on the GFE are unbelievable-believable (*M* = 2.17, *SD* = 1.04), inconsistent-consistent (*M* = 2.13, *SD* = 1.11), and uninformative-informative (*M* = 2.13, *SD* = 1.11) with the lowest mean scores assigned to unbalanced-balanced (*M* = 1.80, *SD* = 1.19), inconclusive-conclusive (*M* = 1.87, *SD* = 1.14), and worthless-worthwhile (*M* = 1.87, *SD* = 1.22).

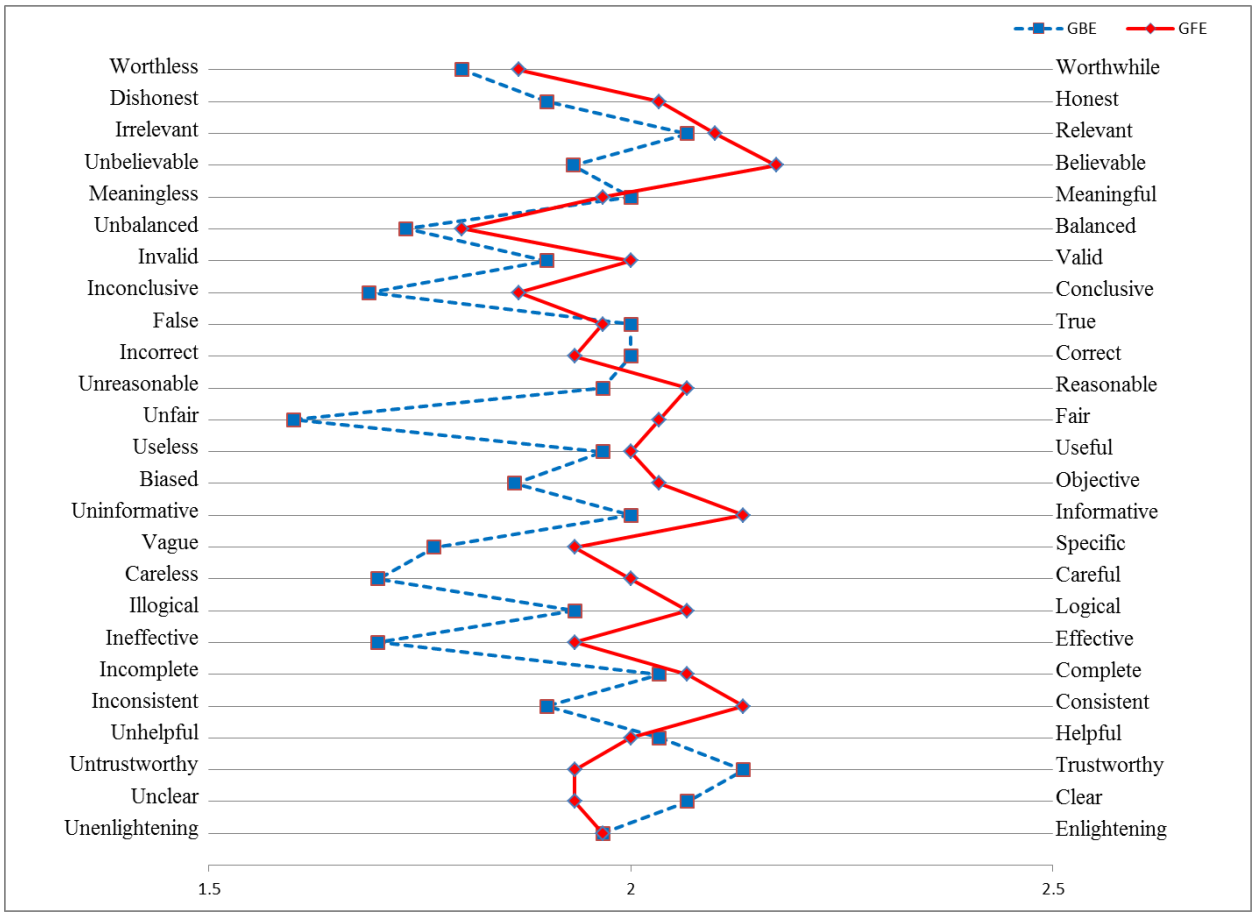

Figure 6. Evaluation users' mean scores on the semantic differential

Many of the responses on the questionnaire's open-ended question referred to writing style and formatting such as liking or disliking the inclusion of the table of contents, charts, and/or rubrics; in addition, there were several comments agreeing or disagreeing with specific evaluation findings and recommendations. Yet there were a couple of substantive open-ended responses. For instance, one respondent found the GBE report useful for its "clear explanation of goals and results" while another wrote, "I preferred the goal-based report for better reflecting how and where we are to improve our trainings." During the course of the evaluation, the goal-free evaluators elected to estimate what they believed to be the training program's goals and objectives based on the OT students' actual training; and one evaluation user recognized the potential for aligning the program's goals by writing that the GFE report was "very helpful" for restructuring the program's goals. There were some evaluation users who were critical of the GFE report. A couple respondents felt that the GFE report included trivial information; for example, one user remarked that since the goal-free evaluators "did not know what OT students were trying to emphasize, [the evaluation] results seem to focus on minor details, not actual goals."

Twenty-six of the 30 evaluation users participated in the follow-up focus group the results of which somewhat conflicted with the semantic differential results as 14 of the 26 focus group respondents reported that holistically they preferred the GBE report. A common sentiment was that they "knew how to read" a report like the goal-based one as they "knew what to look for." Others felt that because the report's purpose was clearer, there was more clarity and applicability in the evaluation's findings. However, also during the focus group, nine of the 26 evaluation users claimed to be in favor of the GFE report. The GFE team examined broad serendipitous outcomes and this was the basis for some evaluation users' preference for GFE; one respondent stated that goal-free nature of the evaluation allowed for investigating "global abilities rather than specific skills." On the semantic differential, the evaluation users rated the GFE report as being highly

informative and one focus group respondent summarize this perspective by stating:

> I can speak to my thought at least on why the goal-free was informative. I think it's because they weren't so focused on the goals, they were more comprehensive, and [included] a lot more observations. Like no one was going to expect the gait belt situation, but they noted that we were flexible and we did that; and I think because they weren't so focused on goals, they [goal-free evaluators] can include those things that came up that day.

Other focus group respondents found neither report more useful than the other; for instance, one OT student stated, "I felt like when I read through one and then read through another online, they were so similar that I just couldn't really separate them." Another evaluation user called the two reports roughly "the same… because the information was so similar." Yet another evaluation user articulated this point during the focus group as she noticed the similarity in evaluation findings and questioned whether knowing the goals of the program prior to the training was even necessary; this respondent stated:

> I thought it was interesting that even though one [evaluation team] had the goals and the other one didn't, they still came to some of the same conclusions and a lot of it was very similar. So that sheds light on, you know, do they need to know the goals beforehand to even reach the same conclusions?

A comparative analysis of the GBE and GFE reports (see Table 7) leads to the conclusion that both evaluation teams conducted their evaluations comparably. Both the GBE and GFE used pre-experimental designs, specifically a one-shot case study design which consists of an intervention (i.e. OT training program) followed by an observation (i.e. program evaluation). The GBE team identified five criteria for judging the training program while the GFE teams had six. Both evaluation teams recognized the training's organization as a critical characteristic in addition to its ability to impart knowledge or educate the trainees. The goal-based team created 13 grading rubrics for comparing the program's performance while the GFE report only included one rubric used with a sole criterion. In terms of the evaluators' data collection, both teams employed direct observation, semi-structured interviews, and structured questionnaires. The two teams used two pre-existing evaluation instruments, a pretest-posttest developed during a prior evaluation and a posttest-only questionnaire created previously by OT faculty. The goal-free team refrained from examining or analyzing any of the results from the structured surveys until after the completion of their data collection. The only differences between the two evaluation teams' data collection methods of consequence were that the goal-based team used an emailed qualitative survey while the goal-free team employed an observation checklist. The GBE team did not include a synthesis of the criteria, standards, and measures instead choosing to profile criteria whereas the GFE team used numeric weight and sum (see Davidson, 2005) as its method of synthesis.

Table 7
Content Analysis of the Evaluation Reports

	Goal-Based Evaluation	Goal-Free Evaluation
Evaluation Design	One-shot case study design	One-shot case study design
Criteria	1. Safety 2. Trainer competence 3. Efficient organization 4. Comprehensiveness 5. Knowledge retention and application	1. Understandable 2. Educational 3. Organized 4. Approachable 5. Flexible 6. Applicable
Standards	Thirteen 4-level grading rubrics across all 5 criteria	One 5-level grading rubric for the criterion *applicable*
Measurement/ Observation	• Direct observation (4 site visits) • Face-to-face semi-structured interviews (9 respondents) • Emailed qualitative questionnaire (1 program director) • Student-evaluator created structured questionnaire on trainee comfort level, knowledge retention, and opinions (15 respondents) • Qualitative questionnaire (12 respondents) • Pre-existing pretest-posttest on the dressing, feeding, and transferring techniques (38 respondents) • Pre-existing structured posttest-only questionnaire on the feeding training (36 respondents)	• Direct observation (3 site visits) • Face-to-face semi-structured focus group (7 respondents) • Face-to-face semi-structured interview (1 respondent) • Observation checklist • Student-evaluator created structured questionnaire on the trainees' application of training (30 respondents) • Pre-existing pretest-posttest on the dressing, feeding, and transferring techniques (38 respondents) • Pre-existing structured posttest-only questionnaire on the feeding training (36 respondents)
Synthesis	Each criterion profiled but no overall synthesis to combine criteria into one evaluative conclusion	Each criterion profiled and then synthesized using numeric weight and sum to combine criteria into one evaluative conclusion

Discussion

Research Question One

The first research question posed in this case study was: What are the methods and procedures employed by this GFE team? Table 8 summarizes some of the demographic methodological and procedural characteristics of this GFE. The GFE team employed several techniques described in the literature and abided by suggestions made by prior goal-free evaluators; described below are three of these characteristics.

First, this GFE used a designated goal screener which several GFE scholars have discussed (Evers, 1980; Matsunaga & Enos, 1997; Scriven, 1972, 1973; Stufflebeam & Shinkfield, 2007; Welch, 1976). The screener for the GFE team was the course instructor. The evaluation course instructor served as the liaison between the goal-free team and the program administrators and staff to eliminate goal and objective-related communiqués. Additionally, the course instructor also helped maintain independence between the goal-based and goal-free teams.

Second, like other goal-free evaluators before them (e.g., Berkshire, Kouame, & Richardson, 2009; Matsunaga & Enos), the goal-free evaluators relied primarily on qualitative data collection methods to gather data from the training's consumer: the camp staff. The goal-free evaluators directly observed training demonstrations as well as made follow-up visits to observe the OT techniques applied in actuality; the evaluators conducted focus groups with camp staff trainees; and distributed an open-ended self-administered questionnaire to camp staff. The evaluators also chose to conduct a face-to-face semi-structured interview with a program administrator. Additionally, like prior evaluators (Scriven, 1974; Youker, 2005), the goal-free evaluators employed an observation checklist to assess training outcomes on camp staff. Two-thirds of the

checklist development occurred prior to the training while the remainder of the checklist creation originated during the OT students' training session. The goal-free team used a structured instrument where training participants completed a Likert scale survey asking them to rate their perceived ability to apply the material presented during the training.

Third, the GBE and GFE operated simultaneously yet separately from each other. This evaluation strategy adheres to Scriven's (1991) position that GFE supplement GBE. In fact, Youker et al. (2014) presented four case studies of GFEs and all four of these GFEs were used in conjunction with a goal-based strategy.

Table 8
Characteristics of the Goal-Free Evaluation

Characteristic	Description
Program evaluated	Training program
Program type/ key interventions	A training of camp staff on basic OT-related techniques designed to meet common needs of campers
Program partnership	OT faculty with third-semester OT Masters students have provided iterations of the training at this camp for 7 years
Program dates and duration	The OT students trained the camp staff on May 29 from 8:30 a.m. to 4:30 p.m.
GFE team members	6 MSW student-evaluators
Pre-evaluation relationship between program and evaluator(s)	A relationship was established between the evaluation course instructor and the OT faculty for the purposes of this study; the evaluation course instructor also visited the OT class to explain the evaluations and answer questions; the student evaluators had no interaction with the OT student-trainers prior to the day of the training
GBE-GFE relationship	Two evaluation teams conducted simultaneous yet independent evaluations: a GBE team and a GFE team
Screener	The evaluation course instructor served as goal screener
Evaluation type	Designed to be a formative, outcome-based evaluation
Data collection methods	Direct observation with an observation checklist, semi-structured focus group, structured questionnaire on knowledge, pre-existing pretest-posttest, and pre-existing structured posttest-only questionnaire
Primarily data sources	The camp staff trainees, OT student trainers, student-evaluator observations
Sampling and Sample size	The intended sample varied depending on the instrument; pre-existing instruments like the pretest-posttest were distributed to nearly every camp staff person who attended the full training while the observations and the focus group had fewer respondents
Approx. time spent in data collection	Following the approval of the evaluation proposal, students had approximately four weeks to complete the GFE's data collection

Research Question Two

The second research question posed in this case study was: From the perspective of evaluation users, is there a difference between GBE and GFE with regard to evaluation utility? The overall conclusion is somewhat uncertain. The semantic differential results provide evidence that the evaluation users felt the GFE report had slightly more utility than the GBE report; however, during the focus group the majority of evaluation users reported that they either preferred the GBE report or found negligible distinction between the two. Moreover, a conclusion from content analysis is that both evaluation reports are similar which likely affected the evaluation users' abilities to differentiate between the reports.

Nevertheless, the results of this case study are consistent with the previous findings that the utility of GBE and GFE does not significantly differ (Evers, 1980; Youker, 2011). Despite the fact that there is insufficient evidence for claiming a significant difference in overall utility between the goal-based and goal-free reports, this does not mean that there were no perceived differences. On the contrary, evaluation users reported differences. It is just inconclusive as to whether these characteristics led to differences that evaluation stakeholders can meaningfully experience and whether these differences directly relate to the GBE or GFE approaches.

Research Question Three

The third research question posed in this case study was: What, if any, do the users perceive as benefits of GFE? Some OT student-trainers perceived that GFE added value in the following three areas: 1) serving in program goal alignment, 2) expanding the pool of potential outcomes, and 3) triangulating with GBE. These three reported benefits are congruent with previously claimed benefits of GFE (Youker & Ingraham, 2014).

First, GFE can be useful for aligning the program goals with its actual activities. According to Patton (1997), a potential product of a GFE can be "a statement of operating goals" (p. 182) as some goal-free evaluators elect to hypothesize what they believe to be the program goals. One evaluation user and respondent summarized this by stating: "The [goal-free] evaluation was very helpful because it provided useful information regarding how goals can be better organized."

Second, as Scriven (1972, 1991) and James and Roffe (2000) have suggested, a potential benefit of GFE is that there is more opportunity for in situ discoveries and unanticipated or serendipitous outcomes that may not have been recognized if exclusively using a GBE evaluation. In fact, some OT student-trainers liked GFE's freedom for uncovering and exploring a wide range of outcome possibilities since they were not restricted to searching for predetermined outcomes from predetermine sources. Thus, they claimed that there was less of a tendency to look for very specific skills rather than "global abilities," as one OT student-trainer put it.

Third, GFE is a form of evaluation triangulation in that it serves as an independent assessment that can either assist in confirming or contradicting the findings of a GBE. It is clear that this GFE supplemented the findings from the GBE allowing for a more comprehensive review of the training program activities and outcomes. Finally, the methods used for this GFE are consistent with methods that have been reported with other GFEs, namely key stakeholder interviews and the use of pre-existing instruments and direct observation (Berkshire, Kouame, & Richardson, 2009; House, 1980; Matsunaga & Enos,1997; Welch, 1978; Youker et al., 2014)

Limitations

Limitations temper results. 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 study. Below is a non-exhaustive list of several limitations of this study.

A limitation when using real programs in a case study of this nature is that the administrator's willingness to participate is probably systematically different than the organization or program that is not willing to participate. The willing program may be more mature, more evaluation savvy, and more confident in its performance and outcomes. Relatedly, this inquiry is also susceptible to social threats to internal validity; for example, in a study of this type, the ever-present Hawthorne effect (i.e. reactivity) on behalf of the training program facilitators—who are also the evaluation users—is an unavoidable potential limitation.

This examination controlled various aspects of the evaluations but not others. For example, this study controlled the evaluation approach and goal-orientation as well as certain parameters of the report such as its format, headings, and page limits, for example. However, there was no attempt to control or manipulate the evaluators' choice of data collection methods nor was there an attempt to manipulate the training program's outcomes.

There were limitations based on the fact that this was somewhat a simulation using graduate students as evaluators. Student-evaluators' motivations and incentives differ from real-world professional evaluation practice. The fact that student-evaluators received graduate school credit, juggled other coursework, and evaluated without financial compensation diverges from professional evaluation practice. Thus, it is arguable that the student-evaluators were not representative of real evaluators with real prior evaluation experience practicing in a real evaluation environment with real evaluation consequences.

This study collected perspectives from only one of the training program's stakeholders, namely the OT student-trainers. Although they were the primarily evaluation users who were responsible for using the evaluation results, other stakeholders' opinions may prove valuable. The participating OT instructors declined the researchers' interview requests, instead deferring to their students. The camp administration and the staff who received the OT training are both an incredibly important evaluation consumers; however the GFE researchers neglected to interview them. The reason the training and the evaluation exist is to satisfy the campers' needs; however, the research team did not consult with the campers or their families as both groups were beyond the scope of the evaluations' primary

users. Nevertheless, future studies should examine utility from the perspectives of these other stakeholder groups particularly the evaluation consumers.

Probably the most significant limitation of this case study is that it includes a small *n* posttest-only examination which essentially negates external validity. However, the case study methodology does offer an in depth examination and description of the GBE and GFE approaches.

There are limitations based on the findings, particularly the seemingly modest effect size and the contradictions in findings between the quantitative and qualitative data analyses. There is too small of an effect size to state definitively whether there is a difference in utility between the evaluation approaches. Despite the limitations, the evaluation users generally concluded that despite minor differences, the two evaluation teams came to relatively similar conclusions as to the quality of the OT training. This serves to validate both of the evaluations' findings and conclusions.

Conclusion

This study accomplished its goals of describing GFE methodologically and procedurally; discussing several differences between GBE and GFE in actual practice; and reporting some actualized benefits of GFE. Evaluation users did not perceive either evaluation as significantly more useful than the other—which in itself is a meaningful finding. Nevertheless, it is important to understand that it is not the purpose of this inquiry to pit GFE against GBE to claim one superior, as Patton (1997) reminds us, "evaluation will not be well served by dividing people into opposing camps: pro-goals versus anti-goals evaluators" (p. 184); rather, the purpose of this study is to examine the evaluation users' perspectives as to how they experienced GBE and GFE utility differently while learning more about GFE in practice. Numerous evaluators prescribe their particular evaluation preferences and practices (Tourmen, 2009) but there are not near enough actual studies on evaluation (Christie, 2012; Coryn et al., 2016). The findings from this case study justify further study of GFE and just maybe it will inspire other evaluation scholars to systematically examine GFE too.

References

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. (2004). Comparing evaluation points of view. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influence* (pp. 3-11). Thousand Oaks, CA: Sage.

Berkshire, S., Kouame, J., & Richardson, K. K. (2009). Making It Work: Evaluation report. Kalamazoo, MI: Western Michigan University, The Evaluation Center.

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.

Campbell, S., & Stanley, J. C. (1963/1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Chen, H., & Rossi, P. (1983). Evaluating with sense: The theory driven approach. *Evaluation Review, 7,* 283-302.

Christie, C. A. (2012). Advancing empirical scholarship to further develop evaluation theory and practice. *The Canadian Journal of Program Evaluation*, *26*(1), 1-18.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

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.

Coryn, C. L. S., Ozeki, S., Wilson, L. N., Greenman II, G. D., Schröter, D. C., Hobson, K. A., Azzam, T., & Vo, A. T. (2016). Does research on evaluation matter? Findings from a survey of American Evaluation Association members and prominent evaluation theorists and scholars. *American Journal of Evaluation*, *37*(2), 159-173.

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.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Evers, J. W. (1980). *A field study of goal-based and goal-free evaluation techniques*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

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-Boss.

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.

Grinnell, R. M,, Unrau, Y.A., & Gabor, P. A. (2011). Program evaluation. In R. Grinnell & Y. Unrau (Eds.) *Social work research and evaluation: Foundations of evidence-based practice,* (9th ed). (pp.521–529). New York: Oxford University Press.

Hellström, T. & Jacob, M. (2003). Knowledge without goals? Evaluation of knowledge management programmes. *Evaluation, 9*(1), 55-72.

Himmelfarb, S. (1993). The measurement of attitudes. In A.H. Eagly & S. Chaiken (Eds.), *Psychology of Attitudes*, (pp. 23-88). Thomson/Wadsworth.

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.

Irving, J. F. (1979). Goal-free evaluation: Philosophical and ethical aspects of Michael Scriven's Model. *CEDR quarterly, 12*(3), 11-14.

James, C. & Roffe, I. (2000). The evaluation of goal and goal-free training innovation. *Journal of European Industrial Training, 24*(1), 12-20.

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.

Julnes, G. (2012). Managing valuation. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New Directions for Evaluation*, 133, 3–15 .

Kundin, D. M. (2010). A conceptual framework for how evaluators make everyday practice decisions. *American Journal of Evaluation, 31*(3), 347-362.

Madaus, G. F., & Stufflebeam, D. L. (1989). *Educational evaluation: Classic works of Ralph W. Tyler*. Boston: Kluwer Academic Publishers.

Manfredi, T. C. (2003). Goal based or goal free evaluation? Growing New Farmers website. http://www.smallfarm.org/uploads/uploads/Files/Goal_Based_or_Goal_Free_Evaluation.pdf

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.

Matsunaga, D. S., & Enos, R. (1997). Goal-free evaluation of the Alger Foundation's Wai'anae Self-help Housing Project: Evaluation report. Kalamazoo, MI: Western Michigan University, The Evaluation Center. Rog, D. J. (2005). Design, evaluation. In S. Mathison

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.

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.

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

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 Publications.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

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.

Powell, E. R. (1982). Sociometric semantic differential assessment. *Small Group Research, 13*(1), 43-52.

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 Publishing Corporation.

Scriven, M. (1974). Prose and cons about goal-free evaluation. In W.J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 34-67). Berkeley, CA: McCutchan Publishing Corporation.

Scriven, M. (1976). Evaluation bias and its control. In G. V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1) (pp. 101-118). Newbury Park, CA: Sage.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage Publication.

Scriven, M. (2005). 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

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.

Stake, R. E. (1967). The countenance of educational evaluation. *Teacher College Record, 68,* 523-540.

Stufflebeam, D. L. (2001). The metaevaluation imperative. *American Journal of Evaluation, 20*(2), 183-209.

Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, & applications* (2nd ed.). San Francisco, CA: Jossey-Bass.

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

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.

Thiagarajan, S. (1975). Goal-free evaluation of media. *Educational Technology, (15)*5, 38-40.

Tourmen, C. (2009). Evaluators' decision making: The relationship between theory, practice, and experience. *American Journal of Evaluation, 30*(1), 7-30.

Vedung, E. (1997). *Public policy and program evaluation*. New Brunswick, NJ: Transaction Publishers.

Welch, W. W. (1976). "Goal Free Evaluation Report for St. Mary's Junior College" (Unpublished Report) Minneapolis.

Welch, W. W. (1978, March). *Goal-free formative evaluation – an example*. Paper presented at the meeting of the American Education Research Association, Toronto, Ontario, Canada.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation*, *19*(1), 21-33.

Worthen, B. (1990). Program evaluation. H. Walberg & G. Haertel (Eds.), The international encyclopedia of educational evaluation (pp. 42-47). Toronto, ON: Pergammon Press.

Youker, B. W. (2005). Goal-free evaluation of the 2005 Kalamazoo Public School Middle School Summer Enrichment Program: Final report. Kalamazoo, MI: Western Michigan University, The Evaluation Center.

Youker, B. W. (2011). An analog experiment comparing goal-free evaluation and goal achievement evaluation utility (Unpublished doctoral dissertation).Western Michigan University, Kalamazoo.

Youker, B. W. (2013). Goal-free evaluation: A potential model for the evaluation of social work programs. *Social Work Research, 37*(4), 432-438.

Youker, B. W. & Ingraham, A. (2013). Goal-free evaluation: An orientation for foundations' evaluations. *The Foundation Review, 5*(4), 53-63.

Youker, B. W., Ingraham, A., & Bayer, N. (2014). An assessment of goal-free evaluation: Case studies of four goal-free evaluations. *Evaluation and Program Planning*, *46*, 10-16.