

A Continuum of Evaluation Strategies
for the Ohio Child Welfare Training Program

Submitted to:

The Ohio Child Welfare Training Program and
The Ohio Department of Job and Family Services

By

Tim McCarragher, Ph.D., MSW
Researcher
University of Akron School of Social Work

Kyle Hoffman
Training Coordinator
Institute for Human Services

Judith S. Rycus, Ph.D., MSW
Program Director
Institute for Human Services

April 3, 2003

Table of Contents

I.	Introduction	3
II.	Previous OCWTP Evaluation Activities	3
	Evaluations of OCWTP Workshops	
	OCWTP Feedback Studies	
	Statewide Training Assessment	
III.	Definitions and Research Methodologies	7
IV.	Training Evaluation Research	10
	A Framework for Evaluating Training	
V.	Potential Evaluation Strategies for the OCWTP	15
	Formative and Summative Evaluations	
	Methods of Data Collection for Formative Evaluation	
	Summative Evaluation Strategies for Measuring Outcomes	
	Methods of Data Collection for Summative Evaluation	
	Advanced Summative Research Design	
VI.	Considerations and Recommendations for the OCWTP	40
	Internal Validity	
	Calculating the Costs of Evaluation Activities	
	Establishing Expected Outcomes for Evaluation	
	Constructing a Chain of Evidence	
	References	48

I. Introduction

The mission of the Ohio Child Welfare Training Program (OCWTP) is to promote the delivery of high quality, family-centered child welfare services to abused and neglected children and their families by assuring that staff in the child welfare system are properly trained. With this mission in mind, the goal of the OCWTP's Evaluation Work Team is to establish a continuous, comprehensive, and coordinated system for the evaluation of the impacts and outcomes of OCWTP training activities.

The 2001-05 Request for Proposal, issued by the Ohio Department of Job and Family Services (ODJFS) called for the development of strategies to evaluate knowledge acquisition, knowledge comprehension, skill demonstration, and skill transfer that occur as a result of OCWTP training. This document presents and discusses a continuum of potential evaluation strategies for the OCWTP's consideration. It describes a variety of potential evaluation research design methodologies and reviews their strengths, their limitations, and barriers to their implementation, to help the OCWTP develop a comprehensive plan to evaluate training provided through the program.

II. Previous OCWTP Evaluation Activities

In recent years, the OCWTP has conducted many feedback studies and other evaluation activities to gather data on the training implementation process, program and training outputs, and the quality and effectiveness of selected OCWTP training curricula and initiatives.

A review of prior evaluation activities is important for several reasons. Prior to evaluating outcomes, an organization must first document the degree to which a program is actually operating in accordance with its program plan. The goal is to compile a body of information about the effectiveness of program operations, and to use this data to identify failures in program implementation which might impact training outcomes. Moreover, the OCWTP has considerable experience using a variety of evaluation methods and strategies, and should utilize and build upon this experience in future evaluation activities. Finally, the OCWTP

has accumulated an extensive body of previously-collected information, which can help in formulating questions for future research, and which can contribute to a chain of evidence that enables OCWTP to more accurately interpret the findings of future evaluations.

A brief summary of prior evaluation activities is presented here to contextualize the continuum for future evaluation activities.

Participant Response Evaluations of OCWTP Workshops

Since 1986, the OCWTP has conducted formative (i.e. process) evaluations of all OCWTP workshops. The workshop evaluation form, which is completed by trainees at the conclusion of every workshop, has been an essential source of data for ongoing OCWTP process evaluation and quality control. Aggregate data from these evaluation forms allows the OCWTP to review and monitor each workshop, each trainer's performance, and the relevance of the training for the group, as well as to identify problems in curriculum content and training delivery. The immediate availability of this data after completion of a workshop prompts OCWTP training managers to provide needed technical assistance to trainers, to modify curricula for content, relevance, or skill level, or to terminate a trainer from the pool of approved trainers.

OCWTP Feedback Studies

OCWTP has conducted a large number of feedback studies about a variety of its curricula and training programs. These have been reviewed in more detail in a report completed in the fall of 2002. The most relevant of these studies are briefly summarized here.

Skill Building Certificate Training Planning Study: Caseworkers and Supervisors Training Programs

This feedback study was conducted to inform the design of the Skill Building Certificate Training (SBCT) programs for caseworkers and supervisors. The SBCT programs were developed to increase OCWTP's skill-building capacity. The study included written surveys completed by caseworkers and line supervisors, and focus groups with caseworker and supervisor key informants from all eight OCWTP regions. The data was used to help OCWTP develop and implement two SBCT programs, and to complete the design work for two additional programs.

SBCT Caseworker/Supervisor Pilot Feedback Study

This study provided feedback on the pilots of the first two SBCT programs. This is the only OCWTP study to date that incorporated an outcome-based research design with experimental and comparison groups. The study included written surveys of caseworkers and supervisors, pre- and post-testing, focus groups, and key informant interviews with team members and trainers.

Building Skills in Family Risk Assessment Feedback Study

The focus of this evaluation was to determine the overall effectiveness of the *Building Skills in Family Risk Assessment* workshop. This evaluation used written surveys with workshop participants to determine their satisfaction with the curriculum, and self-assessments of increased skill level in conducting family risk assessments as a result of having attended the training.

Integration of Risk Assessment Into Caseworker Core Study

This study was undertaken to determine the degree to which caseworkers understood concepts of risk assessment taught in Core workshops. Data collection methods included written surveys of caseworkers, supervisors, and Core trainers. Follow-up telephone interviews were also conducted with Core trainers.

Sexual Abuse Intervention Workshop Series Feedback Study

This study sought to collect data from participants in the seven workshops that comprise the *Sexual Abuse Intervention Series*. The primary data collection instruments were written surveys, which were administered to participants, RTC staff members, and trainers. The standardized OCWTP workshop evaluation forms were also reviewed. The survey examined the participants' level of satisfaction with the training and perceptions of the influence of the training on practice.

Feedback Study on the Caseworker Guide to Core

This feedback study on the *Caseworker Guide to Core* summarized four separate smaller studies that had been conducted during the time the Guide was being designed and implemented. The studies included written surveys completed by participants at the conclusion of workshops using the *Core Guide*, and data generated by focus groups which were conducted at each workshop that used the *Core Guide*. The purpose of the surveys and focus groups was to better understand the usefulness of the Guide and barriers to its implementation.

GOALS-RATE Enhanced Final Report

The *GOALS-RATE Enhanced Study* was undertaken to review the operations of the OCWTP, and was a replication of an earlier *GOALS-RATE* study completed in 1992. The enhanced study focused on a number of areas, including participant demographics, participant satisfaction, perceived learning, perceived cultural knowledge and utilization, transfer of learning, environmental support and barriers to training, training program improvements, utilization of the Individual Training Needs Assessment, and trainer assessment of program effectiveness. A case study approach was also used to provide an in-depth description of how OCWTP training activities had generated or enhanced programmatic changes in recipient agencies.

Statewide Training Assessment

The *Statewide Training Assessment* was undertaken to provide a multifaceted overview of the current trends in child welfare, and the training needs of professional staff working in public children services agencies. Data collection strategies included extensive literature and administrative data reviews, focus groups, survey questionnaires, and telephone interviews.

Previous OCWTP evaluation activities have been used to guide the OCWTP in designing, modifying and improving its curricula and training programs. As OCWTP plans a future evaluation effort that assesses knowledge acquisition and comprehension, and the transfer of knowledge and skill from the classroom to the workplace, it should determine if and how these previous evaluations can be used to guide the design and implementation of future evaluation activities.

III. Definitions and Research Methodologies

Training evaluation is a form of field-based research. Its design and implementation are regulated by principles of social science research and program evaluation. To determine the most appropriate research methods to evaluate training requires a fundamental understanding of general research concepts and methods.

When designing any research project, two decisions must be made: what research design best suits the purpose of the research, and what data collection methods will be used to gather the information. Broadly, the research design addresses the questions of how the research will be structured, who will participate in the research, and when and where the research will be implemented. The data collection methods addresses what specific information will be gathered from respondents, and how the data will be collected. A third variable, sampling, determines how participants will be selected for inclusion in the research. Sampling falls under the category of research design, but is important enough to warrant independent discussion.

Research designs exist on a continuum from broad, open-ended, and uncontrolled at one end, to rigorous and tightly controlled at the other. In general, the greater the degree of rigor, the more likely the study conclusions will be reliable, valid, and generalizable to persons other than those who directly participated in the study. Increasing the level of rigor also increases the amount of work and the cost to complete the research.

Stages on this design continuum are sometimes called pre-experimental, quasi-experimental, and experimental. Pre-experimental research designs are considered the lowest level of design. Their purpose is to gather a large amount of previously unknown information with which to build general ideas and tentative theories. They do not produce statistically sound data or conclusive results. At the other end are highly controlled and rigorous experimental studies, often conducted in laboratory settings, that attempt to isolate and prove a direct causal relationship between two identified variables. Quasi-experimental studies represent an attempt to incorporate greater controls to achieve higher degrees of reliability and validity in the study, without incurring the high costs and work load of purely experimental research. Quasi-experimental studies are often the most rigorous designs possible in field-based research, where it is impossible to control all the variables operating in the real-world environment.

There are several types of commonly used research designs that address who is included in the research, and how often they are sampled. *Cross-sectional* research involves a single measurement of relevant variables in a single group. It is often considered a kind of “snapshot in time.” A cross-sectional design does not have an independent variable, and the group or population is surveyed only once. *Multiple group* research involves gathering the same information from multiple groups at the same time. The groups may be the same or different in composition, and in similar or different circumstances, depending upon the goals of the research. *Longitudinal* research involves multiple samplings of the same group or groups over an extended period of time. The goal is often to document changes over time, or to identify the long-term effects of certain variables. *Pre-/Post-* designs collect data from an identified group or groups of respondents either before or after an intervention, and often, both. Pre-/Post strategies attempt to determine the specific impact of a particular variable or intervention.

Data collection strategies can also be broadly grouped by type. *Survey* methods require participants to complete some form of written response – usually a questionnaire, a test, a survey form, or a standardized inventory/protocol. The Classroom Performance System (CPS) used by the OCWTP uses computer technology to implement survey data collection. *Interviews* require face-to-face contacts between the data collector and the respondent. While written protocols may be used to guide the interviewer, information is provided verbally by the respondent and is recorded by the interviewer. *Focus groups* involve face-to-face contact between data collectors (facilitators) and a group of respondents. *Observation* requires the data collector to watch a respondent performing an activity and rate their performance on a pre-determined set of criteria.

In more advanced levels of research design, a number of more sophisticated data collection methods are available, including the Classroom Performance System and embedded evaluations.

The Classroom Performance System (CPS) is a wireless response system that can provide immediate feedback in the classroom setting. It allows every trainee to respond instantly to questions. CPS hand sets operate using infra-red technology, connecting without wires to a receiver unit provided as part of the system. This new technology can incorporate surveys, formative training evaluation, and pre- and post-tests into the training itself, and can provide both on-site data collection and immediate feedback to participants. CPS is now available at all RTCs in Ohio. Pre-tests and post-tests can be scored in a matter of seconds, and workshop assessments can be quickly obtained to assure that learning is taking place.

Embedded evaluation is a process of incorporating evaluation strategies within the curriculum. Questions may be posed during training that require the trainees to demonstrate their satisfaction, opinion, knowledge level or skill acquisition. Embedded evaluations also provide immediate feedback to the trainees, which can enhance their learning experience. Embedded evaluations are easily incorporated into training using the CPS.

Sampling is the means by which respondents are selected to participate in a study. Key informant sampling involves selecting specific persons according to a predetermined criteria, such as position, prior experience, or knowledge. A sample may be constructed to represent a cross-section of a population, assuring that all geographic areas, types of organizations, and demographics of respondents are represented in the sample. Random sampling involves the statistical selection of respondents, often generated by a computer, with no prior knowledge of the characteristics of the selected respondents. In general, targeted sampling is used when very specific information is sought, but the results cannot be generalized to other populations. By contrast, purely random sampling offers the greatest probability that the findings can be validly generalized to the rest of the population from which the sample was selected.

Finally, researchers use a variety of statistical procedures to improve the rigor of a study. Statistical procedures can quantify the strengths of relationships between variables; test the impact of other, uncontrolled variables on the outcomes; and provide different levels of assurance that the outcomes are related to study variables, rather than a product of other variables or sheer chance.

IV. Training Evaluation Research

Training evaluation falls into the broader category of program evaluation research. Rossi and Freeman (1993) define program evaluation research as, "the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs." At various times, policy makers, funding organizations, planners, and administrators need a body of data to help differentiate effective and useful programs from ineffective and inefficient ones, and to enable them to plan, design, and implement new programs that better achieve their desired outcomes. This data is gathered and analyzed using highly standardized research designs and methods.

To be useful, OCWTP training evaluation activities should be designed within the framework and structure of program evaluation research, as described in the previous section. However, given the scope and complexity of the OCWTP and

its activities, this is a daunting undertaking. To assist the OCWTP in designing an evaluation program of the proper scope and scale to achieve its desired objectives, we present and describe herein a continuum of potential evaluation strategies that are particularly applicable to training evaluation. We also discuss their benefits and limitations, barriers to their implementation, and important considerations for the OCWTP.

A Framework for Evaluating Training

After reviewing the training evaluation literature, the OCWTP Evaluation Work Team identified the American Humane Association's (AHA) prototype for training evaluation as the most useful framework for evaluating OCWTP training activities. The AHA model adapted and expanded upon Donald Kirkpatrick's (1968, 1979) original four-level model of training evaluation. Kirkpatrick's model was the first comprehensive, published typology of training evaluation criteria and methodologies. His work described the most appropriate uses, benefits, and limitations of each type. His typology is widely quoted in the training literature, and has provided the foundation for most of the evaluation research that has occurred in the training field.

Kirkpatrick identified four levels at which training can be evaluated: Level I - Participant Response; Level II - Learning; Level III - Performance on the Job; and, Level IV - Results.

Level I - Participant Response

Level I evaluation samples participants regarding their feelings about and reactions to the training. This is generally accomplished using questionnaires distributed at the end of a training session. However, an unlimited number of questions can be asked, depending on the goals for the evaluation. While participant reaction data represents the opinions and judgments of the respondents, and is therefore likely to be more subjective, there is no reason to believe that participant views are not a sound source of data (Kohn and Parker, 1969). Kirkpatrick (1979) believes that participant response data can be very useful, particularly if the assessment instrument is properly formulated and tabulated, and accurate feedback is encouraged. He contends that trainee input

can be very useful in determining the quality of the conference leader or trainer, and that immediate feedback from trainees can lead to improvement of the training program itself. However, participant feedback is not recommended if the evaluation is seeking an objective measure of training outcomes.

Level II – Learning

Level II focuses on the measurement of knowledge and skills gained through training. *Knowledge* refers to the learning of new principles, facts, and concepts, while *skill* refers to the learning of new behaviors and techniques for accomplishing a task (Parry & Berdie, 1999). The initial learning of basic facts, concepts, and new skills is a necessary prerequisite to application of course content to the job. Measurement at Level II is more complex than at Level I. It requires the use of objective, measurable criteria, valid instruments, and statistical procedures to document with certainty that learning has occurred as a result of the training (Parry & Berdie, 1999).

Level III – On-the-Job Performance

Level III seeks to measure the degree to which new learning acquired during training actually transfers to the job. Level III measures the actual performance of new skills in the job environment. Measurement at Level III requires comparable methodological and statistical rigor to Level II, particularly if outcomes are to be attributed to the training event. Level III evaluation typically involves the use of such instruments as participant action plans, questionnaires, interviews with employers or supervisors, observation of job performance, and the review of administrative data (Parry & Berdie, 1999).

Level IV– Results

Kirkpatrick's Level IV seeks to determine the outcomes of a training program at the organizational level. This usually requires a complex experimental research design, with standardized measures and appropriate controls, that can clearly demonstrate a causal relationship between the training activity and organizational outcomes (Parry & Berdie, 1999). Most training evaluators contend that Level IV evaluation is extremely difficult, if not impossible, to

accomplish in complex organizations, where a myriad of intervening variables cloud the relationship between a particular training event and organizational outcomes. Level IV evaluation is more easily conducted when the desired outcome is more concrete and easier to identify and quantify, such as number of units manufactured, or profit.

The American Humane Association Levels of Evaluation

The American Humane Association (AHA) has recently published a model of training evaluation for the human services that adapts and expands upon Kirkpatrick's original model. The AHA model also explicitly builds a continuum of evaluation that incorporates both *formative evaluation*, which is designed to assess the effectiveness of training processes, materials, and delivery, and *summative evaluation*, which seeks to establish and verify the outcomes of training. It is important that any evaluation continuum include both types of evaluation. Information regarding the adequacy and effectiveness of the training's content, structure, implementation, and delivery are a necessary base upon which we interpret data about outcomes. Outcome data collected on a program that is incompletely or poorly designed and improperly or incompletely implemented can lead to erroneous conclusions regarding the training's effectiveness. Second, outcome-oriented measures, such as tests of knowledge and performance, can provide essential formative feedback on the adequacy of the curriculum, delivery methods, and trainer's performance (Parry & Berdie, 1999), which can help identify needed modifications to the program.

The ten levels in the American Humane Association's model are as follows:

1. *Course level* – includes the evaluation of the training event or program itself, and includes ratings of the content, structure, methods, materials, and delivery of the training.
2. *Satisfaction level* – refers to trainees' feelings about the trainer, the quality of the material presented, the methods of presentation, and the training environment (i.e., room set-up and temperature).

3. *Opinion level* – refers to the trainees’ attitudes toward the value of the training, perceptions of their own learning, perceptions of the relevance of the training, an opinion of how the new material fits within their pre-existing belief system, and their expected utilization of the training on their jobs.

4. *Knowledge acquisition level* – refers to the degree to which trainees have learned and can recall information, including facts, definitions, concepts, and principles. This is most often measured by paper-and-pencil, short answer or multiple-choice tests.

5. *Knowledge comprehension level* - refers to the degree to which trainees understand complex ideas, concepts, and the relationships between these, as well as their ability to recognize how concepts are used in practice. It also includes the ability to use knowledge to solve problems. This level is also frequently measured by a paper and pencil test, but the test items require a greater ability to integrate and apply the course material.

6. *Skill demonstration level* – refers to the actual application of new learning in the performance of a task, within the controlled environment of the training class or activity. It requires the trainee to apply learned material in new and concrete situations, but while still in the classroom. This is often done through simulations, exercises, and other experiential activities.

7. *Skill transfer level* – requires trainees to apply newly learned knowledge and skills in direct practice situations outside the classroom, most often in the job environment. Evaluation strategies at this level include Participant Action Plans, observation of trainee performance on the job, feedback from supervisors, and case record reviews.

8. *Agency impact*, 9. *Client outcomes*, 10. *Community impact levels* – these advanced levels address the impact of training at the agency, client, and community levels respectively. Evaluation at these levels might address the impact of a substance abuse training program on patterns of service utilization, or, the degree of interagency collaboration in case management and referral. A cost-benefit analysis might also be conducted at agency, client, or community levels.

V. Potential Evaluation Strategies for the OCWTP

In Deliverable 7 of the 2001-05 OCWTP Request for Proposal, several potential evaluation strategies were delineated to evaluate outcomes of OCWTP training. These included: 1) pre- and post-test assessments of knowledge; 2) assessment of participant skill development; and 3) assessment of the extent of transfer of learning from training to the job, including factors that promote or hinder learning on the job. These three objectives reflect four of AHA's ten levels of evaluation; knowledge acquisition, knowledge comprehension, skill demonstration, and skill transfer. Conducting evaluation research at all four levels will require the OCWTP to develop different evaluation criteria, research methodologies, data collection and analysis procedures, and differential use of statistics for each of the four identified evaluation levels.

The research strategies included in this section are divided into two major categories: methods of data collection, and methods of research study design. Several potential methodologies will be presented here, with a description of each methodology, its strengths and limitations, barriers to its use, and its practical applications in addressing various levels of evaluation. Data collection methodologies include surveys, interviews, focus groups, observations, embedded evaluation, use of technologies such as the Classroom Performance System (CPS), and administrative reviews. Research design methodologies include cross-sectional, longitudinal, and pre-test/post-test designs.

Formative and Summative Evaluations

There are two general types of evaluations: formative and summative. Formative evaluation is a critical component of an effective training program. It is an ongoing strategy to monitor the training program to see if it is operating as effectively and efficiently as possible (Unrau, Gabor & Grinnell, 2001) and to implement continuous quality improvement. While formative evaluations often concurrently gather information regarding training outcomes, their ultimate purpose is to improve the quality, relevance, and delivery of the training activity or program. Summative evaluations gather relevant data to determine whether the training program has resulted in certain predetermined outcomes. For

example, a summative evaluation of training effectiveness might determine if a skill taught in training is reflected in the daily work of the trainees.

Formative evaluation is most appropriate for evaluating training at Kirkpatrick's Level I, which uses participant feedback to measure the quality of the training event itself and the participants' perceptions of the relevance of the training to their work. Kirkpatrick's Level I corresponds to the AHA model's Level 1 - Course, Level 2-Satisfaction, and Level 3-Opinion.

Methods of Data Collection for Formative Evaluation

Several data collection methods are frequently used in formative evaluations. They include surveys, interviews, and focus groups. Each of these methods will be examined individually, including their strengths, limitations and barriers.

A. Surveys

Surveys are valuable tools for formative evaluation research. Survey instruments can include written questionnaires or survey forms, tests, or standardized inventories and protocols. The main goal of surveys is to gather information and/or opinions from a large enough sample of people to validly describe them as a group. Although surveys require more planning than do focus groups or personal interviews, they also provide data that is more objective and scientific. Because a much larger sample can be used with survey research, the findings from the sampled group can be more accurately generalized to the total population represented by the sample. For example, when trainees are sampled in a precise manner, such as randomly surveying every fifth trainee, there is a high degree of confidence about the validity and reliability of the findings for all trainees (Royce & Thyer, 1996).

However, creating survey questions that yield valid and reliable results can be difficult (Unrau, Gabor, & Grinnell, 2001). To generate valid and reliable results, survey questions must be clearly phrased, well-articulated, easily understandable to the respondent, and must actually measure what they are intended to measure. Many design and development problems can decrease the

validity of results. For example, directions to complete the survey may be too vague; questions may be confusing; and multiple choice questions may have more than one correct answer. While the benefits of well-developed surveys are clear, developing a reliable and valid survey instrument can be a time-consuming and expensive endeavor.

There are two basic types of surveys: non standardized and standardized. Non standardized surveys do not have previously established validity or reliability. In other words, the survey instrument has not been previously tested to determine a high degree of accuracy and consistency in the format of the questions and the potential responses. Non standardized surveys are used when standardized surveys are not available, or are not appropriate to answer the research question. Thus, while non standardized surveys can provide extensive data, we cannot be assured that the data is reliable or valid.

Standardized surveys are scientifically constructed to measure one specific variable (Williams, Unrau, & Grinnell, 1998). An example of a standardized survey would be the Beck Depression Inventory, which has been tested hundreds of times with various populations, and has yielded consistent and accurate results in determining a client's level of depression.

Strengths of Survey Research

The strengths of survey research include:

- There is a high potential for objectivity at a lower cost than other forms of data collection.
- Unlike interviews and focus groups, evaluation participants can read and respond to survey questions at their own pace.
- Seeing the questions in a written form is often an advantage, in that it facilitates comprehension by respondents.
- The anonymity of a written survey may promote more honesty than would be comfortable for respondents in an interpersonal situation, such as an interview or focus group.

- Surveys are less time consuming and relatively easy to administer when compared to other forms of data collection. Because they are easy to implement, they can gather important data from a large number of respondents while it is still fresh in their minds, rather than waiting to schedule interviews or focus groups at some future date.
- If the survey instruments are pre-tested and the study sample is constructed using random sampling methods, the validity and reliability of the results can be very high.

Liabilities of Survey Research

As with all data collection methods, there are also liabilities in using survey research, including the following:

- Standardized measures are difficult to construct and must be pre-tested for reliability and validity.
- Trainees may react negatively to being surveyed, particularly if questionnaires are sent to them in the mail, or they must complete a lot of survey questionnaires for different purposes.
- It is difficult to develop survey items that cover complex topics, that provide in-depth information, or are open-ended enough to allow for a discussion of broader perspectives (Unrau, Gabor, & Grinnell, 2001).
- Mailed surveys often have a low response rate, which requires that many more surveys be sent out than are needed to assure a large enough sample.
- There may be critical differences between those respondents who send back the completed surveys and those who do not, which might skew the results.
- If the sample is chosen for convenience, such as surveying only one easily accessible group of trainees, or only the most available supervisors, the results cannot be generalized to all trainees or all supervisors.

- Because there is no interaction between the evaluator and the trainee, there is no opportunity to clarify confusing items.

Barriers to Survey Research

Barriers to survey research include:

- It is time consuming and costly to develop valid and reliable measures, particularly when attempting to measure multiple variables concurrently.
- There is considerable expense associated with data entry and analysis.
- Because surveys typically sample the respondents' perceptions, beliefs, and opinions, they are not suitable for obtaining objective measures of outcomes.

B. Interviews

Interviewing involves participating in a face-to-face or telephone conversation with selected respondents. One widely-used interview method, often referred to as "key informant interviews," selects specific people who are considered knowledgeable about the chosen research questions.

Interviews can be formal and use a highly structured interview schedule or protocol; or, they can be informal, with a less regimented structure and more open-ended questions. Structured interview schedules are used when there is some prior knowledge of the topic being investigated and a particular kind of information is needed. To obtain such detailed data, questions are included that probe for specific answers with more depth. When very little is known about an issue, an informal, unstructured interview permits more free-flowing discussion and generates information not previously considered. Generally, face-to-face interviews are more often used to ask questions that permit open-ended responses (Unrau, Gabor, & Grinnell, 2001).

Strengths of Interview Research

The strengths of interview research include the following:

- Interviews tend to have a higher response rate than surveys.
- Interviews allow for longer, more open-ended responses, subjects tend to provide more complete and thoughtful answers, and interviewers have an opportunity to record nonverbal information. This tends to yield more in-depth data when compared to survey research.
- Interviewers can clarify questions for respondents, and subjects are often more willing to answer sensitive questions in person than on a written form (Unrau, Gabor, & Grinnell, 2001).

Liabilities of Interview Research

There are also liabilities of interview research.

- During the interview process, the interviewee may distort information through recall error, selective perception, or an unwillingness to disclose certain information in a face-to-face encounter.
- There is a higher chance of introducing researcher bias into the findings, particularly if the interviewer is not regulated by standardized interview questions. The interviewer's tendency to use leading questions or to direct discussion into areas of personal interest can skew the results.
- The interviewee may respond to the personality or style of the interviewer rather than to the objective content of the interview (Unrau, Gabor, & Grinnell, 2001).
- Because of the smaller samples involved in interviews, the findings are less likely to be generalizable.
- It would be difficult to conduct an interview that went beyond the opinion level. While knowledge questions could be asked during an interview, the one-to-one setting may make the trainee feel awkward if they do not have a correct response.

Barriers to Interview Research

Barriers to interview research include the following:

- Interviews are expensive because of the time needed to meet individually with participants, and to analyze the large amounts of qualitative data typically collected during interviews.
- Interviews require well-qualified, highly trained interviewers.
- Structured interview protocols require pre-testing and training of the interviewers in both asking survey questions and recording findings, to assure inter-rater reliability.
- Interviews are not a practical option during training, so they must be scheduled at a time in the future.

C. Focus Groups

Focus groups provide an accepting and non-threatening environment in which a selected group of respondents discuss their opinions and perceptions about a predefined topic of interest. Focus groups are a special type of group in terms of their purpose, size, composition, and the procedures used to lead the session. A focus group is typically comprised of 7 to 10 participants selected because they have certain characteristics in common that relate to the topic of the focus group. Typically, a focus group study will consist of a minimum of three focus groups, but it could involve as many as several dozen groups (Krueger, 1994). Focus groups are more complex than individual interviews because they involve interactions between and among respondents (Krueger, 1994).

The purpose of focus groups is to gather data to explore or test ideas. Focus groups should consist of individuals who are reasonably familiar with the topic of discussion, but not necessarily familiar with each other.

The main task of a focus group facilitator is to balance group discussion so it remains targeted on the questions being asked, but also stimulates members to

discuss more in-depth and comprehensive data. The results of a focus group may show both similar and divergent perceptions of participants (Unrau, Gabor, & Grinnell, 2001).

Strengths of Focus Group Research

A number of the liabilities of interview and survey studies are addressed by using a focus group methodology, including:

- The larger study sample improves the generalizability of the findings.
- The interaction between the focus group participants often prompts more in-depth discussion and responses, clarifies thoughts and ideas, and generates discussion that would be missed if respondents were sampled individually.
- Focus group studies can be less expensive than interview studies because the evaluator can collect data from a larger number of people at the same time.
- The focus group format allows the facilitator to ask open-ended questions to expand discussion, and also to probe for more detailed information when necessary, which is not possible in survey research (Krueger, 1994).

Liabilities of Focus Group Research

Liabilities of focus group research include the following:

- Because focus group participants respond to each other's comments, there may be the potential for "group thinking," where participants begin to develop a common opinion.
- While focus groups have greater generalizability than individual interviews, they still have less generalizability than survey research.

- Results are also significantly more cumbersome to analyze than survey research (Krueger, 1994) because of the sheer scope and depth of the data collected.

Barriers to Formative Focus Groups Research

The barriers to focus group research are similar to those of individual interview studies, including:

- Focus groups require highly skilled facilitation. If multiple focus groups are used with different facilitators, pre-group preparation of facilitators will be necessary to assure standardization between groups.
- The facilitator has less control over the content or flow of the discussion than in an individual interview.
- It is often difficult to schedule a large group of people at the same time and place.
- The focus group format necessitates carefully planned questions to generate the information the evaluators are most interested in.
- Two facilitators may be needed for each group, as it is difficult to take notes while facilitating the group.
- Data analysis is time consuming, since it usually requires identifying and articulating prominent themes from extensive qualitative data.

Research Designs for Formative Research

A. Cross-Sectional Research Designs

As indicated above, cross-sectional research is a single measurement of a variable or variables. Cross-sectional studies examine a phenomenon by examining a cross-section of it at one point in time – a kind of "snapshot" in time. A cross-

sectional survey design does not have an independent variable, and a group or population is surveyed only once.

Strengths of Formative Cross-Sectional Research

- It can provide insight into a topic that has not been exhaustively researched without providing an intervention or a pre-test and post-test. Often, a more rigorous methodology can be used as a follow up once the cross-sectional study's results are analyzed.
- It is useful in situations where the event or activity only needs to be measured once, such as when soliciting opinions or satisfaction.
- There are administrative advantages to cross-sectional research. Because a survey or questionnaire is only administered once, it is less time-consuming than pre-tests and post-tests. It can also measure individual as well as group performance.

Limitations of Formative Cross-Sectional Research

- There can be little confidence that the results are generalizable to the population being studied because the sample is not randomly selected.
- There are no other measures for comparison.
- It is impossible to determine causality or the direction of a relationship between variables, because the variables are only measured once. For example, when examining variables such as education, training, and work experience, it is not possible to determine if training causes a superior work performance, other factors have contributed to superior work performance, or superior workers do better in training.
- While cross-sectional research is relatively simple to administer, it can only measure attitudes, opinions, and knowledge.

- Cross-sectional research in a training context is dependent upon the trainer's ability to administer the survey properly.

Barriers to Cross-Sectional Research

There are two major barriers to cross-sectional research:

- There is often significant cost involved in developing the questions. The questions in a survey or questionnaire must be worded appropriately, and if standardized measures are included, they must be adequately pre-tested to assure validity and reliability.
- There must be a specific plan for data analysis. Some cross-sectional research may only require frequencies and percentages (i.e., 40 trainees, or 55%, indicated that the training would be useful in their work.) Other data analysis procedures would require data entry and more sophisticated data analysis procedures.

B. Post-test Only Research Designs

One Group Post-test Only

The one-group post-test-only design is also known as a "one-shot case study" design. This design provides a single measure of what happens when one group is subjected to one treatment or experience, such as training. The participants are not randomly selected, and the findings are not easily generalizable. An example of a one group post-test-only design would be measuring the degree of trainee satisfaction with the content of the training, or testing the level of trainee knowledge at the end of the training.

Because the strengths, liabilities and barriers of the one-group post-test-only design are similar across the range of pre-test/post-test research designs, they will be discussed under summative evaluation strategies below.

C. Adapted Designs for More Rigorous Formative Evaluations

The evaluation strategies described thus far can adequately measure at Kirkpatrick's Level I, and AHA's corresponding Level 1-Course, Level 2-Satisfaction, and Level 3-Opinion. However, there are adaptations available to strengthen these research designs.

Include Additional Variables in Cross-Sectional Research

In order to rule out other possible explanations of the findings, evaluators may add a number of variables to the study to determine the degree of influence of different variables on one another. For example, when completing a cross-sectional study of knowledge after a training, an evaluator may gather information on trainees' previous attendance at training, their educational level, and their prior years of work experience in child welfare and in social services. These variables can then be included when analyzing knowledge levels to determine if the measured level of knowledge can likely be attributed to the training or may instead be due to these other factors.

Randomized Cross-Sectional Research

Randomized cross-sectional research adds the strength of a randomized sample to the study design. While the strengths and limitations are the same as cross-sectional research, this design requires the sample to be randomly selected from the population, such as surveying the trainees in every fifth workshop conducted using the same curriculum. If the sample is randomly selected, then the data can be generalized to the entire population from which the sample was drawn. This can prove to be more cost effective than sampling every trainee who completes the workshop.

Multi-Group Post-test-Only

A multi-group post-test-only design is the same as one-group post-test design, but more than one group is used. It cannot be assumed that all groups receiving the training are equivalent, nor can it be assumed that any of the groups are representative of the larger population. The participants are not randomly

selected, and the findings are not easily generalizable. An example of a multi-group post-test only design would be evaluating knowledge acquisition in a post-test of various Supervisory Core workshops in several of the OCWTP regions.

Randomized One-Group Post-test-Only

In a randomized one-group post-test-only design, the members of the group are randomly selected to take a post-test. Otherwise, the design is identical to the one-group post-test-only design. Because a random sample is used, however, it is possible to generalize the program's results to the population from which the sample was drawn. However, it is not possible to attribute the measure of the dependent variable (such as knowledge acquisition) to the intervention (training). An example of a randomized one-group post-test-only design would be a survey of opinion from a random selection of caseworkers that attend a standardized workshop.

Longitudinal Case Study

The longitudinal case study is similar to post-test only designs, but it includes more than one measurement of the dependent variable over time. For example, if an evaluator was interested in measuring how the degree of skill transfer changes over time, he or she might include three separate measures of skill transfer at different times after the completion of a standardized workshop. The advantage of a longitudinal case study is that the long-term effects of training can be measured. However, the liability is that many other unrelated variables potentially impact the trainee's skill level over time, thus diluting the strength of the association between skill level and the original training activity.

Summative Evaluation Strategies for Measuring Outcomes

The second major division in training evaluation is summative evaluation, which focuses on determining whether attendance at training results in certain quantifiable outcomes. Most summative training evaluations in the human services have focused on measuring the extent to which trainees master new

knowledge and skills acquired in training, and subsequently implement them on their jobs.

Kirkpatrick's Level II involves quantifying and measuring learning – i.e. the mastery of specific knowledge and skills as a result of training. *Knowledge* refers to the learning of new principles, facts, and concepts, while *skill* refers to the learning of new behaviors and techniques for accomplishing a task (Parry & Berdie, 1999). Kirkpatrick's Level II corresponds to three of the AHA Levels: Level 4 - Knowledge Acquisition, Level 5 - Knowledge Comprehension, and Level 6 - Skill Demonstration. Kirkpatrick's Level III, On-the-Job Performance, corresponds to AHA's Level 7, Skill Transfer. Both models contend that the mastery of basic facts, concepts, and skills are necessary prerequisites to the successful application of course content to the job.

Measurement at Kirkpatrick's Levels II and III is more complex than at Level I because the goal is to evaluate outcomes rather than processes. Summative evaluation requires the use of objective, standardized, measurable, and valid instruments and methodologies specifically designed for the intended purposes (Parry & Berdie, 1999).

Methods of Data Collection for Summative Evaluation

There are several potential methods of data collection in summative training evaluation research, including pre-tests/post-tests, surveys, observations, and the use of administrative data. Each of these methods will be examined, including their strengths, limitations and barriers to implementation.

Pre-tests / Post-tests

The goal of pre- and post-tests is to measure and document the degree of change between a baseline level of knowledge or skill before the training, and a post-training level of knowledge or skill. In theory, if there are no other opportunities to learn between the pre- and post- tests, we should conclude that the training was responsible for any demonstrated changes in knowledge and skill.

However, while this assumption may sound logical, it is not always accurate. Reasons will be discussed shortly.

There is an enormous range in the types of pre-test/post-test methodologies used to evaluate training. A pre-test/post-test design may contain up to four elements: a pre-test, an intervention, a post-test, and the strategy for selection of participants. Different combinations and timing of these four elements can produce dozens of possible research designs. The most widely used designs are presented and discussed below.

One-Group Pre-test/Post-test

In a one-group pre-test/post-test design, the pre-test is used as the basis for comparison with the post-test results. This design is generally used to determine how the independent variable (training) affects the desired outcome (learning) in a particular group. In this model, the goal is to compare trainees' post-training performance with their own pre-training performance and document the *degree of change*. The data can be compiled for each individual respondent or summarized for the group as a whole, depending on how the data is to be used. However, this design does *not* control for rival hypotheses – that is, many factors other than the training itself can affect both pre- and post-test results. These can include demographic factors such as age, educational level, and years of child welfare experience; test-taking comfort and ability; quality of the test construction; clarity of the test questions; the degree of relatedness of the test questions to what is actually taught in the training; the tendency of the trainer to "teach to the test;" and a variety of other factors.

Comparison Group Post-test-Only

The comparison group post-test-only design improves on the one-group and multi-group post-test-only designs by introducing a comparison group that does not receive the independent variable (the training), but is administered the same post-test as the training group. The scores are then compared between the two groups. The assumption is that higher scores found in the trained group would suggest that the training has had an impact. In more rigorous experimental designs, the comparison group is called a "control group." However, there are

differences. While a control group is always randomly assigned, a comparison group is not. An example of a comparison group post-test-only design would be comparing two groups of supervisors; one group would complete a standardized workshop, while the other would not. Both groups would complete a post-test of multiple-choice questions designed to test knowledge comprehension, and their scores would be compared.

Comparison Group Pre-test/Post-test

The comparison group pre-test/post-test design elaborates on the one-group pre-test/post-test design by adding a comparison group. The comparison group does not receive the independent variable (i.e., training), but takes both the pre-test and the post-test. The two groups are not necessarily equivalent, because their members are not randomly assigned. The pre-test scores of the two groups indicate the extent of their differences. An example of a comparison group pre-test/post-test design would be testing for knowledge acquisition in two groups of trainees. One group would take a pre-test, attend a workshop, and take a post-test. The second group would take the pre-test and post-test, without attending the training. One would expect to find greater differences between pre- and post-test scores in the group that attended the training.

Experimental Research

The final group of pre-test/post-test methodologies falls under the category of experimental research, which involves the rigorous control of an independent variable (such as training) in order to test a hypothesis about its effects on the outcomes. This is accomplished by establishing nearly identical comparison groups – one of which attends the training (the experimental group) and the other which does not (the control group). Experimental models are able to establish causality between the independent variable (such as training) and the dependent variable (a score on a knowledge comprehension test). There are three criteria for causality:

- The cause precedes the effect in time.

- The two variables are logically related to one another, such as training and skill demonstration.
- The logical relationship observed between two variables cannot be explained as the result of the influence of some third variable that causes both of the variables under consideration.

These requirements make experimental research relatively rare in the social sciences because of the expense and constraints of including a control group in the evaluation. There are also potential ethical issues surrounding actively withholding training from certain staff members to create the necessary control groups.

Examples of experimental research designs include pre-test/post-test/control group, and the Solomon four-group design, described below.

Pre-test/Post-test Control Group

The pre-test/post-test control group design builds upon other research designs, but includes the random assignment of respondents to either the experimental group or the control group. A randomly assigned experimental group receives the intervention (i.e., training), while the randomly assigned control group does not. This design addresses most threats to internal validity, since the groups are randomly assigned and, therefore, are considered equivalent on all relevant characteristics. In this design, the experimental group would take a pre-test of knowledge comprehension, complete the workshop, and take a post-test. The control group would take the same pre-test and post-test of knowledge comprehension but would not attend the workshop. The distinction in this design is that the trainees are randomly assigned to either the experimental or control group, so other variables such as age, prior work experience, demographics, etc. do not influence the outcomes.

Solomon Four-Group Design

The Solomon four-group design is considered the most elaborate design among the pre-test/post-test designs. The design includes four randomly assigned

groups; the first and third receive the intervention (i.e., training), while the second and fourth groups do not. The first and second groups take the pre-test, and all four groups take the post-test. This design controls for all threats to internal validity, and its findings can be confidently generalized to the population. However, the design is also extremely cumbersome, and is rarely used in social science research because of the expense and staffing challenges of creating four randomly assigned groups.

Quasi-experimental research designs (one-group pre-test/post-test, comparison group post-test-only, and comparison group pre-test/post-test) include a pre-test and/or a comparison group that makes the findings more generalizable, because a greater number of threats to internal validity can be addressed.

Experimental research (pre-test/post-test control group, and Solomon four-group) designs are the strongest type of designs. Because of the random assignment of research subjects, and control groups, the findings are the most generalizable.

Strengths of Summative Pre-test/Post-test Research

In general, the strengths of pre-test/post-test designs are:

- The program can gather information on the current level of knowledge of individual or groups of trainees, and the change in their level of knowledge over time.
- Tests are frequently used to determine whether a particular program or trainer has successfully transmitted information in specific content areas, or has influenced the thinking skills of participants. Pre-test/post-tests can measure satisfaction, opinion, knowledge acquisition, and knowledge comprehension.
- Pre-test/post-tests can also provide feedback to trainers on what is, and what is not being learned by participants in the training.
- The same survey can potentially be re-used over time, provided that the content of the training does not change.

- Pre-test/post-test data can help draw conclusions about both the trainer and about individual participants. Compiled group data can offer insight into how well the trainer and curriculum are conveying the desired knowledge. Individual participant scores could be used to design individual development plans.

Liabilities of Summative Pre-test/Post-test Research

There are also liabilities in pre-test/post-test research:

- These designs can only test satisfaction, opinion, and knowledge. It would be difficult to test skill without also including observations or administrative review.
- Because most of the testing would take place during the training, successful implementation would require trainers who are willing and able to administer the tests. Test administration would also take time away from training.
- If trainers are aware of the nature or content of the tests, trainers may train to the test rather than focusing on the designated curriculum.
- Without thorough pre-testing and validation of individual test questions, the instrument may not produce an accurate representation of the trainees' level of knowledge.
- Some individuals do not respond well to written tests. Test anxiety, reading difficulties, and other variables might negatively impact their test performance and scores. Further, while trainees would be told that their scores would not influence their employment status, testing may still raise their anxiety levels and impact both learning and test performance.

Barriers to Summative Pre-test/Post-test Research

Barriers to pre-test/post-test research include the following:

- Tests of knowledge acquisition and knowledge comprehension would require a floating test question bank, which is a database of validated questions designed to measure the content covered in the curriculum. As the curriculum changes, the questions would have to be modified, which can be time consuming and expensive. This would also require additional computer software to maintain a test bank of questions and to formulate different, but equivalent, versions of the tests to be administered to different groups to prevent trainees from "learning the test."
- There are time constraints involved with ongoing data analysis and report writing, and increased expense when using comparison and control groups, such as mailing questions to trainees in the comparison or control group, or scheduling time for them to complete the pre-test and post-test without receiving the training.

Pre-test and post-test research would be used most effectively to evaluate two of the AHA levels of training: Level 4 - knowledge acquisition, and Level 5, knowledge comprehension.

Level 4 is the most appropriate level for pre-test/post-test research. The most logical type of test would be a multiple-choice, short answer, or true and false test that focuses on key elements of the content covered in the curriculum, administered to groups that attend training in the curriculum. As mentioned earlier, the use of comparison or control groups would make the findings more generalizable, but would also make the evaluation more cumbersome and costly.

Knowledge comprehension could also be evaluated using pre-tests/post-tests. Rather than a multiple choice test, evaluators would develop more elaborate questions that required critical thinking, problem solving, and integrating and applying content covered in the training to practical work tasks. Examples might be essay questions, simulations, or the use of case studies, with questions that focus on determining the most appropriate action to be taken based on given information. This level of evaluation is more complex to develop, would require more time than a Level 4 pre-test/post-test, and would be more time consuming and expensive to score.

Observations in Summative Evaluations

For many training programs, the preferable source of data collection is direct observation. An essential part of any observation effort is a plan for systematically recording the observations made. Observers must be trained how to make observations and how to record them uniformly.

There are three common ways of making systematic observations. The first approach, known as the narrative method, involves the least amount of structure. The observer is simply asked to record events in as much detail as possible, and in the order in which they occur. A second, more structured approach is to provide observers with a data guide – a set of questions they are required to answer from their observations. A data guide can resemble a survey instrument, with blank spaces between the questions for recording purposes. Such a recording instrument simplifies analysis considerably, and increases the likelihood that the observers will provide consistent information. The third approach is to use some form of structured rating scheme. Ratings can be purely descriptive, such as a checklist specifying the proportion of time devoted to different kinds of activities (Rossi & Freeman, 1993).

Structured observations occur under controlled conditions to collect precise, valid, and reliable data about complex interactions. An impartial observer is trained to watch particular persons or events and to look for specific actions and behaviors. The observation can take place in a natural setting (such as in the field, with a supervisor observing an assessment) or in an artificial setting (such as in a training class), but the conditions and timing of these observations are always predetermined. The trained observers carefully record their perceptions of what they see; they are not directly involved with the people or the event being observed (Unrau, Gabor, & Grinnell, 2001). The key to successfully conducting an observation is a clear format for scoring the degree of skill demonstrated. A carefully constructed format allows for a more precise measurement.

In terms of training evaluation, observation would provide the OCWTP with an opportunity to evaluate the higher levels of learning, such as skill demonstration and skill transfer. Because these levels may be difficult to measure during the

training itself, it would be necessary to conduct observations of skill transfer after training has occurred. These observations can be made by trainers, job coaches or mentors, the employees' supervisors, or OCWTP personnel.

Strengths of Summative Observational Research

Observation provides direct information about a trainee's behavior on the job, and provides an excellent opportunity to observe higher levels of learning. While it predominantly measures skill, it can also indirectly measure knowledge (Parry & Berdie, 1999).

Liabilities of Summative Observational Research

The liabilities of observational methods of research are as follows:

- The quality of the data depends largely on the quality of the observation and its documentation.
- Observation methods can produce data that are difficult to summarize and analyze. The less structured the observation method and the more complex the program, the more troublesome these difficulties become (Rossi & Freeman, 1993).
- If the desired behaviors are not standardized, clearly articulated, and agreed upon by multiple observers, the level of inter-rater reliability is often very low, and the measures are less trustworthy.
- Observation requires high levels of training of observers. This can be time consuming and costly.

Barriers to Summative Observational Research

The barriers to observation include:

- The time and expense required to conduct the observations;
- Direct observation methods are not easily taught;

- The expense of developing a valid and reliable format for measuring and analyzing the observation;
- The availability of trained observers to conduct the observations;
- The potential intrusiveness of conducting an observation in the field.

Using Observation to Evaluate Skill Demonstration

Skill demonstration may be measured during training using role-plays. Unlike a pre-test/post-test of knowledge, a role-play would require a trainee to integrate the skills they are learning into a practical application. While role-plays are relatively common in training, effectively evaluating the role-play is another matter. An effective evaluation of the trainee's skill would require a standardized measurement tool, time to conduct the observation, and a trained observer.

It is also possible to develop a pre-test/post-test design for observation. For example, at the beginning of a workshop on legal issues, trainees could be asked to role-play a court setting, and a structured observation would be completed during the role-play. Another role-play would be conducted after the training is completed, and the pre-test and post-test scores on the observation would be compared.

Advanced Summative Research Designs

Kirkpatrick's third level of training evaluation focuses on transferring new learning to the job. This level corresponds to the AHA Level 7 - Skill Transfer. This advanced level of summative research is the most complex and requires the most time and expense. Other methodologies described earlier can be applied to this level with modifications to the research design. The primary methods of training evaluation research at this level include observation and the review of administrative data.

Using Observations to Evaluate Skill Transfer

The evaluation of skill transfer could occur in the field as part of an on-the-job coaching or mentoring activity. Supervisors or coaches/mentors could be used as observers, provided they were trained in both observation and documentation. For example, if a supervisor's rating was part of the caseworker's overall evaluation, the supervisor might accompany the caseworker on a home visit, to court, or to an interview, and conduct an observation at that time. There are a number of assumptions in this example, including that the supervisor is trained in observation, knows the nature of the behaviors being observed, that the family consents to the observation, and that there is a standardized measure developed to validly and reliably conduct and record the observation.

Administrative Data in Summative Evaluations

The review of administrative data involves the examination of various written records in order to measure the degree to which skill demonstration and skill transfer have occurred. Almost all human service programs are required to keep records. However, these programs vary widely in the quality of their records and their organizational capacity to effectively maintain and store them. One of the major concerns in administrative data review is whether recorded data are thorough, accurate, and reliable. However, they can provide some information related to the quality of job performance (Rossi & Freeman, 1993). Examples of administrative data that could be reviewed include case notes and recordings, risk assessments, family case plans, Individual Training Needs Assessments, and trainee/participant action plans.

Three key considerations should govern the design and use of program records in administrative review. First, a few items of data gathered consistently and reliably are generally more helpful than a more comprehensive body of information of doubtful reliability and inconsistently collected. Second, recording forms should be structured as checklists whenever possible, so recorders can check off various items rather than write out narrative information. Such a procedure not only minimizes the time required to record the data, but also yields data in the form most convenient for analysis. Finally, completed

records should be reviewed carefully and as soon as possible for consistency and accuracy. Reports on data gathered from administrative review may be presented in either narrative or statistical form (Unrau, Gabor, & Grinnell, (2001).

The same constraints related to observation also apply when using administrative data. These include the need for a valid and reliable assessment tool, the time and expense associated with conducting and analyzing reviews, and the need for training the raters who will conduct the reviews. There is also a potential issue of confidentiality when evaluating client records.

Strengths of Administrative Data Review Research

The strengths of administrative data review include the following:

- The ability to evaluate elements of skill mastery and transfer, which is not possible with surveys, interviews, focus groups, and written tests.
- Much of the data needed for the review has already been gathered as part of the worker's ongoing employment, so no additional time is needed to collect data; and, the evaluators do not need to intrude on the worker's time to perform the evaluation (Parry & Berdie, 1999).

Liabilities of Administrative Data Review Research

- Case records in many service organizations may have uncertain validity, reliability and availability. Some agencies keep excellent records, while others do not (Rossi & Freeman, 1993).
- Since case recording is usually completed by the worker who performs the tasks, the data in the record reflects the worker's perception of activities and outcomes, rather than those of an objective observer.

Barriers to Administrative Data Review Research

- Training of recorders and development of standardized measures are required. This may be time consuming and costly.

- Supervisors or evaluators need to allocate sufficient time to locate and read the needed documentation.
- Developing data collection instruments and analyzing the data is time consuming and expensive.

VI. Considerations and Recommendations for the OCWTP

The continuum of research designs and data collection methodologies presented in this document provides a general overview of the options available to the OCWTP to help determine the most appropriate next steps in its evaluation of OCWTP training. In this section, additional considerations and issues will be identified, and recommendations will be presented.

Internal Validity

Internal validity is an important factor to be considered when choosing an evaluation methodology. Internal validity is the degree to which an effect observed in a study was actually produced by the experimental stimulus, and is not the result of other factors (Royce & Thyer, 1996).

For example, when evaluating knowledge comprehension using a post-test only design, trainees would complete a series of multiple-choice questions at the completion of training. However, their scores could as easily be the result of the knowledge acquired on the job or in previous training. It would be difficult, if not impossible, to attribute the trainees' scores to the training, unless other factors are ruled out or controlled by the research design. Once again, it is important to balance threats to internal validity against the other pragmatic factors, such as available resources and ethical considerations.

There are nine major threats to internal validity. They are:

1. **History:** Outside events and variables can confound study results. For example, if a trainee has already been required to testify in court prior to

attending Core 100, her experiences may influence her test scores more than attending the training.

2. **Maturation:** The passage of time can affect the results. For example, in a longitudinal study of skill transfer, caseworkers may learn over time to identify risk factors associated with abuse, rather than as a result of training.
3. **Testing:** The mere experience of completing a measure such as a pre-test can influence the results of the post-test. For example, taking a pre-test for a workshop could prompt the trainees to look up answers before they complete the post-test.
4. **Instrumentation:** There may be problems with the instrument used and the potential for lack of standardization. For example, observers monitoring skill transfer of trainees could interpret the variables differently, or modify their criteria without realizing it.
5. **Statistical regression:** When evaluating trainees who had very low scores on a pre-test of knowledge acquisition, the odds are that this group will show a greater degree of improvement on a post-test than trainees who had a high score on the pre-test. This is a common problem in research.
6. **Selection biases:** Using comparison groups may not have any meaning unless the two groups being compared are really comparable in composition and have similar characteristics. For example, there would be little benefit to comparing skill demonstration of supervisors with a graduate degree in social work with caseworkers who have an undergraduate degree.
7. **Experimental mortality:** Subjects often drop out of an experiment before the study is concluded. For example, it might be problematic to complete a longitudinal study of skill transfer of caseworkers over a five-year period, given the high employment turnover rates of caseworkers.
8. **Ambiguity of the direction of causal influence:** There may be uncertainty regarding which variable influenced which. For example, does a high degree of satisfaction with the training influence a higher level of

knowledge, or does a higher level of acquired knowledge increase trainee satisfaction?

9. Diffusion or imitation of treatments: Two groups being compared may not be that different. An example would be comparing the scores on a Core 101 post-test between members of a training group, and a group of caseworkers who did not complete the training, but who reviewed the handouts or attended a unit briefing on workshop content.

Threats to internal validity must be considered when choosing a process of evaluation. Formative research that measures course, satisfaction, and opinion levels generally has a low degree of internal validity. In comparison, evaluations that incorporate control groups, random selection of evaluation participants, and standardized measures have a much higher degree of internal validity. Obviously, adopting a methodology with higher internal validity is more time consuming, requires more effort on the part of both the evaluator and the trainees, and is more expensive.

Barriers to Validity in Field Based Research

Even when opting for a methodology with a higher degree of internal validity, there are additional barriers that must be considered by the OCWTP. They include the ethical issue of withholding training from comparison group participants, the timing of training, and the inability to regulate the work environment.

Withholding Training

When attempting to evaluate knowledge acquisition, it is possible to provide the training to an experimental group of trainees, while withholding the training from a control group. The findings would provide a high degree of generalizability, but there are clear ethical concerns in withholding training from child welfare workers. There are similar ethical considerations involved in randomly assigning staff to either training or control groups, which is necessary for highly structured summative evaluation. However, this problem could be

addressed by scheduling members of a comparison group into needed training immediately after evaluation data has been collected. This would require careful planning and scheduling by the OCWTP, but it is certainly an option.

The Timing of Training

An additional barrier to collecting accurate data is how long workers are on their jobs prior to attending training. For example, while many caseworkers attend Core training immediately after being hired, that is not the case in all counties. Some caseworkers may have substantial work experience before attending training, which would impact their level of knowledge and skill when they attend the training. It is important to identify methodologies that can consider this factor in measuring outcomes. For example, in an effort to evaluate knowledge acquisition, a post-test only methodology would not determine if post-test scores reflect the effects of the training, or learning that has occurred on the job prior to attending training. A pre-test/post-test design would provide more reliable data. Including items in the survey about previous education and work experience would also allow OCWTP to use statistical procedures that could control for the effects of these other variables.

The Inability to Regulate the Work Environment

A variety of factors in trainees' work environments are likely to impact skill transfer. If the environment does not support the use of newly-acquired skills on the job, skill transfer may not occur, even if the trainee has the capacity to perform the skills after having attended training. Therefore, when implementing summative evaluations that emphasize skill transfer, barriers in the work environment must be considered in the evaluation study. Such factors include the availability of the supervisor, the supervisor's prior work experience, the supervisor's ability to provide educational supervision, the existence of policies and procedures that conflict with what was learned training, the amount of time available for one-to-one supervision, and the degree of support received by trainees to practice new skills. Summative evaluations of skill transfer should always include the identification of these barriers. Further, OCWTP should not generalize research findings about transfer of job skills from evaluations of

workers in a single county agency, or in agencies with similar work environments.

Calculating the Costs of Evaluation Activities

It is difficult to estimate the costs of any evaluation project without specific study parameters. According to Sharon Milligan Ph.D., Co-Director, Center for Urban Poverty and Social Change at Case Western Reserve University, evaluation costs are driven largely by the scope and scale of the evaluation study or studies, and the personnel and technology resources of the system needed to conduct the evaluation. Generally, costs increase as the complexity of the evaluation methodology is increased to strengthen the validity of the findings. Costs also increase with increased need for data entry and statistical analysis, the development and pre-testing of standardized measures, and an increase in the number of subjects included in the evaluation sample.

Below is a hierarchy of possible evaluation methods and their general level of cost (Parry & Berdie, 1999):

Low Cost Evaluation Examples: Participant satisfaction surveys; statistical counts of outputs, such as numbers of staff trained, and breakdown of data by workshop, location, job duties. This category also includes formative evaluation surveys wherein trainees rate the quality of the training, the curriculum, the trainer, and the training's relevance to their jobs and skill levels.

Low-Moderate Cost Evaluation Examples: Post-test-only measures of knowledge and skills.

Moderate-High Cost Evaluation Examples: Pre-test/post-test designs; evaluations that require follow-up measures or the tracking of individual trainees; on-site observation of skill transfer; extensive review of administrative data.

Highest Cost Evaluation Examples: Evaluation strategies that require comparison or control groups; and evaluations that attempt to correlate training to agency, client or community outcomes.

The costs of an OCWTP continuum of evaluation strategies can be estimated after the OCWTP determines the amount and type of evaluation information needed, the data collection methods to be used, and the number of cases required for reliability and validity.

Establishing Expected Outcomes for Evaluation

In planning any evaluation program, it is critical that OCWTP establish a common understanding about the expected outcomes of evaluation activities, define those outcomes in measurable terms, and prioritize those questions with highest systemic impact to be addressed first. As part of the initial evaluation process, the OCWTP must consider many questions, including:

- What is the purpose of the research?
- What are the specific research questions?
- What resources are available to conduct the research?
- What are the obstacles to conducting the research?
- How can these obstacles be overcome?
- Where and when should the research be conducted?
- What data should be collected?
- Who should participate in the study?
- What variables need to be measured?
- How should the variables be measured?
- What other variables, if any, need to be controlled and how should this be accomplished?
- What are the ethical considerations that must be considered?
- How should the data collected be organized and analyzed?
- How will the research findings be disseminated?

These questions must be considered when defining measurable outcomes, developing evaluation questions, and then choosing a research methodology. Each methodology has its advantages and disadvantages, and the OCWTP must weigh practical concerns, such as time constraints and available resources, when determining what data is needed for an effective evaluation.

Constructing a Chain of Evidence

The term "chain of evidence" is utilized in evaluation research to indicate the integration of various findings from a variety of research designs and data collection methodologies. Because of the barriers and limitations of any research strategy, causality (i.e., proving that training results in a greater degree of knowledge and skill among trainees) is difficult, if not impossible, to conclusively establish using any single evaluation strategy. Building a chain of evidence is one method to counter this problem. This is done by using a number of different methodologies to evaluate the same training at different levels.

For example, a group of trainees could first be asked to rate their own learning as a result of training. The trainees could be administered a multiple-choice pre- and post-test of knowledge to identify changes in knowledge as a result of the training. An embedded evaluation could be developed that tests how well a skill is being mastered during the training exercise. A cross-sectional survey of their supervisors could follow, to determine the supervisors' perceptions of the change in their workers' knowledge and skill levels as a result of training. Finally, observation of the trainees in their job setting could indicate whether the trainees can demonstrate the skills learned in the training.

The limitations of data collection strategies, discussed herein, would weaken the conclusions of any one of these methods if used alone. However, taken together, a sequence of evaluation strategies would help establish a chain of evidence that strengthens conclusions. The OCWTP should consider using this principle when developing all its evaluation strategies. In general, using several sequenced evaluation strategies on a smaller sample group will provide more accurate data than the large-scale use of any single evaluation method.

The OCWTP should continue to conduct formative evaluations. These evaluations have provided the OCWTP with invaluable information about the quality of its training and trainers. The OCWTP feedback studies to gather input from other stakeholders have also been invaluable in the design, delivery and revisions to training curricula and programs. In addition to these formative evaluations, the OCWTP should carefully consider the cost/benefit ratio of large summative research projects. It is possible to determine whether staff acquire

knowledge and apply knowledge and skill on the job. But because of the expense involved in this type of evaluation research, the OCWTP may want to pursue the use of “chain of evidence” research methodologies on smaller, more selective samples to gather data on the outcomes of higher priority training programs.

REFERENCES

Frechtling, J., & Sharp, L. (Eds). (1997). User-friendly handbook for mixed method evaluations. Washington D.C: Division of Research, Evaluation and Communication, National Science Foundation

The Institute for Human Services. (2001). Brief report and catalog listing of materials available to inform future OCWTP evaluation activities. Columbus, Ohio: Author.

The Institute for Human Services. (2002a). OCWTP statewide training assessment. Columbus, Ohio: Author

The Institute for Human Services. (2002b). Review of OCWTP feedback studies 1998 – 2001. Columbus, Ohio: Author.

The Institute for Human Services. (2002c). A review of the TrainTrack training management system for use in program evaluation. Columbus, Ohio: Author.

Kirkpatrick, D. (1959). Techniques for evaluating training programs. Journal of the American Society of Training Directors, 13.

Kohn, V. & Parker, T.C. (1969). Some guidelines for evaluating management development programs. Training and Development Journal, 23 (7).

Krueger, R.A. (1994). Focus groups: A practical guide for applied research. Newbury Park, California: Sage Publications.

Parry, C.F., & Berdie, J. (1999). Training evaluation in the human services. Washington, D.C.: American Public Human Services Association.

Rossi, P.H., & Freeman, H.E. (1993). Evaluation: A systematic approach. Newbury Park, California: Sage Publications.

Royce, D., & Thyer, B.A. (1996). Program evaluation: An introduction. Chicago: Nelson-Hall Publishers.

Rubin, A. & Babbie, E. (2001). Research methods for social work. Belmont, CA: Wadsworth/Thomson Learning.

Rycus, J.S., & Hughes, R.C. (2001). Levels of learning: A framework for organizing inservice training. Columbus, Ohio: Institute for Human Services.

Unrau, Y.A., Gabor, P.A., & Grinnell, R.M. (2001). Evaluation in the human services. Itasca, Illinois: F.E. Peacock Publishers, Inc.

Williams, M., Unrau, Y.A., & Grinnell, R.M. (1998). Introduction to social work research. Itasca, Illinois: F.E. Peacock Publishers, Inc.