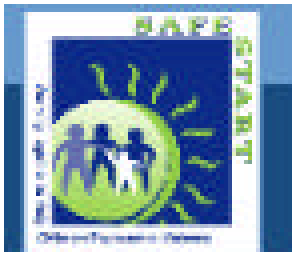
Suggested Methods for Evaluating Safe Start Training Outcomes

September 2, 2003



Safe Start National Evaluation Team: the Association for the Study and Development of Community, Caliber Associates, and the Office of Juvenile Justice and Delinquency Prevention.

PREFACE

This report on the evaluation of training outcomes was developed by the Association for the Study and Development of Community (ASDC) for the Office of Juvenile Justice and Delinquency Prevention (OJJDP) for the Safe Start Initiative under a contract with Caliber Associates. ASDC staff contributing to this report include: Inga James (Associate), David Chavis (Co-Project Director), DJ Ervin (Senior Associate), Kojo X. Johnson (Associate), Larry Contratti (Research Assistant), Kien Lee (Senior Associate), and Louisa Conroy (Project Assistant). ASDC would like to thank Jill Hunter-Williams for her contribution to this report.

TABLE OF CONTENTS

1. Purpose of Paper

According to the National Civic League's (2003) most recent Safe Start Assessment Plan, all Safe Start sites are providing training, or intend to provide training, to a variety of Safe Start stakeholders. In order to ensure that the training is meeting the needs of the recipients and the sites, as well as furthering the goals of the Safe Start program, this paper presents specific suggestions for training evaluation.

Generally, training modules of Safe Start sites focus on the impact of children's exposure to violence (CEV) (including both witnessing violence and being a victim), interventions for CEV, specific issues of child development relevant to CEV, and issues of referral to assistance for CEV (National Civic League, 2003). The training often targets outcomes that pertain to the enhancement of community systems in the identification and treatment of children exposed to violence. Specifically, outcomes pertaining to community systems include a more effective response to children who have been exposed to violence, such as:

- Timely identification of children exposed to violence;
- Appropriate referrals to assisting agencies;
- Increased access to services;
- Cultural- and gender-appropriate practice;
- Improved case management; and
- Efficient service delivery.

2. Training Framework

This paper proposes several approaches to evaluate training based on the four training levels developed by Kirkpatrick (1959). Table 1 details Kirkpatrick's levels:

Table 1. Kirkpatrick's Four Levels for Evaluating Training.

Name	Description
Reaction Level	Measures the level of participant satisfaction with training. For example, Safe Start trainers may survey police officers who attend a session outlining the impact of exposure to violence on children, asking how well the officers believe the training meets their needs, how useful the training will be in their work, and how well the training was organized and delivered.
Learning Level*	Measures the degree to which participants absorbed the material presented in training. This approach generally takes the form of pretests and posttests that assess the level of knowledge obtained by participants (e.g., police officers' knowledge about the impact of exposure to violence on children).
Behavior Level†	Measures the transfer of training concepts to applicable real-world situations. This type of evaluation would measure the degree to which participants engage in the behaviors targeted in training. For example, if a Safe Start site provides officers with CEV training, one would expect that officers would become more sensitive and responsive to children who have been exposed to violence.
Results Level	Measures the actual cost-effectiveness of the training and the return on investment produced by the training for the organization or training provider. For example, this evaluation would pose questions such as, are Safe Start sites that receive regular training experiencing less turnover, receiving more stable funding, or encountering cost-reducing phenomenon?

* The learning level of training evaluation is considered an internal validity concept in that it reflects how well the participant can perform certain tasks within the specific training venue.

† The behavior level of training evaluation is an external validity concept given it requires that the participants be able to transfer their new skills, knowledge, or attitudes to a different situation.

According to Alliger and Janak (1989), these four levels of training are related to one another to varying degrees. First, there is little or no relationship between reaction criterion and other levels (e.g., satisfaction with a training program has very little relationship to whether a subject is learned, etc.). However, there is a small positive relationship between learning and behavior criteria (e.g., there is some evidence that when a topic can be easily recalled on a test, it is also transferred appropriately to the real world). Furthermore there is a small positive relationship between behavior and results (e.g., using the applicable training concepts in the appropriate situation can result in a positive cost-benefit ratio), as well as a moderate positive relationship between learning and results (e.g., "book learning" is positively related to the cost effectiveness of the training program).

The findings described in Alliger and Janak (1989) highlight the importance of the learning and behavior levels when evaluating training. Specifically, because participant perceptions of the training are not associated with participant knowledge, participant workplace behavior, or organizational results, the information provided at this level is limited. Similarly, the Alliger and Janak data show that the learning criteria can provide an adequate indication of the results criterion, minimizing the need to focus on the results level. Furthermore, although a cost benefit analysis should be included in any long-term evaluation plan, such an analysis necessitates a longer time frame than that within the purview of this paper. Given these reasons, this paper will focus on the three initial training levels (i.e., reaction, learning, and behavior). The following sections outline ways in which evaluators can apply Kirkpatrick's levels to the evaluation of Safe Start training.

3. Evaluation Techniques

3.1. Reaction Level

Because this level is concerned only with the participants' own assessments of the training session/program, a single-posttest self-report survey is an appropriate data collection method. This approach is what Campbell and Stanley (1963) refer to as the *One-Shot Case Study*. Strengths of this approach include the wide-spread applicability and convenience of the design. Moreover, surveys can be administered soon after the conclusion of the training, making evaluation easier and yielding higher response rates. A primary weakness of the reaction approach is its lack of association to knowledge gained, behavioral work product, or organizational benefit. Furthermore, the lack of randomization, comparison group, and pretest allow for threats to internal and external validity (Campbell & Stanley, 1963).

Specific areas to be assessed at this level may be:

- Participants' perceptions of the usefulness of training;
- Participants' perceptions of the quality of training;
- Participant satisfaction with training; and
- Other participant perceptions of interest.

This design is appropriate for all training sessions regardless of the subject matter, participant profile, training format or other considerations.[‡]

3.2. Learning Level

The most common method of evaluating the degree to which training participants understand and absorb training concepts is a single-subject pre/posttest design. Campbell and Stanley (1963) refer to this approach as the *One-Group Pretest-Posttest Design*. The strength of this design is that it obtains a baseline measure of the criterion variable, which addresses internal validity issues related to selection and mortality (Campbell & Stanley, 1963). This is a rather weak design, however, due its inability to eliminate many threats to internal validity. For example, one cannot know if taking the pretest affected the posttest results (see Campbell & Stanley, 1963; or Cook & Campbell, 1979, for more).

To ameliorate the weaknesses of such a design, Eckert (2000) suggests a slight variation on the pre/posttest, single-subject design. He advises that instructors design a 30-question paper-and-pencil test of the training material. The questions are then randomly divided into two 15-question tests (e.g., Test A and Test B). Then, members of each training session are randomly placed into two groups, and Group 1 receives Test A as its pretest, while the Group 2 receives Test B. At Time 2 (posttest), the groups are given the opposite test (e.g., Group 1 receives Test B and Group 2 receives Test A). Of course, as with any evaluation design, this method works best with large groups of trainees. An alternate design, which would work for when training sessions are comprised of small groups, might be employing this method using two different training sessions. For example, if a site is providing a training module to two different groups of police officers, Test A could be used as the pre-test for one group and as a post-test for the other group (likewise for Test B).

Specific areas of knowledge to be assessed at this level may be:

- Indicators of violence exposure;
- Community resources to serve children who are victims/witnesses of violence;
- Ways to access services;
- Cultural dimensions to violence;
- Gender differences in views of violence; and
- Any other training objectives that may be added to this design as applicable.[§]

3.3. Behavior Level

This criterion requires that we examine whether training recipients, including both referring agencies (e.g., law enforcement agencies, child protective agencies, day care workers, etc.) and Safe Start sites themselves, are able to utilize their training in their work with children exposed to violence. The majority of Safe Start training goals fall into this category. For

[‡] The Sitka, Alaska, site has a relevant reaction survey on the Safe Start website (http://capacitybuilding.net/safestart.html) which may be of use to other sites.
[§] Baltimore, Maryland has learning level test forms also on the Safe Start Evaluation website for reference.

example, if sites are able to increase access to services and provide cultural- and gender-appropriate practice after training, they would be demonstrating behavioral. Generally, there are two distinct methods to gathering information about the degree to which transfer of training occurred: Single-Group Design and Two-Group Design.

3.3.1. Single-Group Design

The first method involves using measures of workplace behaviors to assess the extent to which participants were able to successfully apply the training information to their work with children exposed to violence (Garavaglia, 1993). This approach is amenable to either the One-Shot Case Study or the One-Group Pretest-Posttest designs mentioned above (Campbell & Stanley, 1963). Although these designs have serious limitations (see previous section), they can provide information about the obstacles encountered in using training materials on the job that may not be gathered any other way.

The dependent measures may take multiple forms (such as examining case documents, conducting consumer interviews, measuring consumer attrition, or evaluating staff job performances). For example, to use staff job performance to measure transfer of training within a pretest-posttest design, evaluators could develop action or implementation plans in conjunction with trainees' supervisors (Garavaglia, 1993). Prior to completing the training, participants could develop action plans that outline the ways in which they intend to implement the skills learned in their training. The plans could include a timeline and specific steps for achieving goals and could be completed as part of the training module. A supervisor could then monitor the trainee's plan for compliance and difficulties. At a point in the future, then, the evaluator could interview the supervisor as to the trainee's success in fulfilling the plan. This method has the additional benefit of providing program supervisors with a supervision tool that is directly related to program outcomes and the training created to ensure those outcomes.

3.3.2. Two Group Designs

The second method involves methodologically stronger designs involve assessing carryover of training to the workplace by comparing the performance of the referring agencies receiving training with a comparison group (e.g., a group not receiving training or another training group). Researchers have long recognized the Solomon Four Group Design as the most appropriate way to evaluate organizational training (Campbell & Stanley, 1963). Unfortunately, most Safe Start sites cannot randomly assign which groups receive training and which do not. Given these limitations, three potential designs are (Campbell & Stanley, 1963; Cook & Campbell, 1979):

- Separate-Sample Pretest-Posttest Sample;
- Untreated Control Group Design with Pretest and Posttest; and
- Posttest Only Design with Nonequivalent Groups.

Each of these approaches uses nonequivalent groups (i.e., the decision regarding which sample received training and which did not was not decided randomly). These models are presented in order of strength of design, with the designs presented first being the strongest and those presented last being the weakest, and least desirable. Again, the dependent measures could include document review, consumer interviews, or consumer attrition.

Separate-Sample Pretest-Posttest Samples. This design relies on the use of two training groups to be compared. Instead of using the training recipients in the measurement of training outcomes, Safe Start program participants can be surveyed. For example, prior to implementing a training module, two groups of program consumers (i.e., families receiving services) can be randomly selected. Group 1 is administered a satisfaction survey prior to the training session, and Group 2 is given the same survey after the training. The results of the surveys can then be compared and differences analyzed.

Untreated Control Group Design with Pretest and Posttest. A second design for measuring behavioral change is to adopt a two-group design to evaluate performance before and after training. Within the context of Safe Start, one way to examine the impacts of training is to compare referrals to Safe Start by training recipients with those not receiving training before and after training. As mentioned previously, two common training goals of Safe Start sites are (a) timely identification of children exposed to violence and (b) appropriate referrals to assisting agencies. Given these goals, document reviews should find briefer time lapses between violence exposure and referral to a helping agency and fewer inappropriate referrals (i.e., those that do not fit the established profile of CEV, as defined by the Safe Start team) for those receiving training than for those not receiving training.

Posttest Only Design with Nonequivalent Groups. The weakest design involves administering posttests to two groups – one that received training and one that did not. Because this design lacks randomization and pretests, reasonable causal inferences about any differences between groups is not possible (Cook & Campbell, 1979).

In sum, there are a variety of designs available from which evaluators can choose to assess the Safe Start training at the respective sites, taking into consideration the needs of the site. The strongest designs involve comparison groups as well as pretests and posttests. Research indicates that the utility of the information gathered at the learning and behavior levels is greatest.

4. Challenges and Recommendations

There are four major challenges to evaluating training programs within the Safe Start project.

- Program directors and evaluators have limited control over who receives training. Furthermore, it is generally not ethical or viable to withhold needed training in order to provide a control group. As a result, evaluators must rely on quasi-experimental designs, which can make it more difficult to draw inferences from the data.
- Sites have restricted resources to assess training, which can make stronger designs impractical.
- A third challenge involves measuring short-term and long-term learning. For example, if a training program encompasses several modules of training, all with the goal of increasing cultural sensitivity, how does one adequately measure the impact of each training module, apart from the others, particularly on the behavior level?
- Finally, it can be difficult to measure and track training effects.

Developing and implementing a strong, valid training evaluation protocol is a challenging but necessary task. Evaluators can adopt three strategies to strengthen the interpretability of their training evaluations. For example, to enhance both internal and external validity of the training evaluation, it is imperative that training be assessed relative to its contributions to program outcomes (i.e., be centered on the behavioral and results levels). It is not sufficient to rely on conventional satisfaction surveys (i.e., the reaction level) and pretests and posttests focused only on assessing knowledge learned. As highlighted earlier, the reaction level has limited utility and interpretability, and the learning level, although loosely associated with organizational results, often depends on very weak designs.

Another way to increase validity is to operationalize clearly the outcomes of interest. This task requires identifying intermediate outcomes and ensuring that these outcomes not only contribute to the long-term impacts, but also that they are measurable and testable. In addition, evaluators should adopt a multi-method approach to evaluate training programs in order to minimize threats to validity. Using a variety of designs, such as those described in this paper will increase the validity of the results by capitalizing on the strengths and addressing the weaknesses of the different designs. A multi-method approach also can entail using different measures (e.g., examining consumer views as well as referral documents) and focusing on different training levels (e.g., learning and behavior levels). Although administering a survey during or immediately following a training session is much simpler than tracking records and cases or conducting a consumer survey (with the inherent response rate problems), a multi-pronged approach provides sounder data. The Alliger and Janak data highlight the utility of involving both the learning and behavior levels. Targeting the learning level allows one to infer some extension to organizational outcomes, and targeting the behavioral level provides information about specific changes identified within training goals. Thus, developing and implementing a comprehensive training evaluation program requires continuous monitoring and on-going data collection.

5. Summary

The purpose of this paper is to provide a comprehensive means for assessing the strengths and weaknesses of the Safe Start training program. Kirkpatrick's (1959) framework offers an excellent structure for determining whether the training contributes added value to the Safe Start program. The paper proposes a variety of research designs appropriate for Safe Start sites using this framework. The examples and ideas presented should be used to guide training evaluation, and, as noted previously, the actual evaluation plan must be tailored to meet the needs of each site and each training module. Safe Start sites face many challenges in evaluating training, including limited control over sampling and restricted time and resources. Evaluators can strengthen the validity and interpretability of their training research by focusing on the learning and behavioral outcomes of training, clearly operationalizing these outcomes over time, and adopting a multi-method approach.

References

Alliger, G.M. & Janak, E.A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 42,* 331-342.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin.

Eckert, W.A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation, 21,* 185-193.

Garavaglia, P.L. (1993). How to ensure transfer of training. *Training & Evaluation, 47,* 63-68.

Kirkpatrick, D.L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training Directors, 13,* 3-9, 21-26.

National Civic League. (2003, March 5). *Annual Training and Technical Assistance: Safe Start Assessment Plan.* Washington DC: Author.