Eighth Annual

2005

# National Human Services Training Evaluation Symposium

**May 25–27, 2005**

**University of California, Berkeley**

# Proceedings

***Sponsored by:***
***the California Social Work Education Center***
***in conjunction with***
***the California Department of Social Services,***
***the American Humane Association, and***
***the National Staff Development and Training Association***
***of the American Public Human Services Association***

# Proceedings
# *of the*
# Eighth Annual
# National Human Services
# Training Evaluation Symposium

---

The material in this volume is based on presentations and discussions at the Eighth Annual National Human Services Training Evaluation Symposium, held May 25–27, 2005, at the University of California, Berkeley.
*This volume is also available online at http://calswec.berkeley.edu/CalSWEC/CWTraining.html.*

*Co-sponsors*
California Social Work Education Center (CalSWEC)
California Department of Social Services
National Staff Development and Training Association of the American Public Human Services Association, Washington, D.C.
American Humane Association, Englewood, CO

*Published by*
California Social Work Education Center
University of California, Berkeley
School of Social Welfare
Marchant Building, Suite 420
6701 San Pablo
Berkeley, CA 94720–7420
*http://calswec.berkeley.edu*

*Editors*
Barrett Johnson
Michelle Henderson
Mark Thibedeau

---

## Table of Contents

## Acknowledgments

This year CalSWEC was again fortunate to work together with a talented group of individuals and organizations to make the NHSTES a success. We are especially grateful to the event's co-sponsors: the California Department of Social Services, the National Staff Development and Training Association of the American Public Human Services Association, and the American Humane Association.

Each year a very active steering committee provides sage guidance on the symposium program, as well as other aspects of the event. This year's members included Anita Barbee, Jane Berdie, Dale Curry, Midge Delavan, Bill Donnelly, David Foster, Shaaron Gilson, Michelle Graef, Henry Ilian, Mindy Ing, Michael Lahti, Chris Mathias, Michael Nunno, Cindy Parry, Debra Peel, Doug Pratt, Shradha Tibrewal, Naomi White, and Leslie Zeitler.

I would especially like to thank Leslie Zeitler, who provided stellar overall coordination of the symposium again this year. Other CalSWEC staff who were indispensable to the planning and implementation of the symposium included Rebecca Paris, Michelle Henderson, Loraine Park, Monica Asfura, and Terry Jackson. Karen Ringuette, Michelle Henderson, and Mark Thibedeau provided great assistance as co-editors of the proceedings.

Of course, special thanks also goes out to our presenters and facilitators, who again provided us with a great learning environment to move our field forward.

Barrett Johnson, *LCSW*
*Regional Training Academy Coordinator*
CalSWEC
Principal Editor of the *Proceedings*

# Introduction

This year our symposium focused on evidence-based practice, and how we can build evidence that training has an impact on outcomes for the clients we serve. We were privileged to have Eileen Gambrill facilitate discussions at the beginning and end of the symposium on what evidence-based practice really means, and the issues it raises for us in social work.

As training evaluators, we are caught in a bit of a bind about building evidence through evaluation and research. On one hand, we are advocates of more rigorous evaluation of training programs, so that we can promote efficiency and have a greater impact on outcomes for those we serve. On the other, we know that our evaluation efforts are frequently imperfect—we often lack the capacity for randomized controlled studies, and intervening variables (such as organizational factors and caseload sizes) often diminish our ability to establish direct links from training to practice to outcomes.

This is a tension we must continue to address. The implementation of (often imprecise) outcome measures and indicators in areas of practice such as child welfare services has shifted the culture of human services agencies. Managers and executives increasingly try to assess all programs in terms of their impact on set outcomes—including training. The federal Child and Family Services Reviews (CFSRs) support this trend.

An example from California's CFSR process illustrates this point. As the coordinating body for California's child welfare training system, CalSWEC played an integral role in meeting the goals of California's Program Improvement Plan (PIP). I attended many meetings with state staff to help formulate and respond to the training needs identified in the PIP. One particularly important and notable meeting brought together representatives from the counties, the state, and the federal government to discuss proposed tasks for the PIP. After several hours, we finally arrived at the items specifically related to the training system, and a top executive from one of California's most influential counties immediately raised his hand to speak. I thought he might speak

about the importance of training or talk about the excellent university-based in-service training program that the county uses. Instead, he said, "I'm not sure we really know that training has any impact on outcomes. Should we really even focus on it? Does it really matter?"

At this point I sputtered something about all the work we had done to establish a chain of evidence for the impact on outcomes, and stressed the importance of the *Framework for Child Welfare Training Evaluation* that had been included in drafts of the PIP. *(See the paper and discussion on the framework beginning on page 15.)*

This incident illustrates for me the importance of the symposium and all the work we do to better evaluate training. Through more rigorous evaluation, we not only can improve the effectiveness of training, but also make the case that training matters.

Using a model of evidence-based practice also may help us in this effort. As Dr. Gambrill illustrated so well in her opening and closing discussion, evidence-based practice promotes the use of the best evidence we have available, including research evidence and clinical expertise. It also integrates client values and preferences, which is so important in meeting the needs of our diverse clients.

As we move toward the symposium's ninth year in 2005, I am certain that we will continue to grapple with building evidence of training's effectiveness. Framing our efforts on an evidence-based model will help us to do this more effectively

Barrett Johnson, *LCSW*
*Regional Training Academy Coordinator*
CalSWEC
Principal Editor of the *Proceedings*

## Discussion: NHSTES Future Search 2015

*Theme:* Trends in Child Welfare Training Evaluation
*Title:* NHSTES Future Search 2015
*Facilitators:* Teresa Hubley, *M.P.A., Ph.D.*, and E. Douglas Pratt, *D.S.W., LCSW*

---

The purpose of the opening group exercise was to help participants maximize what they gain from the symposium by first enhancing their personal connections, having fun, and establishing trust. The exercises were also intended to identify common professional strengths and needs, and consider future developments in training evaluation. The expectation was that this process would result in a vision statement for the National Human Services Training Evaluation Symposium (NHSTES) in 2015.

The presenters divided participants into small groups to share their ideas, record them on newsprint, and then report them back to the larger group. The first exercise involved sharing what participants had been working on that they were most proud of in the past year.

Originally, the presenters were going to ask participants to discuss the current trends and recent past trends in training evaluation, including the social, scientific, economic, technological, political, and environmental trends. These trends would also include the main theories and main characters involved. As there was limited time to complete all of the tasks, the presenters moved to the third task.

The presenters asked participants to think ahead ten years into the future. The exercise involved participants visualizing where they will be, some of the barriers they have faced, and some of the milestones that have occurred, both positive and negative, which have come up in trying to reach their goals. Participants were asked to imagine what would be happening in training evaluation that is feasible, desirable, and that people would be motivated to do. They were asked to list what they may be doing in terms of methods, practices, programs and roles.

## Topics of Discussion

### I. Similar Discussions Emerged within the Groups

- Two main themes or topics jumped out from each group's reporting out: seamlessness and technology.
- There was also discussion of changing cultures, both organizational and the larger culture.
- Change cannot occur in a vacuum and change must occur at multiple levels simultaneously.
- There also seems to be an underlying assumption that there is going to be more information available faster to everybody and that we will know how to use it. One may look at that as a barrier or as a challenge.
- We must be careful about what technology does for us and what it might allow us to hide behind and hide from; what parts of learning necessitate human interaction and dialogue as opposed to the things that technology brings us, such as sharing information, data and access to training.
- Additionally, looking at who we are going to train is a big issue. It is important to consider how the "who" is changing and may change dramatically in the next 10 to 15 years.
- In thinking about positive changes and developments we are anticipating, it is important to remember that the history of social services is full of surprises. People cannot predict what changes will occur but they can predict that, at some point, major negative events will occur. It is vital to think about how to persevere and make positive gains in the face of negative events.

Newsprint from each of the groups was collected and collated to contribute to a vision statement, presented on the last day of the symposium. The vision statement for the NHSTE Symposium in 2015 was written as: *Relationships and technology work together so research on what works is integrated with all our levels of training evaluation, and administrations, stakeholders, families are seamlessly mainstream.*

## Discussion: How Do You Measure an Attitude or Value?

*Theme:* Addressing Fairness & Equity in Training Evaluation
*Title:* How Do You Measure an Attitude or Value?
*Presenter:* Katherine Cahn, *M.S.W., Ph.D.*

---

### Topics of Discussion

Dr. Cahn posed the question, "Who are we being as child welfare practitioners and evaluators?" She responded by stating, "All social work, but particularly child welfare starts with the self. It starts with who I am." This is as true for evaluation as for any other aspect of child welfare. Cahn conveyed a sense of urgency about the child welfare system in that Children of Color are significantly disproportionately represented in terms of length of stay, multiple moves, and other negative outcomes. Children of Color are less likely to be well-served by the system. She questioned the reasons for this disproportionality and discrepancy in the quality of service, and charged participants to explore what evaluators, as part of the system, can do about it. Her discussion centered on how to approach these two issues and the importance of continuing this discussion with other professionals in the field.

#### I. Disproportionality in Child Welfare

- The percentage of African Americans in the general population nationally is 13% and the percentage of African American children in child welfare is 47%.
- Dr. Cahn described a recent gathering in King County Washington. At the request of county leaders, evaluators provided basic statistical information to illustrate the discrepancy in services for Children of Color. In this particular county, 7% of the child population is Native American or African American. Yet, among the children in foster care for four years or more, half are Native American or African American.

- In King County, a Juvenile Court judge took on this issue of disproportionality and pulled together all the stakeholders in the county to try and do something about it. There were 65 leaders, including the head of Juvenile Court, some key judges, the head of CASA, many key people in the child welfare agency, people from the African American and Native American communities, foster parents, public defenders, and people from the Urban League, who sat down together to discuss the disproportionality of Children of Color.
- All the stakeholders attended a two-day workshop together sponsored by the People's Institute for Undoing Racism and Beyond to learn a common language around institutional racism and privilege. This experience illustrates that it is possible to get everyone on board but it took the work of many people to raise the conversation and commit to the work.
- It is urgent that this conversation take place. It matters that we understand how to do our work better for all of our children but we must address the fact that our system is working worse for Children of Color.

## II. Institutional Racism

- Our entire child welfare institution is living the power inequity. The operation of that power is institutional racism.
- We need to feel at home with the word "racism," start telling the truth about institutional racism, and teach our staff to understand that even though we are all well-intentioned and want the best, there is a dynamic that we walk into and that claims us when we walk in the office door.
- We need to understand power and privilege, and how they operate.
- We need to learn to be allies when and where we have privilege and to empower ourselves when we are in the non-privileged position.
- We need to be good and powerful at being allies when we are coming from positions of privilege so that we are undoing that every day and at every step of our lives.

- We must ask ourselves if our agencies are committed to undoing the institutional racism of the institutional structures of the agency, and are we committed to addressing this issue in the training classrooms that we are evaluating.
- For an evaluator, the issue is what do the people who meet with our clients need to know, need to do; what abilities do they need in working, not in a bi-cultural environment, but in a multi-cultural environment.
- Another element is to think about class as well, because not only are our attitudes about racial, cultural or ethnic groups potentially an obstacle to equity but attitudes towards class, within and across groups can also be a barrier.
- As trainers and evaluators, we are always part of training. By the questions one asks, by the data that one gives to trainees that surfaces the discrepancies, each of us are in very powerful positions to move this conversation along.

**III. Evolving from Cultural Competency Models**

- Many of us have been involved in this kind of conversation from the very beginning. The vocabulary for this conversation and the theoretical understanding of what is going on about this conversation has faltered recently, at least in many communities.
- An article on training for cultural competency, written only five years ago by a respected national training organization, focused on the approach that if everybody understood that everybody has a different culture and they talk about it, they have achieved culturally competence. There came a point where many were doing this in training and it was no longer sufficient to address the dynamics of racism.
- What is much more powerful and important is the introduction of the ideas of social justice, the critical thinking analysis and the introduction of the power analysis.
- What came out of some of the early work on the outcomes for Children of Color is that if it were just about color, then one would expect to see African American or Native American workers producing better outcomes because they were of the same race as the children they were serving.

However, we don't see better outcomes from race-matching staff to clients.

- There is a power inequity inherent to child welfare. No matter what your background is when you walk into a child welfare job you are wielding enormous power. On top of that position power, if you have privilege from being White, or by virtue of your gender, class, or the language you speak, by being straight not gay, or whatever it is that you have that is ascribed power in the dominant culture, there is something that you have access to that the children and families that you serve do not.
- Class is something that we have not addressed in cultural competence training but is critical to this discussion.

## IV. An Evaluator's Role in Systemic Change

Dr. Cahn posed the question: "What do we know as evaluators and what can we say about what we have learned as evaluators of training?"

- If our goal is to have training be part of the solution to the disproportionately poor outcomes for Children of Color (or framed in a positive way, if we want to have training be one of the pathways to a system that is fair and provides equally good services appropriate to each child and family's needs), then how do we need to proceed as evaluators?
- We need to begin looking at: 1) what we evaluate, 2) how we evaluate and 3) the context of evaluation.
    1) The "what" involves:
        a) How do we know what competencies we are looking for? For example, do competencies include the competency of being culturally responsive, understanding power and privilege, how to be a "stranger" and work cross-culturally?
        b) How do we know these competencies are getting learned in the classroom? Are we measuring cultural responsiveness as well as other competencies?
        c) How do we know competencies are transferred to practice out into the field?

2) How do we know this made a difference for children? Are we looking at differential outcomes? The "how" involves:
    a) What are the methods we use to conduct evaluation?
    b) Are there ways that those methods change or would we proceed differently so that we are actually getting the answer to the question and doing it in a culturally respectful way?
3) The "context" involves:
    a) What about us?
    b) What about our evaluation organizations?
    c) What about our training organizations?
    d) What about the schools of social work or the child welfare agencies in which we operate?
    e) Since "change begins with us," are we evaluating ourselves?

- An example of an evaluator's role may involve evaluating the capacity of in-service training programs to deliver a workforce that is qualified and competent to provide fair and equitable services. The evaluator must define what fair and equitable services are. California, for instance, has defined what they mean by "fair"; it does not mean only equitable or the same cookie-cutter services. It means the right service to each family.

## V. Obstacles in Training and Evaluation

- When evaluating a training on race and culture, responses on the worker satisfaction questionnaire may not be an accurate indication of whether the training was effective. Culturally responsive training or training about class, power, and privilege is not necessarily a comfortable training. This type of training gets personal and people's defenses come up. Trainees may fill out the satisfaction questionnaire and respond in various ways that reflect this discomfort, including:
    1) "This training was not relevant to my job."
    2) "The trainer was too strident. She had a chip on her shoulder."

3) "The presentation was disorganized. I want to see something more logical."

It is important to understand the context for these comments and not assume the training was ineffective.

- Often, people who train in this area have a different training style as compared to someone who is training (for example) on the three steps of risk assessment.
- The second evaluation issue is who turns in the survey. As evaluators, we rarely partialize our satisfaction surveys. A mean score of 4.5 means to us that "it was a good training". But we don't hear from trainees who are not speaking or those who didn't turn in their questionnaire. Who is saying what is important, particularly if we are not conducting an evaluation on paper but are doing it in a group. Look for who doesn't speak, answer questions, or who suddenly has to leave the training. Find out more from those who are traditionally silent.
- Another obstacle is evaluating knowledge or demonstration of competence in the classroom. You have the problem of trying to define what culturally competent is.
- One person may define a culturally competent worker as someone who, for example, can recognize that the marks on a child are burns that are caused by cultural healing practices versus those that are signs of ill intent towards a child. That may be an important distinction to know; however we are never going to get enough information to cover all variations in language, culture, class, first generation, second generation and newly arrived immigrants. We are never going to have a workforce that knows enough in all of the different areas.
- It is far more important to teach staff to be sensitive to understanding and learning cultural norms and how race and power play out in the gathering of information and drawing of conclusions in child welfare work. Staff needs to understand how to be a "stranger."

## VI. Measuring A Way of Being

- The learning that we are really talking about is affective learning. You really want to learn how to be a stranger, which leads one to ask questions including:

1) How do you be a stranger?
2) How do you measure how somebody is being?
3) How do you measure not knowing or the capacity to not know?
4) How do you measure someone's willingness to not know?
5) How do you measure "getting it"?
6) How do you measure someone who has learned to be an ally and has decided to take the risks?
7) How do you measure someone who has decided to let go of some of that internalized oppression and quit having other people's problems be their problems, and let them sort it out?

- Even at the level of knowledge, we have some problems of not knowing what the competencies are and that the competencies, if we did know them, are ineffable. They are *being* competencies, not *knowing* or *doing* competencies.
- The other side of the dilemma is, Does the agency want to know? Are they curious? Do they want their workers to receive training that leads to people being outraged and taking action until the institution changes? It may be that this is not the kind of training the agency wants.
- The next level of evaluation looks at whether the trainees use skills learned in the training. Yet, what would the skills be? What are the skills that people would use? Do workers come out as systems change agents? Again, you need a more robust and clear idea of what the training is meant to do. You want to have the right people at the table answering that question. As training evaluators, we have an enormous amount of power to ask questions and hold up mirrors.
- The final dilemma is, What difference did it make? \ What is the difference you want made? Can we look at better outcomes for kids, as that is the ultimate goal? You can have an intervention that makes equally better outcomes for all kids and you have not fixed the problem.

## VII. Considerations in Evaluating Attitudes and Values

- Evaluators usually conduct evaluations with pen and paper. But do all people even relate to an experience, turn it into a thought, and translate it into something written on paper?
- A problem with verbal evaluations is that they are generally done in groups and there are things about which one must be silent.
- Evaluators can draw on some of the research methods being done in communities of color.
- Another dilemma is whether to ask people abstract or concrete questions. Would it be better to evaluate trainings using concrete questions, role plays, where somebody could demonstrate that they are culturally competent? Would asking abstract questions, where participants respond with a theoretical response, work? This may only show that they are good at abstract thinking and may not actually show that when they are in a room with a child or family, they demonstrate the desired skills.
- In conducting research in King County, our research team made sure that all of the focus groups were run by people of color because someone who had not directly experienced racism or classism might miss some important content. It is important to think about who is going to collect the data, who is actually going to be doing the listening, and who will be interpreting it.
- Finally, we must understand ourselves in context and ask: Are we evaluating ourselves? Who are the trainers? Who are we as evaluators? Evaluators tend to be people who have participated in the dominant culture enough to achieve a doctoral degree. The ideas we have learned can be culturally or class determined.
- Do we value certain kinds of methodologies and not other kinds of methodologies? To whom are we obliged? To a certain degree, we can assign that to ourselves. There is a customer who is paying our check and we need to deliver a very good result for them, both rigorous and difference-making.
- Are we taking on advisors from communities of color? Do we have consumer voices?

## VIII. Conclusion

- If we want to be culturally competent, there is a way to set up and design our organizations so that we are beholden to the people with whom we want to make a difference. It means listening to and being taught by them.
- There are very good cultural audits available; they range from asking about the pictures on your wall and whether your staff are diverse, to who holds you accountable. The tipping point is that until approximately 25% of your agency's staff is from underrepresented groups, particularly visible minorities, there is not likely to be a safe enough space in the organizational culture to even have the conversations that you want to have.

# Training Evaluation in a Large System: California's Framework for Child Welfare Training Evaluation

Cynthia Parry, *Ph.D.,* and Barrett Johnson, *M.S.W., LCSW*

**Abstract**

This paper discusses several issues associated with evaluating training in a large, diverse, and de-centralized system. California's experience in developing a statewide training evaluation framework is described in order to illustrate the tension between the need for standardization imposed by a large-scale evaluation, and the need for county-based and regional training entities to be sensitive to local context. Issued explored include: item development; item functioning, content validity; and interpretation of findings.

California has a state supervised, county-administered child welfare system, with a regionalized system for providing in-service training. The state funds the California Social Work Education Center (CalSWEC) to coordinate statewide training, as well as four university-based Regional Training Academies (RTAs) to provide training in each region. Los Angeles County funds the Inter-University Consortium (IUC) to provide training in that region, and some counties also provide most of their own training, including core training for new workers. As a consequence, the RTAs, the IUC, and some large counties have traditionally developed core curricula independently, and have developed different content and models for delivery of training. Prior to the move toward standardization described here, the RTAs and the IUC all had different content, delivered over different time periods, with different levels of evaluation of training.

This began to change in 2002, when the California Department of Social Services (CDSS) called upon CalSWEC, the RTAs, and the IUC to develop a strategic plan for training evaluation for the state. Together with representatives from several California counties and national consultants, these entities formed the Macro Evaluation Team, later part of the Statewide Training and Education Committee (STEC), to guide the development of a multi-year plan for training evaluation in California. This work was given added impetus by the federal Child and Family Service Review (CSFR) conducted in 2003. The CSFR review required minimum standards for ongoing training, and statewide, standardized, core training for line workers and supervisors. Additionally, the subsequent Program Improvement Plan (PIP) incorporated the work of the Macro Evaluation committee in the form of two related tasks:

> In consultation with CalSWEC, CDSS will develop a common framework for assessing the effectiveness of training that is aligned with the federal outcomes. (CDSS, p.220).
>
> CalSWEC and the RTA's will utilize the results of the evaluation of the models of mentoring to develop a mentoring component which will be included in the supervisory core curriculum. (CDSS, p. 222).

At the time of the NHSTES, the Macro Evaluation Team (with the consultants) has used the work of the strategic planning process to develop the *Framework for Training Evaluation.* The framework is in the process of implementation, as the new *Common Core for California Child Welfare Workers* is rolled out across the state. Implementation is in process for six of the seven levels of training evaluation outlined in the framework.

## Overview of the Framework

California's training evaluation framework is based on the notion of levels of training evaluation that build on one another to form a "chain of evidence" for the effectiveness of the program as a whole, as well as evidence of the effectiveness of particular courses (Parry, Berdie, & Johnson, 2004). The chain of evidence refers to establishing a linkage between training and desired outcomes for the participant, the agency, and the client such that a reasonable person would agree that training played a part in producing that outcome. In child welfare training it is often

impossible to conduct studies that would establish a direct cause and effect relationship between training and a change in the learner's behavior or a change in client behavior, since these studies would involve random assignment. In many cases, ethical concerns prevent withholding or delaying training (or even a new version of training) from a control group.

To achieve a chain of evidence about training effectiveness, it is useful to develop a structured approach to conducting evaluation at multiple sequenced levels (lower levels being those most closely associated with training events). It is possible to gain a picture of the effectiveness of the training program as a whole by targeting evaluation of a specific course to the level that is most appropriate based on course competencies and learning objectives, while building the total framework to include evaluation at multiple levels over all course offerings. This approach conserves resources and instructional time, by making routine collection of evaluation data at all levels unnecessary. Evidence from individual courses shows whether staff attend the required training, perceive it as helpful and relevant, acquire knowledge and skills, and transfer these skills to the job. Taken as a whole, the evaluation evidence supports the case that training is positively affecting practice, and provides a foundation for linking training outcomes with outcomes for children, youth and families.

### *Levels of Evaluation*

There are several conceptions of levels of training evaluation. The framework is built on a modification of the American Humane Association (AHA) model for levels of training evaluation (Parry & Berdie, 1999). The AHA model expands Kirkpatrick's well-known four-level model to 10 more narrowly focused levels. These levels are tied directly to cognitive and affective levels of learning described in Bloom's Taxonomy of Educational Objectives (Bloom, 1956) and explicitly connect the formative evaluation of curriculum to other training outcomes.

AHA's model consists of 10 levels: formative feedback, satisfaction, opinion, knowledge acquisition, knowledge comprehension, skill demonstration, skill transfer, agency impact, client outcomes, and community impacts. These levels are arranged in order; however, it is not necessary to evaluate at all lower levels before conducting an evaluation targeted to higher-level competencies or evaluation questions. Lower levels are more directly related to training, and easier to measure. At higher levels,

training is one of a number of factors potentially impacting a given outcome, and direct connections to training are very difficult to document.

***California's Levels***

For the purpose of the California framework, some of the levels in this model have been collapsed, for a total of seven levels. "Knowledge acquisition" and "knowledge comprehension," for example, are combined into one level called "knowledge." The design of the knowledge testing captures both acquisition and comprehension. These levels and corresponding framework features, decisions, and requirements are described more specifically below.

**Level 1: Tracking Attendance at Training.** A system for tracking attendance at training is needed to ensure that new caseworkers are being exposed to training on all of the competencies needed to do their jobs, and that this occurs consistently across the state. This level is not included in AHA's or Kirkpatrick's model for evaluation for training, generally because it is considered implicit that an organization will track who completes training. In California's county-administered system, there was no capacity for collection of statewide data on completion of core training prior to the 2003 CFSR. Accordingly, STEC recommended that counties report to CDSS annually how many new child welfare workers and supervisors had completed twelve months on the job, and the proportion of these that had completed the new standardized common core. The RTAs and the IUC will report individual data on who completed the common core to the counties in their region semi-annually. The counties will report aggregate data to CDSS. STEC has recommended that new hires be required to complete Core training within 12 months.

**Level 2: Course Evaluation.** STEC initially identified five areas of core training as priorities for a standard evaluation. These included: Human Development; Critical Thinking in Child Welfare Assessment; Child Maltreatment Identification; Engaging Families in Case Planning and Case Management; and Placement and Permanence. Later, STEC recommended that these areas all have standard content, and added an additional area with standard content (but no knowledge evaluation): Framework for Child Welfare Practice in California. Collectively these content areas are known as the "Big 6." Under the direction of the Content Development Oversight Group (CDOG), a subcommittee of STEC,

the RTAs, the IUC and CalSWEC each took responsibility for leading the development of one of the Big 6. Each organization also provided feedback and oversight over the content areas that they were not leading. This collaborative model allowed for extensive review of the curricula by all entities that were required to train it, including observation of pilots. While this model was time-intensive, it simultaneously created an incentive to participate in the process and provided course-level evaluation. A standard form was also used to collect course-level data from the pilots and the initial roll-out of the training. Dubbed the "Plus/Delta Form," this tool allowed trainers and observers of the training to comment on strengths and recommend changes for each segment of the curriculum. Data could then be aggregated to inform and structure subsequent revisions of the curricula.

**Level 3: Satisfaction/Opinion.** The RTAs/IUC and counties currently use a variety of forms to collect participant feedback on the quality of training. The Macro Evaluation Team deemed standardization of this form unnecessary for the framework, since it would involve a great deal of effort to change forms used by different universities and institutions. CalSWEC and one of the RTAs provided a form as resource, and the RTAs and the IUC were given the option of linking the satisfaction forms to the standard demographic form and the knowledge—and skill-level data if they wished—simply by using the same unique identifier. Interestingly, satisfaction data was not emphasized as a large part of the evaluation framework, even though this was the only level of evaluation that was implemented by all of the RTAs and the IUC prior to the implementation of the framework. (One RTA collected knowledge and some skill data, and the IUC collected knowledge-level data prior to implementation of the framework.)

**Level 4: Knowledge.** Knowledge testing is conducted for the purpose of providing feedback for course improvement and demonstrating knowledge competency of workers in aggregate. The framework specified knowledge-level evaluation for four of the Big 6. Pre- and post-tests were developed for: Human Development; Engaging Families in Case Planning and Case Management; and Placement and Permanence. Post-tests only were developed for Critical Thinking in Child Welfare Assessment, since this was a one-day training and a pre-test would have used too much classroom time. Initially, the same test form is used for

pre- and post-testing, although continued development of the item bank will enable the use of alternate forms in the future. Pre- and post-testing will be conducted until items are validated, at which point this design will be revisited. At that time the pre-test may be eliminated except for a random sample of training classes, if data show that trainees consistently don't know the material prior to training (thus making continued verification of this fact unnecessary), and post-test scores continue to fall in an acceptable range for indicating mastery of the material.

CalSWEC, together with its consultants plus subject matter experts from the RTAs and the IUC, developed and will maintain a bank of items to construct knowledge tests. To date approximately 250 items have been developed. These items have undergone extensive editorial review by the RTAs and the IUC, counties, CalSWEC, and consultants with child welfare and test construction expertise. A research or policy basis has been established for almost all of the item content (with the exception of some that seem based on conventional wisdom), and statistical item analyses are underway. Items may be added and validated on an ongoing basis as curricula are updated or new methods of training are implemented.

Data are collected from each participant but reported back only in aggregate to counties and CDSS. STEC chose this model with the goal of collecting data to improve the training system and provide evidence of its effectiveness in producing desired learning outcomes for workers as a whole. Knowledge tests under this model are carefully developed to be both reliable and valid, but do not require the level of rigor necessary for making decisions leading to supervisory actions. System-wide knowledge testing intended for making decisions about individuals was felt to make too many demands on limited resources, including classroom time. The self-generated identification code is used to link pre- and post-tests for analysis and to link test data to demographic data. Participants are informed of the purpose of the evaluation, confidentiality procedures and how the results will be reported and used, and test administrators have written instructions and training in how to administer and debrief evaluations and monitor the generation of the identification codes.

Following the pilot phase, CalSWEC will provide the RTAs and the IUC with statewide aggregate data and data from their own

trainings. They will use the results to determine the extent to which knowledge acquisition is occurring. Initially, however, the data will be used only to validate the items' performance, assuring that they are functioning properly.

**Level 5: Skills.** All the RTAs and the IUC are also implementing an application of knowledge/skills-based embedded evaluation for one content area of the Big 6: Child Maltreatment Identification. Embedded evaluation builds on existing exercises or designs new tasks to simultaneously teach and evaluate the trainees' level of skill acquisition while they are in the classroom (McCowan & McCowan, 1999). This avoids the expense of more typical skills-based evaluation, which entails observing trainees some time after they have finished the training activities. Embedded evaluation therefore promotes efficiency while it enhances both trainee learning and the quality of the evaluation data. Embedded evaluation can also help to document the extent to which learning is taking place in the classroom, and help to predict how much transfer could reasonably be expected to take place under optimal conditions in the field.

The Macro Evaluation Team, with the assistance of the consultants, developed an embedded evaluation for Child Maltreatment Identification that requires trainees to read and review case scenarios, and then assess whether physical abuse has occurred according to California law.

This evaluation, like the evaluation described in Level 4, is used for course improvement and demonstrating competency of workers in aggregate (not individuals). Since there is little likelihood that the majority of participants will have this skill prior to training, the evaluation is conducted at the end of training only. (Pre-testing of skills using performance tasks is often impractical since it is very time consuming, technically difficult, and costly.)

The same unique identification code—linked to the demographics form—is used for the Level 5 evaluation. Participants turn in evaluation forms before the trainer processes the evaluation exercise, keeping a carbonless copy for use during review with the trainer. For item security purposes, no scenarios or test forms leave the classroom. As in Level 4, the RTAs and the IUC and counties will have access to statewide aggregate data and data from their own trainings, and will use results to determine the extent to which skill acquisition is occurring.

**Level 6: Transfer.** There are two main activities included in the framework under Level 6. The first is a part of the curriculum development process described in Level 2; suggested transfer of learning activities will be developed for the Common Core as part of the curriculum development process for the each area of the Big 6. The second is an evaluation of the role of mentoring programs in transfer of learning. The results of this evaluation, which involves evaluating mentoring or field training programs at some of the RTAs, will be used to recommend additional transfer activities for the rest of the state.

This evaluation involves two phases. Two RTAs have participated in phase one of this project, which assesses the extent to which mentoring services:

- increase perceived transfer (by workers and their supervisors) of Core knowledge and skills,
- increase worker satisfaction with the job and feelings of efficacy, and
- contribute to improved relationships with supervisors.

Both mentored workers and new caseworkers who do not receive mentor services are being asked to rate their skills, the supervisory support they receive, and their job comfort and satisfaction at the beginning and end of a six month mentoring period. Their supervisors are also being asked to rate their worker's skills and the supervision they provide. If mentoring is effective, the evaluation should show a larger skill gain for the mentored workers than the workers who are not mentored.

After phase one is completed, one RTA will continue with phase two of the evaluation, which will focus in-depth on one skill and assess the skill in the classroom and on the job.

**Level 7: Agency/Client Outcomes.** Using the entire framework, California can begin to build the "chain of evidence" necessary to evaluate the impact of training on outcomes. Linking the outcomes of training with program outcomes is a complex process that requires the careful assessment of multiple competing explanations for any change that is observed. Building the supporting pieces necessary to show that training has an effect on worker practice is a first step in linking worker training to better

outcomes for children and families. California has chosen to focus on this necessary groundwork in the framework in order to build a firm foundation for future efforts to evaluate at this level.

## Discussion

### *Resource Issues*

Implementation of the framework involved significant resource allocation for each level. These resource needs were identified as part of the framework, in order to guide the development of budgets for evaluation activities across organizations and over several funding cycles. Identified resource needs are outlined below, for each level.

**Level 1: —Tracking.** Prior to the PIP, counties were not required to track completion of core training. Databases are therefore required to track completion of core for the counties, the RTAs/IUC, and the state. The RTAs and the IUC already tracked who completed training, and reported it to counties. This information was not aggregated by counties and reported to the state, however. Methods for tracking data vary widely in a large system with small and large organizations, ranging from paper files to sophisticated electronic databases. However, all the entities involved need to commit personnel time to maintain the databases and monitor submissions for timeliness and quality (both locally and centrally). Protocols and training for individuals involved in maintaining and submitting data also are needed.

**Level 2—Course Evaluation.** As mentioned previously, each lead entity needed to provide personnel and/or consultants for forming a workgroup, reviewing and revising competencies and learning objectives, reviewing literature as needed, developing new curricula as needed, and adhering to decisions and protocols regarding quality assurance and curriculum format. Each lead entity also committed personnel to participate in joint meetings guiding and tracking progress and developing the formative evaluation tools. Developing curriculum that everyone in the state could agree to provide is time-intensive, and required a great deal of effort by CalSWEC to facilitate a coordinated development and approval process.

**Level 4—Knowledge.** At the local level, RTAs/IUC and counties have expended staff time to review items for the item bank, and attend item bank training. Staff time is required to prepare and distribute paper tests, and to manage storage and

transmission of data (either data entry or scanning of forms). Quality assurance processes also must be developed and maintained to ensure data are read correctly. There are also costs associated with copying and mailing of paper forms, and trainer time for learning test administration and transmission procedures. Again, CalSWEC expended considerable resources supporting and facilitating the development of the knowledge items. This included consultant costs, costs to underwrite item-writing activities, and staff time to develop and assist in implementing protocols for administering the tests.

**Level 5—Skills.** Additional resources needed at this level include trainer/subject matter expert time, associated costs for consulting on evaluation design and scoring rubrics, and trainer time for learning to administer and debrief the evaluation. As with the knowledge tests, CalSWEC also assumed costs of gathering and analyzing the data, and producing reports for all of the entities to use to improve the training based on the evaluation.

**Level 6—Transfer.** Local resource requirements at this level of evaluation have included significant staff time to participate in planning the evaluation, review the design and instrumentation, complete and track completion of data collection instruments, enter data, and participate in project meetings to assess progress. A significant amount of consultant time was also required to design the evaluation, develop the evaluation instruments, develop databases, enter data, conduct analyses and write reports.

### *Evaluation Issues*

For knowledge and skill testing, the ability to demonstrate a clear relationship between the assessment tool and the curriculum helps provide evidence that the training is responsible for any growth in knowledge observed, and helps to build the chain of evidence that training impacts outcomes. In developing the framework, a great deal of time and energy was expended to determine what was needed for valid and reliable measurement at each level.

#### Item Development

First, quality test items are needed that are clear, accurate, unbiased, and relevant to curricular competencies and learning objectives. CalSWEC, the RTAs, and the IUC, and counties have invested considerable time and resources in developing and testing a bank of knowledge test items, as well as an embedded assessment of skill related to child maltreatment identification.

This is a particularly time-intensive process when the items must be approved and reviewed by multiple organizations that are going to provide the training (i.e., the RTAs, the IUC and the counties).

The item writing and review process for the knowledge tests began with development of questions addressing the common core competencies and learning objectives for the four areas identified for this level of evaluation. Some questions required recall of specific facts, definitions and requirements; others required trainees to use these facts to make a judgment or comparison after reviewing specific case scenarios. Items underwent extensive editorial review by the subject matter and evaluation experts identified by the Macro Evaluation Team. They were modified when wording was unclear or not consistent with California practice, or when more than one potential correct answer was found. Literature and policy reviews were conducted to provide an evidence base supporting item content whenever possible. A large majority of the items were documented with specific research, but where research was not available the best evidence was gleaned from the clinical expertise of subject matter experts. The embedded skill evaluation instruments were developed jointly by CalSWEC, the Public Child Welfare Training Academy (the lead agency for that curriculum), the trainer/curriculum writer, and the evaluation consultants. These instruments also underwent extensive editorial review prior to pilot testing.

**Item Functioning**

As part of the implementation of the new core curricula and the evaluation framework, CalSWEC and its consultants are analyzing all the knowledge-level items to assess their functioning. For knowledge test items, this means calculating item difficulty and discrimination indices, and assessing possible differential functioning across racial/ethnic groups, gender, and regions of the state. Items that are at the extremes of difficulty (either too difficult or too easy[1]), or poorly discriminating[2] will be dropped or modified. As sufficient data are collected, items will also be

[1] These items don't help to assess relative knowledge since everyone tends to answer them either correctly or incorrectly, and also, contribute to unreliability in measurement because they are prone to errors from carelessness or guessing.

[2] Poorly discriminating items are those that more able students tend to get wrong more often than expected, or less able students get correct more often than expected. Such a pattern is indicative of a problem with the item such as confusing wording or more than one correct answer.

evaluated for differential functioning across characteristics of test takers that are unrelated to knowledge, such as gender and race/ethnicity. Items that are significantly more difficult (answered incorrectly more often) for trainees of a particular racial or ethnic group, gender, or region relative to others will be modified or dropped from future tests.

This is a particularly important step in the context of a large and diverse state such as California. For test results to be interpreted consistently statewide, it is important to ensure that the tools that are used to measure the effectiveness of the training are not subject to differential interpretation based on regional practice or demographic differences that are unrelated to the effects of the training.

A similar process will be employed to assess the functioning of the embedded skills evaluation. In this case, performance of the students is evaluated by comparing it to a standard or criterion set by the subject matter experts guiding the development of the curriculum. In this criterion-referenced evaluation model, items that fewer than two thirds of participants answer correctly are flagged for further examination by the expert group. They are then analyzed to determine whether the source of the problem is related to the item itself, the stimulus scenario, the curriculum material, or the curriculum delivery emphasis.

**Content Validity**

Content validity relates to how well the test as a whole shows a clear relationship to the objectives and content of the course. Evidence of content validity usually rests on the linkage of test items to course content through a test plan, in which items are written to measure specific competencies and learning objectives. The relative proportion of items for a given subject depends on the emphasis given to that content area in curriculum. In California, the item-writing process began prior to the PIP and the establishment of standard content. Items were therefore written to address common competencies and learning objectives, under the logic that wide agreement on the items would dictate what was covered in the curriculum. As the PIP progressed, however, the curriculum development process described under Level 2 above allowed for enough standardization of the curriculum to link each item to specific content that was trained statewide.

***Relevance and Coverage.*** Two aspects of validity have been cited as important to judging a performance or skill task: relevance

and coverage (Forster and Masters, 1996). Relevance addresses the degree to which the evaluation addresses the skills and outcomes focused on in the classroom. Item development for the embedded evaluation of Child Maltreatment Identification occurred as a part of the curriculum development process and was directly tied to training content. The task required the participants to use a decision making model taught in the class to determine which of twelve factors (e.g., explanation of injuries, location of injuries, etc.) in a written scenario were possible indicators of abuse. Participants then make a decision about whether or not abuse likely occurred, and document how they support their decisions. The development of the evaluation tool to mimic the decision making process taught in class serves to provide evidence of content validity.

The second aspect of validity is coverage, or the degree to which the evaluation samples the range of expected outcomes in the learning situation. Coverage issues were addressed by the use of four scenarios, some of which were designed to be abuse situations and some of which were not, to examine the participants' ability to apply the model in a variety of situations.

**Implementation Fidelity**

Effective evaluation involves not only valid test items that are explicitly linked to a written curriculum, but consistent and faithful delivery of the curriculum and implementation of evaluation procedures. Two processes were put in place to address implementation fidelity. The course-level review of the curriculum implementation (outlined under Level 2, above) helped to analyze differences in training delivery across regions. Standard evaluation protocols were also developed, which spell out procedures for test construction and administration. For example, the protocol for knowledge testing calls for use of pre- and post-tests of 25 to 30 items each, and recommends that 45 minutes be allowed for pre-tests and 30 minutes for post-tests. It also specifies test administration and security procedures. To maintain security of the item bank items, participants turn in their tests before leaving the classroom, and trainers have been asked not to debrief specific items, although they may answer general questions. Trainers have been given detailed instructions for introducing, conducting and debriefing the assessments, as well as for explaining the unique identifier process and how results will be used.

### Consistent Interpretation of Evaluation Results

In order to generalize the evaluation results to a large and varied state such as California, data must be interpreted consistently by all of the entities administering and evaluating the training. The state needs the ability to create multiple forms of a test that function equivalently, and that still allow for a consistent interpretation. This is important for several reasons: it accommodates some local variation in policy and practice; it allows for updates to both curricula and test forms over time, without sacrificing the ability to aggregate information from across the state; and it allows tracking of achievement of trainees from year to year.

***Item Banking.*** This need has been addressed partially by the development of an item bank, or pool of test items that have been keyed to learning objectives and/or competencies, scaled and validated. Item banks are frequently used in large-scale educational testing programs to form multiple versions of a test with known properties. Item banking software has been reviewed and a program called EXAMINER has been selected and purchased. RTA, IUC, and county representatives have received initial training on its use and will receive further training. They also have been provided copies of the EXAMINER software and may choose to construct their own knowledge items and item banks for any of their courses they wish to evaluate.

***Rasch Modeling.*** Analysis using Rasch modeling (Wright & Stone, 1979) will provide a basis for interpreting test results by accounting for differences in responses on a given test, as well as differences in the educational and skill levels of the participants in different classes and locations, within a common scaling framework. This model offers several advantages.

Ideally, measurements of a person's ability or attitudes should not depend on the characteristics of the items that make up the test or the range of ability of the group. Rasch is a probabilistic model, where probability of success on an item depends on the person's ability and the item difficulty. The mathematical formula chosen by Rasch to define this relationship allows the estimation of item difficulty independent of the abilities of the sample of persons taking the test (Wright and Stone, 1979).

It is also desirable that the scale represented by the test scores should be "equal interval". This property allows the conversion of scores on several tests to a common scale (Brown, 1976), and

provides for score differences that have the same meaning whether at the low end, the high end, or the middle of the scale. This latter property is important for measuring change—for example, a difference from pre to post-test of 5 points should represent the same relative improvement whether it is from 50 to 55 points (out of 100), or from 90 to 95 points.

Several common types of test scores are not equal interval. For example, larger raw score differences are needed to show an improvement from the $90^{th}$ to $95^{th}$ percentile than from the $50^{th}$ to $55^{th}$. The mathematical units defined by the Rasch model are logits, which form an equal interval scale. A person's ability in logits is their natural log odds of succeeding on items of the kind chosen to represent the zero point on the scale. An item's difficulty in logits is its natural log odds of eliciting failure from persons with zero ability.

The Rasch model has other advantages for measuring pre- to post-test change within the California framework. First, since it provides estimates of trainees' knowledge that are independent of the characteristics of the items used to make up the scale, if someone chooses not to answer some of the items, the model is able to accommodate the missing data and still provide an estimate of the person's overall knowledge that is comparable to estimates for the rest of the group. Similarly, if some items are dropped and new items are added to a test to reflect curriculum updates, Rasch analysis can still provide estimates of trainee ability that can be meaningfully compared to estimates from previous versions of the test. Finally, since Rasch estimates of a trainee's ability are also independent of the overall distribution of scores obtained by a particular group of trainees, the error traditionally associated with scores at the high or low extremes of the scale is reduced. Thus, scores can be compared meaningfully across different groups with varying degrees of prior knowledge and ability.

The Rasch model will be used to explore reliability and item functioning on for all tests derived from the item bank, and to provide summary pre- and post-test scores for each participant that can be compared by t-test or multiple regression techniques.

## Conclusion

Both development of the framework and the evaluation designs and tools that it calls for have presented complex logistic and

technical challenges. They have also involved the skills, experience and commitment of a broad range of professionals. At this time the common curriculum modules and evaluation tools are being tested and refined, and the strategies that have been devised for dealing with the issues and needs of a large and diverse system are being implemented. Much work remains to be done to fully realize the framework, and additional issues and challenges are likely. Given the complexity of the training system in California, however, the framework has served to structure the process of curriculum development and training evaluation. This structured approach has greatly improved the likelihood of a successful implementation of the statewide curriculum, moving California closer to tying training to changes in practice and improved outcomes.

## References

Bloom, B. (Ed.) (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals.* New York: Longman, Green.

Brown, F. (1976). *Principles of Educational and Psychological Testing (2nd Edition)*. New York: Holt, Rinehart and Winston.

Forster, M., & Masters, G. (1996). Performances: Assessment Resource Kit. The Australian Council for Educational Research, LTD. Melbourne, Victoria.

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

McCowan, R., & McCowan, S. (1999). Embedded Evaluation: Blending Training and Assessment. Research Foundation of SUNY/Center for the Development of Human Services: Buffalo, New York.

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

Parry, C., Berdie, J., & Johnson, B. (2004). Strategic Planning for Child Welfare Training Evaluation in California. In B. Johnson, V. Flores, & M. Henderson (Eds.), *Proceedings of the 6th Annual Human Services Training Evaluation Symposium* (pp 19-33). Berkeley, California: California Social Work Education Center.

Wright, B., & Stone, M. (1979). *Best Test Design: Rasch Measurement.* Chicago: Mesa Press.

## Discussion

*Theme:* Training Evaluation in Large Systems
*Title:* The California Framework
*Presenters:* Cynthia Parry, *Ph.D.,* and Barrett Johnson, *M.S.W., LCSW*

---

Cynthia Parry and Barrett Johnson presented on the development of California's framework for child welfare training evaluation. They provided a context in which the framework developed in response to the Federal Child and Family Services Review and shared challenges the state faced in working collaboratively with stakeholders in such a large training system. The discussion centered on the process of developing consensus in a large training system, ways of evaluating transfer of learning on the job, and consensus-based versus evidence-based practice.

### Topics of Discussion

**I. Finding Consensus in a Large System**

- The purpose of developing a framework off of which we could hang different evaluations was to address the many different cultures and organizations in California's large training system and establish a way for people to have a conversation about a statewide, standardized core curriculum.
- The process of coming to an agreement on the standardization of the curriculum was a laborious process.
- CalSWEC and the California Department of Social Services co-chaired the Statewide Training Education Committee meeting every other month. There were many repeat conversations from one meeting to the next, but that seemed inevitable in reaching consensus with so many different organizations working together on the curriculum and on the evaluation.

- For training systems that are much less complex than in New York City or the state of California, the framework still may apply and assist in planning for evaluation.

## II. Evaluating Transfer of Learning on the Job

- The plan is to evaluate transfer of learning on the job through the context of mentor programs.
- The goal was to do something more than self-report. It would have been ideal to actually see what people are doing with the skill taught in the classroom and observe how skills are applied in the field. Yet, we quickly realized that we did not have any means to observe and evaluate this.
- More curriculum work was needed to see how the skill was actually taught in the first place. Presently there is a two-phase evaluation of the mentor project:
  1) The first phase is self-report.
  2) The next step is trying to get pre- and post-meetings with the workers and their supervisors regarding their skills and their relationships.
- The evaluation also measures satisfaction with the mentors. This instrument was built from scales that Don Bowman and colleagues in Texas put together in their wisdom study, which was a large, workload burn-out/turnover study that included items related to training.
- Other items were liberally stolen from Dale Curry that have to do with worker transfer of learning and looking at differences pre- and post-training.
- The Northern Academy is experimenting with a slightly different model, where they have a transfer of learning specialist in the classroom to collect pre- and post-data. The Central Academy is using a model with mentors and the evaluation will compare the two models.
- In theory, there is a comparison group in which there is no special activity happening. The aim is to see if people in the mentoring programs have a greater estimation of their skill levels and feel a greater sense of efficacy, comfort, and job satisfaction. The evaluation will also explore the impact of the worker-supervisor relationship when a mentor is involved.

- Mentors are in the field with workers. They are Academy employees but are in the county system, so they have access to observe workers in the field for the engagement piece and have access to review case plans.
- The goal of the mentor evaluation is to look at the innovation that Central Academy implemented with having university-based mentors in the counties and be able to say more definitively that it increased the efficacy of the training.

### III. Consensus-Based versus Evidence-Based Practice

- This discussion and Dr. Gambrill's presentation raise the issue of the level or threshold of evidence that we want to accept.
- We often act on consensus. Consensus is a form of evidence but it may not meet the threshold of what one wants to act on. If consensus of something is reached, we may then want to build on that and change the research question to move it forward.
- There are extremists on either end of the evidence-based practices debate: social constructionists who think that there isn't any truth, and positivists who think evidence-based research can only be done in a particular way.
- In the field of human services, we need to have a conversation about how to proceed and what to do, including talking about the errors. We learn from our mistakes and have to get those out in a more forthright, acceptable manner.
- The philosophy of evidence-based practice stresses transparency. If something is based upon consensus, it cannot be called empirical in the sense that the relationships between the indicators and an outcome have not actually been empirically tested.

# Aiming at a Moving Target: A Multidimensional Evaluation of Supervisory Training—Two Years Later

Robert Highsmith, *Ph.D.,* and Henry Ilian, *D.S.W.*

## Abstract

The effort to evaluate a major training initiative brought unexpected consequences in the failure, over a two-year period, to achieve an adequate sample size, despite numerous efforts to boost the response rate. Lessons derived from the experience include: the design and execution of the evaluation need to align with the culture of the organization/agency; a draft of the final report should be written prior to beginning the evaluation in order to facilitate identification of major challenges; language used in the initial evaluation proposal should avoid raising expectations that are not achievable; and outside consultants should be intimately acquainted with the institutional culture of the organization being evaluated.

## Introduction

Final reports prepared for clients of training evaluations typically highlight the investigators' success in accomplishing their objectives. They tend to present a picture of research efforts carried out smoothly, with little difficulty in implementing whatever methodologies were employed. Hints may appear in the recommendations for further study about what may have panned out differently than expected, but little help is offered to shepherd investigators around the inevitable potholes that are encountered in most studies. This is so even though these evaluations are, of necessity, carried out within formal organizations that are often difficult environments in which to conduct any kind of research or evaluation. There is no way of knowing what proportion of research projects undertaken within organizations—including

training evaluations—do not work out as planned. Nevertheless, much can be learned by analyzing the experiences of those that—despite attentions to the usual precautions—yield outcomes different from what had been targeted.

One research tradition that offers guidance in making use of these insights is Action Research (Hindle, Checkland, Mumford, ,& Worthington, 1995). Although somewhat controversial because of its dual concerns with research and bringing about organizational change, Action Research typically takes as its subject for analysis the research methodology itself, and any unexpected developments in the course of carrying out the study. In one example from this tradition, Popper (1997) analyzes the unintended consequences of the introduction of a new employee evaluation system, which resulted in an outcome opposite from what he had expected. Cole and Knoles (1993) likewise examine the areas in which a study of teacher development did not go according to plan.

In terms of organizational demands, certain types of evaluations are more difficult to carry out than others. Kirkpatrick's (1959; 1983; 1996) four levels of training evaluation embody increasing levels of challenge in developing measures and collecting data. Assessing trainee satisfaction (Level 1) requires only a questionnaire administered at the end of the training session and is wholly within the control of the training entity. Measuring knowledge acquisition (Level 2) entails the more complex task of developing a valid and reliable test, but here too data collection occurs at the end of the training while the trainees are still in the classroom. Measuring transfer of learning as evidenced by change in behavior (Level 3) and the impact of training on the achievement of the organization's goals (Level 4) require developing and validating appropriate measures and collecting data once trainees return to their work locations. These more complex tasks require the active cooperation of managers for whom evaluating training is only one of many competing priorities. Consequently, evaluation beyond Level 2 is less frequently undertaken (Boverie, Sanchez, Mulcahy & Zondlo, 1995; Dyer, 1994; Eseryel, 2002).

Donovan and Hannigan (1999) cite a 1997 survey conducted by the American Society of Training and Development Benchmarking Forum that indicated that while 88.9% of

responding organizations used trainee satisfaction questionnaires, and 27.9% did testing, only 13.4% and 4.3%, respectively, examined transfer of learning and organizational impact. The survey also reported a belief among participating organizations that there is a diminishing return from Level 3 and 4 evaluations and perhaps even spurious or dubious results. Likewise, Todesco (1997), assessing the state of training evaluation in Canada, observes that "evaluation at the 3 and 4 levels, the transfer of learning and the impact on business, is still viewed by managers as too time-consuming and costly." Nickols (2005) believes this is the case because, after more than 40 years, the human resources development (HRD) community has failed to capture the commitment and support of relevant constituencies.

Despite these challenges, as Clarke (2001) notes, the critical importance of staff performance in determining the effectiveness of social service agencies argues strongly for measuring effectiveness training that is geared toward changing behavior (Clarke, 2001). Examining evaluations of training within social service agencies in the United States and the United Kingdom published between 1974 and 1997, Clarke underscores the urgency of bringing robust methodologies to evaluating training's impact on behavior. He concluded that, with respect to changing behavior, methodological limitations in most of the studies reviewed allowed "only tentative conclusions to be drawn" (Clarke).

This paper discusses insights culled from an evaluation that yielded different outcomes from the expected ones. The evaluation was intended to apply methodological rigor to gauging the impact of a major training initiative: to provide all of the approximately 1600 current, and all future, line supervisors in the New York City child welfare system with core training (SupCore) that develops a set of specified supervisory coaching and interpersonal helping skills. The aim is to improve supervisory practice throughout the child welfare system. Included in the training—and the evaluation—are supervisors employed both by Children's Services, the city's child welfare agency, and the large number of private agencies with which the city contracts for foster care, adoption and preventive services. The evaluation, which began in January 2003 and is still in progress, is being carried out by the Assessment and Evaluation Department of the James Satterwhite

Academy for Child Welfare Training, the internal training component of Children's Services.

In what follows, we briefly describe the research questions, sample and methodology used in the study. We discuss next our expectations in conducting the study and the challenges encountered in pursuing each objective targeted by the evaluation. We then present some positive consequences achieved in confronting the challenges. We conclude with a summary of what we learned that may be helpful to investigators facing similar challenges.

## Research Questions

The following questions were to be addressed by the evaluation:

- Do supervisors who attend the Supervisory Core Program assimilate its philosophy, develop appropriate competencies and skills, and apply these in their practice and on the job?
- If there are significant differences in behavior on the job, to what extent can these differences be attributed to supervisors having attended the training?
- What aspects of the Supervisory Core training (i.e., in-class training vs. technical assistance) explain observed differences in the way supervisors practice supervision on the job?
- Are there demographic characteristics that account for significant differences among supervisors with respect to their application of Children's Services philosophy, skills, and practice on the job?
- How do supervisors, their managers, and subordinates perceive the impact of the Supervisory Core training on those who participate?

## Method

### *Design*

In order to get as full as possible an indication of the state of the participating supervisors' practice, we chose to use a 360-degree assessment (Keenan, 1996; Nowak, 1993), in which participants evaluate their own efforts and are, in turn, evaluated by their subordinates and supervisors. In the classic 360-degree design, peers are also included, but we chose not to include this

element for two reasons: (1) to decrease the complexity of assembling a sample and (2) because the supervisors' work is largely independent, we anticipated that peers would not possess sufficiently detailed knowledge to provide a meaningful assessment.

In order to determine whether the training has a lasting effect, and whether the effect strengthens or weakens over time, the design called for administering the study instruments at three-month intervals over the year following the training.

To establish whether the training was responsible for any observed change in supervisor practice, the design called for a control group of supervisors scheduled to have the training late in the cycle. These supervisors would complete the study instruments on the same schedule as those who had already taken the training, and then repeat the process when it became their turn to have the training, thereby providing a comparison group but also a time-series design. This would strengthen our ability to draw conclusions from the study by ruling out noncomparability between the experimental group and the comparison group.

In order to assuage concerns among participants about confidentiality, we chose to present all data in aggregate, rather than by individual supervisor or unit.

### *Instruments*

The study made use of three measures:

- A one-page, machine-readable questionnaire to assess skills learned from the training;
- A one-page, machine-readable questionnaire to assess the impact of the training on the work climate of each supervisor's unit; and
- The pretest/post-test developed for the training.

The language used in the two questionnaires was taken from the descriptions of the skills and elements of work unit culture given in the printed curriculum.

### *Sample*

An initial proposal for the evaluation called for a sample of 150 supervisors, which would yield a 95% confidence interval. This was rejected as being too difficult to achieve. Additionally, although a random sample would have been preferable, labor/management concerns made this unfeasible. Therefore in the plan we adopted, we chose to recruit participants as volunteers who, in turn, would recruit their managers and up to five

caseworkers, at their discretion. The target sample was 90 participants/supervisors along with their managers and their caseworkers. Among these we planned to have a comparison group of 15, who, when they were scheduled for training, would join the experimental group.

## Results

### *Duration of the Study and Budget*

We naively expected the evaluation to be completed on time (within 18 months) and within budget. However, changes in the delivery schedule of training reduced the number of possible recruits, extending the deadline. Although the study began early in 2003, the fact that contractions in the New York City budget in the aftermath of 9/11 resulted in increased workloads meant that line-supervisors were less available to volunteer for the study and less willing or able to sustain their commitment over the required year. As a result, we experienced lower than expected recruitment. Among those who volunteered, not everyone followed through in recruiting their supervisors and caseworkers to return questionnaires for the post-test and subsequent post-training iterations. This necessitated extension of the deadline and unplanned expenditures, exhausting the initial budget before a sufficient number of participants had been recruited. A supplemental grant was secured, which aimed to increase response rates from approximately 30% to 85%.

### *Efforts to Increase Response Rates*

We also naively expected the response rates would increase because of our efforts to improve them. Heroic efforts, however, under the leadership of an experienced outside consultant, were unsuccessful. This was due to several factors: normal attrition of participants and their managers and caseworkers over the year following the training; reassignments from one unit to another and promotions; and dampened enthusiasm for “extras” due to the aforementioned contractions in staff and resources. The redirection of priorities for Children’s Services following the appointment of a new commissioner also resulted in termination of several agency contracts, further causing our sample to decrease. Moreover, institutional resistance to voluntarist initiatives like the Supervisory Core Evaluation was not successfully surmounted. As a result, frequent mailings of hard copies of evaluation materials, e-mail and voice-mail appeals, and on-site visits for the purpose of

assisting participants to complete questionnaires and tests were unsuccessful in raising response rates; in fact, they declined. (Some improvement in response rates occurred when an experienced CPS caseworker took over the management of the evaluation and used her intimate inside information of the workings of the agency to facilitate the evaluation.)

### *Participants' Recruiting of Caseworkers and Managers*

We naively expected, moreover, that the participants would be successful in recruiting their managers and caseworkers to participate each quarter. Managers were too busy with other priorities to respond to the Supervisory Core Evaluation, and some participants reported that their managers did not regard it as a legitimate part of the unit's work. Accordingly, managers' participation rates were considerably lower than caseworkers' and participants' rates, resulting in participation rates for everyone below what would have been achieved had the managers actively supported the evaluation. Not surprisingly under these conditions, participants found it difficult to recruit and/or sustain caseworkers' participation in the evaluation.

### *Unanticipated Outcomes*

Despite the naïve expectations, several positive and unanticipated outcomes have emerged from the evaluation. The study has generated over 900 observations of data to date that, at the very least, can inform other investigations and that can be used to test, with varying degrees of confidence, intuition regarding the impact of training on practice. The study also reinforced the importance of skills targeted by the training by providing periodic reminders for participants, managers and caseworkers to reflect upon and rate how well participants were practicing them. The SupCore Evaluation has also created a community of practitioners within Children's Services and its contract agencies that is willing to participate in, advocate for or otherwise involve itself in future evaluations. The expertise of the investigators and their ability to work collaboratively together have been enhanced as well by the evaluation; especially how to choose and work effectively with consultants.

## Discussion

As we responded to the challenges of the Supervisory Core Evaluation, we relearned a number of principles to keep foremost in mind for future investigations. First, the evaluation design needs to align with the culture of the organization or agency in which it is

conducted. Since the culture of Children's Services is hierarchical with priorities decided at the top and lower levels of the organization charged with implementing them, initiatives like the Supervisory Core Evaluation require the imprimatur of the highest possible levels of management within the targeted divisions to achieve desired rates. Had we sought and received the active support of the head of the Division of Child Protection, the largest single front-line division, in terms of staff, in Children's Services and the contract agencies, the response rates would have been substantially higher.

Second, a draft of the final report written prior to the start of the evaluation should be attempted to facilitate the identification of major challenges posed by the study and the requirements for addressing them. For example, a draft would have made us much more aware of our dependence on the good will of respondents who rely upon the managerial levels above them for signals regarding what's important. We would have invested considerably more effort in securing support at the managerial level before launching the evaluation. Such an exercise would also have forced early awareness and resolution of nettlesome issues about data collection, like selection of the appropriate unit of analysis. A draft also may have alerted us to the over-reliance on the good will of managers and caseworkers in providing primary measures of training effectiveness, and encouraged us to address the issue frontally.

Third, the language used in the proposal should avoid raising expectations that are unachievable or creating tensions about the implications for what's to be delivered. Promise only what you have a high likelihood of success in delivering. Goals, objectives, and outcomes should be written in the most precise language possible to maximize the investigators' ability to operationalize them and to ensure shared understanding of their meaning by stakeholders.

Fourth, to maximize the likelihood that the evaluation will deliver on its aspirations, select outside consultants who are intimately acquainted with the institutional culture of the organization being evaluated. Technical expertise is a necessary but not sufficient condition for success in working with consultants.

## Conclusion

Despite many challenges, the Supervisory Core Evaluation provided many insights about training evaluation in a large and complex system. In many ways, our expectations in conducting the study and in pursuing each objective targeted by the evaluation were not met. Nonetheless, unanticipated positive outcomes emerged from the challenges and what we learned may be helpful to investigators facing similar situations.

Two challenges remain as we move toward completion of the Supervisory Core Evaluation, one immediate and the other long term. Although we cannot claim, as we had originally hoped we would be able to do, that our methodology can lead to definitive results immediately, we need to mobilize all our creative powers to address the gaps in the data resulting from the low response rates, especially among managers. There are always gaps in databases, so our challenge is one of degree, not kind. In the longer term, we need to design evaluations that are synchronous with the fluid institutional context within which the child welfare system is likely to be operating for the foreseeable future.

## References

Boverie, P., Mulcahy, D.S., & Zondlo, J.A. (1995). Evaluating the effectiveness of training programs. John A. Zondlo /ACCESS Learning Systems. Retrieved May 15, 2003 from http://www.Zondlo.com/access/eval.htm

Chisolm, R.F., & Munzenrider, R.F. (1989). Evaluating a public-sector productivity improvement effort: An OD approach. *Public Administration Quarterly, 13*(1), (Spring), 91-111.

Clarke, N. (2001). The impact of in-service training within social services. *British Journal of Social Work, 31*(5), 757-774.

Cole, A.L., & Knowles, J.G. (1993). Teacher development partnership research: A focus on methods and issues. *American Educational Research Journal, 30*(3), 473-495.

Donovan, P., & Hannigan, K. (1999). Context and causation in the evaluation of training: A review and research outline. IMI (Irish Management Institute) Working paper. Retrieved May 15, 2003, from http://www.iminew/hrd/ARTICLES.HTM

Dyer, S. (1994). Kirkpatrick's mirror. *Journal of European Industrial Training, 18*(5), 31.

Eseryel, D. (2002). Approaches to evaluation of training: Theory & practice. *Educational Technology & Society, 5*(2), 93-98.

Keenan, J.P. (1996). Multi-source (360 degree) feedback: A case study and evaluation. In R.J. Ebert (Ed.), Proceedings of the 1996 Annual Meeting of the Decision Sciences Institute, (pp. 354-357). Retrieved June 11, 2003, from http://www.leadership-international.org/case/360degree.pdf

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

Kirkpatrick, D.L. (1983). Four steps to measuring training effectiveness. *Personnel Administrator, 28*(11), 19-25.

Kirkpatrick, D.L. (1996). Great ideas revisited: revisiting Kirkpatrick's four-level model. *Training & Development, 50*(1), 54-58.

Nickols, F.W. (2005). Why a stakeholder approach to evaluating training. *Advances in Developing Human Resources, 7*(1), 121-134.

Hindle, T., Checkland, P., Mumford, M., & Worthington, D. (1995). Developing a methodology for multidisciplinary action research: A case study. *The Journal of the Operational Research Society, 46*(4), 453-464.

Nowak, K.M. (1993). 360-degree feedback: The whole story. *Training & Development, 47*(1), 69-72.

Popper, M. (1997). The glorious failure. *The Journal of Applied Behavioral Science, 33*(1), 27-45.

Todesco, A. (1997, September). From training evaluation to outcome assessment: What trends and best practices tell us, a progress report prepared by Angie Todesco Learning Services Directorate Public Services Commission, Revised. Training and Development Canada. PSC (Public Service Commission of Canada). Retrieved May 15, 2003, from http://www.edu.psc.ctp.gc/tdc/index_e.htm

## Discussion

*Theme:* Training Evaluation in Large Systems
*Title:* Aiming at a Moving Target: A Multidimensional Evaluation—Two Years Later
*Presenters:* Robert Highsmith, *Ph.D.,* and Henry Ilian, *D.S.W.*
*Facilitator:* Norma Harris, *Ph.D.*

---

Robert Highsmith and Henry Ilian presented the last report on a multidimensional study launched three years ago evaluating supervisory training. Their presentation focused on what, despite their best efforts, didn't work. They provided a brief overview of the study, presented the final report written prior to the completion of their study that projected successes, and shared discoveries that caused their results to fall short of their expectations. They candidly shared their technical discoveries and invaluable lessons learned in the process. The discussion centered on how to address past, present and future challenges in training evaluation differently to achieve greater success.

### Topics of Discussion

**I. Technical Discoveries and Lessons Learned**

- Write very precise goals and deliverables.
- Be clear about the unit of analysis.
  1) The focus was on supervisors and training for supervisors but what we really wanted was data from their units. It was a mistake to expect supervisors to recruit managers and workers in their units.
  2) If only the supervisors had been asked to participate, without the rest of the units, there may have been a much better response rate. The problem seems to have been asking people to put in effort to persuade others to do something that they themselves might not have been fully persuaded to do.

- Avoid relying on processes that foreclose success.
  1) Workers do not like to evaluate their bosses' performances and managers do not like to evaluate the performances of those who report to them.
  2) The study was designed to get a reading on the effectiveness of the training, relying upon what amounted to an evaluation of the supervisors.
- An evaluation design needs to be compatible with the organizational culture.
- Limit promises to what you expect to deliver.
- Select consultants who are acquainted with the culture surrounding the evaluation.

**II. Effectively Involving Stakeholders**

- Given the scope of this evaluation conducted at different levels of the organization, how does one effectively involve, if not all the stakeholders, at least enough stakeholders at the different levels to avoid some of the pitfalls experienced in this study?
- Involving top management may have helped; if the head of the Division of Child Protection had actively supported this project, she may have sent signals through this very hierarchical organization to energize participants about the fact that this study will be done. There are five borough directors who probably should have been more involved.
- The feedback given by people was that there is no time to complete this evaluation because the "work has to be done." If we could have used stakeholders to help staff see that this is part of the work, we may have had more success.
- We should have reached out to managers to convey to the supervisors that "This study is part of the work."

**III. Addressing Institutional Culture**

- Children's Services is a large system and there are many different groups and components. Each office is likely to have a different institutional culture from the next office. Perhaps one way to design future evaluations would be to do it in modules (e.g. evaluations of special events,

programs, changes), rather than an evaluation of a whole system. In a system that large, the fluidity is difficult to adjust to. By the time you get to the second wave of your study, everything may be different.

- In such a large system, it seems vital to actually go to the offices, meet some of the stakeholders and participants, shake hands with them or have lunch together. Face-to-face contact can make a profound impact on someone's willingness to support a study.
- There was enormous fluidity and lack of stability in the organization. Although rapid turnover of caseworkers is a national phenomenon and is expected, we didn't expect supervisors to move around quite so quickly. Yet, supervisors often moved around due to promotions, transfers within the agency, and new assignments. This movement prevents the organizational stability necessary for institutional knowledge and memory, which must be considered for future studies.
- Develop a communication plan that goes along with the evaluation plan. Detail the how, what, when, where, and why of what it is you are trying to accomplish. Developing a formal communication plan allows one to continue to review it throughout the evaluation process to meet the established goals.
- A 360-degree evaluation, which sounds great at General Motors, has a very different meaning in a child welfare organization that in the past has been very quick to penalize and point blame but not quick to celebrate individual successes..
- An incentive may have prevented the dramatic drop-off in participation. The study was a heavy investment in a potentially fearful exercise. Finding ways to provide added inducements to supervisors who also have to ask other people to participate would have provided an extra measure of recognition to those supervisors that might have made the difference.
- In spite of our developing an incentive package to keep supervisors who volunteered interested from iteration to

iteration, one of the painful realities we discovered early on was that we would not be able to offer any of the incentives due to New York City Conflict of Interest regulations that prohibit city employees from accepting extra compensation for doing tasks that are part of their jobs.

# Discussion: Measuring Training Practice, Child, and Organizational Outcomes

*Theme:* Linking Training to Outcomes for Children and Families
*Title:* Measuring Training Practice, Child, and Organizational Outcomes
*Presenters:* Anita Barbee, *M.S.S.W., Ph.D.,* Becky Antle, *M.S.S.W., Ph.D.,* and Susan Kanak, *M.B.A.*
*Facilitator:* Claudia "Kyle" Hoffman

---

Anita Barbee, Becky Antle, and Susan Kanak discussed higher levels of training evaluation in child welfare. They focused primarily on the Louisville Child Welfare Training Evaluation Model as an example of how higher levels of training evaluation are successfully implemented. They discussed both barriers and ways that evaluators and researchers can overcome them in their organizational structures, in terms of collecting good data, linking training to transfer of learning and client outcomes, and engendering collaboration at all levels.

## Topics of Discussion

### I. Barriers and Supports

- Information systems enable this high level of evaluation. In Georgia, for example, it will be 10 years before they can look at outcomes because they don't have a system in place to gather the data. They are in the process of building the skills pieces, evaluating knowledge and some transfer of practice skills.
- Organizational buy-in from the beginning is critical. Kentucky had the buy-in from our child welfare system. We had to build those relationships, though, and it was slow. We had to resell ourselves.
- Keeping the same drumbeat going across time is very difficult when administration shifts occur. The new administration in Kentucky took 1 ½ years to figure out what was happening because they are a different political

party. All the files were cleaned out, so they were starting from zero. In some ways, evaluators are the institutional memory of the state because we have the data. We need to help them stay the course so they don't get distracted by politics, and to continue to build on all the hard work we have done.

- Kentucky is a state-administered system with approximately 250 supervisors and 2,000 workers. Training takes place consistently across the state. Having a smaller system is advantageous to being able to do this type of training evaluation. It is easier to collect the data and get the buy-in because we are working with a smaller group.

## II. Linking Training to Outcomes

The following questions were posed: "How did you connect, training to change in practice—to some measure of change in permanency? What measure of permanency did you use?"

- In Kentucky, all of the casework data are entered into a computer-based system and the state has it set up to extract data to measure federal permanency indicators, such as length of time in care, number of placements and achievement of permanency outcomes. The state produces monthly reports on this data for every region and they can extract it at the team and individual level.
- Evaluation at different levels in Kentucky:
  - √ We administered a number of surveys to the individuals who went through the training to test their knowledge.
  - √ We measured transfer in the first study through an archival review.
  - √ We identified case records for the persons who went through the training and for the control group.
  - √ We received the state management data for both individuals who went through the training and those in the control group, and evaluated all their casework outcomes.
  - √ Many systems allow only for a cross-sectional study; however, we track individual data over time, so that we know each person's training history, when they had it, who conducted it, and what they learned on their tests.
- We are starting to do some comparisons of the federal measurement methods for permanency and longitudinal

measurement methods. According to the literature, using the federal measurement methods is controversial because they collect data at a point in time, which can create biased outcomes. We are finding differences in permanency outcomes based on those two different measurement methods.

### III. The Impact of Multiple Case Workers in Evaluation

- Discussions focused on the possibility of collecting data on individual workers' caseload and client outcomes and attributing the outcomes to that worker, if clients have more than one child welfare worker.
- The cases in the Kentucky study were overseen by the individual workers at the particular time they were carrying the case. The researchers cleaned out the cases that had multiple workers to focus in on the worker who went to the training.
- There are some studies that look at how changing workers affects permanency for children. A Wisconsin study found that multiple case workers reduced permanency for children. One of researcher's doctoral students did a study of the predictors of kids who stay in care and found that if children change workers to work with specialty teams focused on permanency, they reach permanency faster. Other factors regarding case transfer that impact permanency are the number of different workers a child has, who they are, what their training is, and what their expertise is.
- One of the factors important to having an effective training system is to reduce staff turnover. The Children's Bureau has been made aware of these workforce issues and several symposia have been convened to address retention. States cannot meet the CFSR goals and ignore the workforce, as turnover is a predictor of negative outcomes.
- The training system and personnel system need to begin to work more closely together. The variables that make a great child welfare worker and the skills that one needs have been measured. It has been harder for the larger system to develop a method for choosing workers who are motivated, have congruent values, care about children, and have the necessary skills. The issue is larger than understanding that

training has an impact. It means getting the system to acknowledge that who does the work matters.

## IV. The Most Useful Pieces of the Process in Kentucky

- The state made the training mandatory.
- Training supervisors and their teams together was very effective. Kentucky is trying to have more of their trainings given regionally in teams. They are doing more refreshers and more reinforcement of training across the whole state. The mentoring model has been used but not in every region, in spite of their efforts to expand it into other regions, as the data reflects that reinforcement of training actually makes a significant difference.
- Further support for this practice model has been implemented over the past decade. The evidence was needed because it is difficult to use a solution-focused approach in child welfare.
- The administration in Kentucky is now more interested in research. They want as many variables as possible linked to see what is happening. They want more evidence-based practice for the training content. This study helped to win them over in terms of believing that this training evaluation is important and that having control groups for studies matters. They also see the importance of having evidence that supports their requests for additional resources from the legislature.
- The administration has placed an emphasis on training transfer in response to the data collected from this project.
- The ultimate goal is to be able to modify the training activities in the curriculum to see if changes in certain kinds of outcomes occur over six months or a year. There is great interest in doing more simulation activities and specific skill training. Workers should gain a wealth of knowledge, values, and skills in training; yet, what they need to learn are the most critical behaviors, which anchors help to narrow down. Fifty anchors were pared down across all the duties of child welfare to identify the most critical behaviors and the training was modified accordingly. This data will help to illuminate what in the field a worker needs to be able to do to make good decisions.

# Evaluating Trainees' Testifying Behaviors

Michelle I. Graef, *Ph.D.,* and Megan E. Potter, *Ph.D.*

## Abstract

The development and implementation of an evaluation tool for assessing the testifying skills of new Protection and Safety Workers is described. The project illustrates the process and the challenges inherent in trainee performance assessment, such as: working with subject matter experts to generate behavioral anchors; the use of holistic versus dimensional ratings; the difficulty of creating anchors for important, yet unobservable, factors; the practical realities of establishing inter-rater agreement; and the pros and cons of rating in real-time versus using videotapes.

## Background

### *UNL–CCFL and NHHSS*

The University of Nebraska–Lincoln, Center on Children, Families and the Law (CCFL) provides pre-service training to all new Protection and Safety (CPS) workers employed by the Nebraska Health and Human Services System (NHHSS). This six-month employment practicum period features a combination of classroom and field training and limited case assignments. This paper discusses the evaluation of one of the "Working within the Legal System" curriculum units, which is designed to develop trainee skills in providing effective courtroom testimony.

### *The Context and Structure of the Training*

Over the course of several classroom sessions, trainees learn the facts of a simulated case and develop a request to file a petition. The lawyer trainers prepare the class for testifying through lecture, discussion, and demonstration, using video and movie clips of the best and worst practices in witness testimony. Then, over the course of two days, mock adjudication and disposition hearings are conducted in a classroom. Each trainee has

the opportunity to testify twice in each of the two hearings. Lawyer trainers play the roles of Judge, County Attorney, Defense Attorney, and Guardian Ad Litem. Trainees take turns testifying for five to nine minutes, and each trainee's testimony is videotaped. Trainees experience questioning by the different attorneys, depending on the type of hearing and type of examination (direct, cross, re-direct, re-cross).

One of the lawyer trainers evaluates the trainee, in real time, on one set of performance dimensions. At the conclusion of the trainee's testimony, the trainer provides immediate feedback on these factors to the trainee and training group as a whole. Upon completion of the feedback in the court classroom, the testifying trainee takes his or her videotape to a separate room where a second lawyer trainer is waiting. Meanwhile, the courtroom testimony resumes with the next trainee. In the video room, the trainer evaluates the trainee's performance using a second set of performance dimensions. The trainer reviews the tape with the trainee and stops the tape to point out noteworthy behaviors. When the second review is complete, the trainee returns to the courtroom to watch other trainees and wait for his or her next testimony.

### *Impetus for Change*

As part of an overall redesign of our pre-service training program, CCFL took advantage of the opportunity to expand the regular training performance reports that were previously created by field training specialists. Supervisors receive detailed, individualized reports of trainee classroom and field training performance at regular intervals, and each report summarizes trainee performance on knowledge and skills assessments that occur during that training period. Trainers for this curriculum unit had created their own informal evaluation process. However, the information that was being collected lacked clarity and could not be interpreted by supervisors in a meaningful way.

## Designing the Evaluation

### *Former Evaluation Tool*

The former evaluation tool included 10 performance dimensions that were labeled, but not defined, and some dimension labels were not easily understood (e.g., Controlling Own Answers). The 10 dimensions were clustered into two sets of five, with the sets being primarily distinguished by when they were evaluated

(courtroom vs. video), rather than by cohesiveness of content. Each set included an Overall Competence dimension.

The format of the former evaluation tool included a 5-point graphic rating scale for each dimension, with qualitative and numeric anchors on the endpoints as seen below.

Proficient (5) ---------- Needs Improvement (1)

Scores included not only discrete values from 1 to 5, but also ratings of 3++ or 4-, depending on the rater.

### *Defining Existing Dimensions*

The initial process included defining the existing performance dimensions. A variety of methods were used, including questioning subject matter experts (SMEs; in this case, lawyer trainers), reviewing training materials, and reviewing evidence from legal scholars regarding the elements of effective witness testimony.

A few challenges arose in the process of defining the existing dimensions. For example, much of what SMEs said they looked for was not actually behavioral in nature (e.g., confidence) or was not readily observable (e.g., listening). Some legal constructs like *Articulation of Case Theory,* were important, yet difficult to describe in behavioral terms. SMEs often had a mental image of what constituted effective testifying but found it difficult to describe. Finally, the SMEs could not define either *Overall Competence* dimension, although they all insisted they were both essential to retain.

The initial definitions of the existing dimensions took the form of approximately 40 sets of positive and negative behaviors, distributed across eight performance dimensions. (The former evaluation tool included 10 performance dimensions, but the 2 *Overall Competence* dimensions were ultimately eliminated, resulting in 8 dimensions.) The following example from the *Response to Questions* dimension illustrates one set of positive and negative behaviors: For a rating of 1, formerly anchored by "Needs Improvement," the new anchor was "Continues to talk when there is an objection to a question or when the judge interrupts." For a rating of 5, formerly anchored by "Proficient," the new anchor was "Stops talking immediately when there is an objection to a question or when the judge interrupts; waits for a ruling before responding."

### *Selecting a Format*

After defining the existing dimensions in terms of numerous behaviors, possible formats were explored. The first consideration was a behavioral summary scale format, shown in Figure 1.

**Figure 1. Behavioral Summary Scale Example**

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Continues to talk when there is an objection to a question or when the judge interrupts | ? | ? | ? | Stops talking immediately when there is an objection to a question or when the judge interrupts; waits for a ruling before responding |

Even though 40 sets of behaviors had been developed, they weren't enough to satisfy this format, and many more behavioral anchors would have been required for ratings 2, 3, and 4. The bigger constraint was that, in a given testimony, trainees could potentially exhibit the entire range of behaviors, which would make selecting a single rating impossible. Therefore, the more important question to address was the frequency with which trainees exhibit particularly desirable or undesirable behaviors. With this in mind, a behavioral observation scale (BOS) approach was selected. With a BOS, each rating requires assessment of only a single effective or ineffective behavior. Using the previous example, the desirable behavior, "Stops talking immediately when there is an objection," can be selected and rated on a 5-point frequency scale, from "Never" to "Always."

### *Tailoring the Format*

To adopt this BOS format, one behavior from each of 40 sets of positive and negative behaviors was selected to target for assessment. Positive behaviors were chosen whenever possible, but negative or ineffective behaviors were sometimes more important to assess. A 3-point frequency scale was developed, anchored by "Never/Rarely," "Sometimes," and "Frequently/Always." A

boldface cue word was added to each behavior for quick reference. A rating of "Not Applicable" was added for those behaviors that trainees might not have an opportunity to exhibit, depending on the type of hearing or exam. Spaces were created for note taking, and a booklet format was used to eliminate the need to flip between pages during the rating process. Appendix A includes the final rating tool.

### *Testing the Format and Training the Raters*

Before implementing the new evaluation tool, trainers practiced using it by rating videotapes of previous trainees. The trainers watched each video together as a group, made independent ratings of the trainee, and then discussed their ratings to identify points of agreement and disagreement. As a result of the practice session, the 3-point rating scale was expanded to 5 points to capture variability in skills, and the point values for each rating were included in each rating checkbox to remind trainees about the desirability of the listed behaviors (i.e., more points awarded for more frequent display of positive behaviors; more points awarded for less frequent display of negative behaviors).

## Scoring the Evaluation

The scoring system includes a process for arriving at both a score for each of eight performance dimensions and an overall total score. One goal was to have scores that would be easily interpreted by trainees and supervisors. We suspected that raw scores would often be converted to percentages of the total score for each interpretation. To eliminate this need, we decided to develop a scoring rubric based on a 100-point scale, which automatically indicates the percentage. The inclusion of the "Not Applicable" category precluded meaningful summative scores because the total possible points would vary, depending on how many times "Not Applicable" was checked. Therefore, scoring had to be an average of the points earned only on those behaviors that were rated. To arrive at an average that varies from 1 to 100, the points assigned to each behavior had to vary from 1 to 100, depending on the rating selected.

In the final scoring system, each behavior is awarded 0, 25, 50, 75, or 100 points. For desirable behaviors, more points are awarded for higher frequency (e.g., Never = 0 points; Always = 100 points); for undesirable behaviors, more points are awarded for lower frequency (e.g., Never = 100 points; Always = 0 points). The

scores for each dimension are an average of the points earned across behaviors within the dimension, and the total score is an average of all points earned across all rated behaviors, rather than an average of dimension scores.

## Implementing the Evaluation

The new evaluation system was recently implemented, with promising results. Trainees are introduced to the form when they receive classroom training on how to testify, and they see their ratings after each testimony. The trainers are enthusiastic about the new tool and believe it improves the training and the evaluation process. Their acceptance of and investment in the new tool is considered a noteworthy outcome that is essential to the ultimate success of the rating tool.

## Challenges Still Ahead

Although the initial feedback has been positive, there are still a few challenges that lie ahead. One goal is to incorporate assessment of trainees' performance improvement over the course of the four testifying opportunities, including how they use feedback to improve in the areas where they struggled. One difficulty is that there are threshold effects, such that some trainees start out doing very well and others have lots of room for improvement. Another challenge is that exam difficulty varies (cross examination is more difficult than direct examination) within and across trainees. Routine assessment of patterns of change will require resolution of these challenges.

Currently, there are insufficient data to assess interrater agreement or interrater reliability. Because each trainee's testimony is only rated by one trainer and because trainers take turns being raters, it is essential that trainers have high agreement; a trainee's ratings should not vary as a function of who is providing ratings. To ensure high agreement, additional practice sessions with videotapes are required because there are not enough trainers to have more than one rater during live training sessions.

In the current scoring system, each behavior is equally weighted in its contribution to the total score, even though the trainers believe that some behaviors are more important than others. Although it is possible to assign different scoring weights

to the behaviors, there is question about the value of doing so, and this issue will therefore need to be further explored.

Finally, there is still some question about the most efficient means of converting raw ratings data to final scores that can be included in an electronic report. Various software are being considered for its ease of use and compatibility with other assessment and database software.

# Measuring Competence in Physical Restraint Skills in Residential Child Care*

Lorna Bell, *Ph.D.,* and Cameron Stark, *Ph.D.*

## Abstract

This study aimed to design and test an instrument that could be used to measure the competence of residential child care workers in using a physical restraint technique that was developed by Cornell University and taught in its Therapeutic Crisis Intervention (TCI) programme. At the beginning of the study, a video was made of ten trainees each practising the restraint technique. Thirteen trainers were asked to use their expertise to rate the performance of the trainees, which provided baseline data on the degree of variation in the trainers' judgements. A training instrument was then developed and used by the trainers to assess the trainees.

Two new assessors rated the performances of at least 80 trainees practising the restraint on video using the measuring instrument to assess the inter- and intra-rater reliability of the instruments.

The findings of the study yielded a wide variation among trainers when they used their expertise to evaluate the performance of trainees. There was greater consistency among the trainers when they used the measuring instrument to assess the performance of trainees compared with when they relied solely on their expertise to make judgements. The two new assessors who used the measuring instrument to assess trainees achieved a significantly greater level of agreement on scores than would have been predicted by chance.

## Introduction

### *Literature Review*

Children in residential care can exhibit disturbed and violent behaviour, which can manifest itself in aggression directed towards property, the child him/herself, other children, care staff, and other

* British English style has been retained in this paper.

adults. It is difficult to get an accurate assessment of the total number of all these aggressive acts, but a survey of 120 children's home units in Scotland, conducted on behalf of the Scottish Office by Harvey, J. in 1992, suggests that 39% of homes (49) experienced physical violence against staff members during a one-month period. Eight units reported five incidents or more, with the maximum number of incidents in a single unit being 23.

It is clear that the occurrence of aggressive behaviour is not an unusual event in residential units. In order to protect the aggressive child and others, residential child care staffs need to learn and develop ways of controlling the behaviour of such children. Although it has been the subject of debate and controversy for a number of years (see Utting, 1991), the means by which such control can be achieved in a safe and non-oppressive way is still unclear.

The Scottish Office survey of 1992 (Harvey) identified a range of sanctions used by staff to control children's behaviour which included restricting leisure activities (82% of units); being sent to bed early (74%); physical restraint (67%); control of pocket money (62%); extra tasks (46%); isolation (20%); withholding normal clothing (11%); grounding/staying in unit boundaries (6%); reduction in family contact (5%); loss of home visits (4%) and reparation for damage (4%). As early as 1981, however, Millham et al. concluded that the use of many of these types of sanctions was ineffective or inappropriate. However, staffs' lack of control over the children in their care can has a harmful effect. Staffs do need more ways to maintain control over children in care; however, in order to maintain safety, use of physical restraint is one means to retain control in an everyday situation.

The Department of Health (1993) describes physical restraint as "the positive application of force with the intention of over-powering the child" while Ross (1994) defines it as "a method of controlling violent behaviour, usually by means of two adults holding the young person who is exhibiting a high degree of violence and/or distress."

There are legal constraints on physical intervention to restrain or restrict liberty in the child care field. In Scotland, Section 12 of the Children and Young Person Act of 1937 makes it an offence for anyone over age 16 who has charge or care of a child under 16 to ill treat or expose that child to danger or the likelihood of danger. A person who physically restrains a child could also be guilty of criminal assault under Scots common law and/or be sued through the civil courts for assault or injuries to liberty. *(For a*

*detailed exposition of the legal implications of physical restraint in Scotland, see Watson, 1995.)*

In England, such legal constraints derive principally from the Children Act of 1989. They are explained in chapters 1 and 8 of Volume 4 of the associated guidance and also the Children (Secure Accommodation) Regulations of 1991 (SI 1991/1505) and the Children (Secure Accommodation) (no. 2) Regulations of 1991 (SI 1991/2034). The common law position on unlawful restriction of liberty and the criminal law relating to assault are also relevant.

There are also government guidelines with regard to the use of physical restraint. In 1993 in England, the Department of Health (DOH) published *Guidance on Permissible Forms of Control in Children's Residential Care* to provide some guidelines to staff on the use of restraint. Ross (1994) dismissed these guidelines as unhelpful as they are "far too vague" and their "assumption was seen as downright dangerous—leaving workers vulnerable and children at risk." In Scotland, the Skinner Report (1992) recommends that "the Social Work Services Inspectorate convene a working group to draw up guidance on sanction and control in residential child care." In April 1995 the Association of Directors of Social Work produced a document, *The Management of Conflict in Residential Childcare Establishments,* which reiterated the principles governing the use of restraint in the DOH (1993) guidelines and stated that "physical restraint should not be used to prevent a child from absconding, apart from extreme circumstances ...." (p 1).

As a consequence of these guidelines and laws, local authorities in England and Scotland have been exploring ways in which they could equip their residential care staff with the necessary knowledge and skills to appropriately control the young people in their care. Some authorities have elected to train staff in "Control and Restraint (General Services)," which was developed from "Control and Restraint" used in the Prison Service; others have trained staff in "Therapeutic Crisis Intervention (TCI)." This latter programme was developed by Cornell University in the United States in 1983 as a training package that could be presented to participants by in-house trainers who had previously attended a "TCI Train the Trainers" programme.

Since these studies have tended to focus on the acquisition of knowledge, evaluation of the effectiveness of the training of physical restraint skills has been limited. Although some studies have been conducted in the United States and a recent evaluation of TCI training was conducted in one Scottish local authority (Bell,

1995). There is very little information on whether or not trainees actually acquire physical restraint skills during training. Similarly little is known about whether or not these skills are retained months or years after training. Some of the methods that could be used to assess the effectiveness of skills training are fraught with methodological difficulties that make them unreliable. For example, one might attempt to compare the number and severity of injuries sustained by staff and children before and after training, since the use of correct physical restraint should reduce the number of injuries. Alternatively, one could compare the performance of untrained staff and trained staff in the field. However, Bell and Mollison's study (1995) indicated that there were so few incidents involving the use of physical restraints in one Scottish local authority in the six-month periods before and after the training that such comparisons would not be statistically significant. Even with more incidents, it would be extremely difficult to show that a reduction in injuries or a better performance by trained staff was the result of training and not the result of some other variable, such as a change in recording incidents, the size of the child, the size of the adult, or the location of the incident. It might also be possible to examine staffs' perceptions of their own competence. A study of Cardiopulmonary Resuscitation (CPR) skills by Marteau, Wynne, Kaye, & Evans (1990) indicated that training *does* increase practitioners' self-confidence. Unfortunately, the study also shows that staffs' increased self-confidence is not matched by an improvement in their competence.

Physical restraint skills are unusual in that there is often no established method for monitoring standards. Where a technique has become widely used, competence is assumed to be assured by training trainers in the technique and these trainers are then trusted to train others. The competence of the trainers and their professional standards are assumed to ensure that trainees will achieve an acceptable standard of competence.

CPR, which involves both mouth-to-mouth ventilation and chest compression, is another area where a complex set of skills are taught and where instruction is usually by a trainer. Research into this area suggests that there is a substantial reduction in competence in resuscitation skills six months after the training has taken place (Berden et al., 1993). The research also demonstrates that trainers differ in their assessment of competence and may rate their own trainees higher than the trainees are rated by another expert. Attempts to improve the consistency among trainers in their assessments have led to the development of standard scoring

systems (1993). The aim of this study was to develop a standard scoring system for physical restraint skills so that acquisition and retention of such skills could be measured more accurately.

The Therapeutic Crisis Intervention (TCI) programme was selected as the subject for this study because many trainers had been trained to instruct practitioners in this model of physical restraint in Scotland and in England. These trainers were in the process of training large numbers of practitioners throughout Scotland. For example, in 1994 Grampian Region planned to train all its residential child care staff in TCI over a period of a year.

In its Trainers Manual (Family Life Development Center, 1993), Cornell writes that the aim of the Therapeutic Crisis Intervention programme is "to provide child care workers with the skills and knowledge so that they can become the catalyst through which the child changes old habits, destructive responses, and maladaptive behaviour patterns. The goal of this training programme is to train child care workers to help children develop new responses to their environment that will enable them to achieve a higher level of social and emotional maturity" (p. ix).

The objective of the programme is to "provide positive, therapeutic, practical, and proven methods for managing children in crisis…. The programme focuses on preventing and de-escalating crises and is designed to build skills in these areas. Information and skill-building activities concerning intervention approaches; self-awareness; awareness of the child and the environment; behaviour management; communication; conflict resolution; and life-space interviewing are included. The life-space interview is a non-blaming supportive interview conducted with the child after the incident in order help the child understand what happened and to identify better ways of coping with similar situations in the future. The physical restraint techniques and principles presented in this training are done in a manner that conveys a sense of caring and protection to the child, and that maintains the dignity of both the adult and the child." (p. ix) Thus a major part of the programme is aimed at enabling workers to understand the nature of crises and to help them develop skills in de-escalating crises and undertaking life-space interviews and only a part of the programme is aimed at providing training in physical restraint skills.

TCI offers training in three restraint techniques: the team restraint, the single person restraint, and the basket hold restraint. In a three-day programme these restraints will be described and

then practised by participants on the afternoon of the second day and the morning of the third day.

The following is a brief outline of the team restraint suggested by Cornell University. (Note: They *must not* be used as a guide to practising the techniques.) In the team restraint two adults approach the child from the front and grasp the child's arms above the wrist with their hands. They then slide their arms under the child's armpits, gently pull the child's arms across their bodies, and secure the child's arms against their chests. The adults then take a step forward and kneel on the floor. This brings the child to the floor backwards. but the child's fall is broken by the adults touching the floor first and bringing the child down between them. The adults then roll the child over so he/she is face-down on the floor. One adult secures the child's arms and the other secures his/her feet.

## Research Aim and Objectives

### *Goal of Research*

The goal of the research as to design and test an instrument to measure competence in physical restraint techniques.

### *Research Objectives*

- To obtain data on how trainers would currently assess skills in physical restraint techniques.
- To assess the reliability of trainers current assessments.
- To get trainers to agree on the stages of a restraint technique; the possible errors trainees can make in using the technique; and the level of seriousness of each error.
- To design an instrument to measure competence in physical restraint techniques using the above information obtained from trainers.
- To obtain data on the trainers use of the instrument and assess its reliability.
- To test the reliability of the instrument through assessing its use by new assessors.

## Methodology

### *The Delphi Technique*

There are no published research studies which allow an objective assessment of the importance of the specific parts of the restraint techniques to their overall success or failure. Our aim was to develop a consensus on what errors could be made by trainees in using TCI restraints and on what ratings should be given to each of the errors. The Delphi technique appeared to be useful as it aims to structure

group discussion and opinion. It was developed in the 1950s by the Rand Corporation in order to generate discussion and enable a group to make judgements on a specific topic so that the decisions reached could claim to represent the groups' views. (Goodman, 1987). This method has been used in areas such as nursing research (Bond and Bond 1982) and primary health care needs (Burns et al., 1990; Hutchinson & Fowler, 1992).

The Delphi technique requires a panel of experts on the social or technological trend(s) of interest. Panel members are asked to provide independent forecasts of events they expect to occur and to identify the assumptions on which they base their forecasts. These independent forecasts are summarised in statistical form and key assumptions are identified. These summaries and assumptions are then provided to all members of the panel, and each member is asked to provide a new forecast based on this new information. These new forecasts are summarised and reported to the panel members, who are again asked to revise their forecasts. This iterative process continues until a consensus is obtained or no further changes occur in the individual forecasts. In practice, it has been observed that the Delphi technique seldom requires more than three or four iterations. Over the years several modifications have been made and, although originally it was a face-to-face method, more recently it has developed as a mailed method.

Several problems have been identified with the use of the Delphi technique. The use of experts has been challenged (Sackman, 1975) and, as Goodman (1987) points out, it can be difficult to define an expert in some subject areas. A clear rationale for the use of experts rather than a random sample of practitioners is therefore required. Mullen (1983) argues that a common method used in the Delphi technique, that of selecting items in a first round and then rating them, does not allow participants to reject an item altogether. Appropriate use of the technique should allow participants state a position on a particular item and to have that position related to the others in the group. Scheibe et al. (1975) argues that, as the technique promotes agreement (Sackman 1975), the stability of agreement should receive careful attention as a median score may mask a bimodal or flat response distribution. Given these concerns, we decided to use a modified version of the Delphi technique.

***The Measuring Instrument***

The measuring instrument as designed using the framework developed by Berden et al. (1992) who designed a valid and reproducible system for assessing competence in basic cardiac life

support skills. They defined five criteria for an effective measuring system, namely:

1. Inadequate techniques must be reflected by a fail score
2. Skilled persons should achieve a pass score
3. The effect of training must be reflected by an improvement in the score
4. Inter- and intra-observer unreliability must be negligible
5. The system must be simple to apply

It was intended that the design of the instrument in this study would match these criteria by the adoption of the following methods.

***Methods***

The Delphi technique was to be used to establish: a restraint technique, the errors that a trainee could make in using the technique, and the level of seriousness of each error.

Fourteen experts were invited to participate who were defined as experienced practitioners who had been trained by the Cornell University staff who developed TCI. Following their training, the practitioners were viewed by Cornell as able to train other people in TCI techniques. All the practitioners were involved as TCI trainers and had recently attended a refresher course. A U.K. trainer who was recognised by Cornell University as competent to run TCI Train the Trainer courses identified the practitioners as competent TCI trainers. The practitioners were from a range of statutory and voluntary agencies; 12 were from Scotland and two were from England. They were contacted by telephone to give them details of the project and to enlist their cooperation. All the trainers who were approached agreed, usually very enthusiastically, to take part in the project. Permission for each practitioner to participate in the project was sought from the head of the organisation for which the practitioner worked. In the remainder of the report these practitioners will be referred to as "experts."

**Stage 1**

Ten trainees were videotaped performing the team restraint. A copy of the videotape was sent to each participating expert, who was asked to rate the quality of each trainee's skills in the restraint on a five-point scale: *excellent, very good, good, poor* and *very poor*. They were asked to use their own expertise to make this assessment. Inter-rater agreement was examined statistically.

**Stage 2**

The experts were then asked to:

- Identify the stages of the restraint;
- For each stage to identify all the errors that could be made in carrying out the restraint;

- Give a rating of *minor, moderate, severe,* or *unacceptable* for each of the identified errors.

The experts were told *not* to include minor technical errors. They were expected to focus on errors that would either lead to injury or would lead to a failure to complete the restraint satisfactorily.

**Stage 3**

The results of Stage 2 were collated to show the distribution of responses, rather than a median or mean, and circulated to the experts. They were asked to review their original response in the light others' responses and to indicate whether or not they agreed or disagreed with the stages and errors identified and with the error ratings in the collated response chart.

**Stage 4**

The results of Stage 3 were collated. Where 75% or more of the experts agreed on stages or errors, these were accepted as final judgements by the experts and these stages or errors were kept as part of the measuring instrument. In collating the ratings of errors, where 75% of the experts were within one rating of each other, the arithmetical average was accepted as the judgement and these ratings were kept as part of the measuring instrument. The stages, errors, or ratings for which there was not 75% agreement were listed and recirculated to the experts, with the new collated responses. They then were asked to reassess their judgement. This process was continued until 75% agreement was reached for all stages, errors, and ratings. If this level of agreement had not been reached after collated responses had been circulated three times, then the majority error rating was accepted.

**Stage 5**

A measuring instrument for the restraint was then developed composed of all the errors and ratings for errors that had received 75% agreement from the experts. In order to quantify the errors, 5, 10, or 15 penalty points were assigned to *mild, moderate, or severe* aberrations, respectively. Twenty (20) penalty points were assigned to errors that would likely result in failure of the attempt or could cause injury to the restrainer or restrainee; these were deemed an *unacceptable mistake*. In order to grade the performance of a trainee, we assumed, as did Berden (1992), that no more than three minor, one mild, and one moderate, or one severe mistake would be acceptable to consider the performance adequate. A pass score was therefore defined as 15 penalty points or less.

The original videotaped scenarios were scored by the experts using the measuring instrument to assess whether or not it achieved greater consistency among them than their first assessment.

**Stage 6**

Many videotaped scenarios of each technique were produced. Two new assessors scored these scenarios using the measuring instrument, and the reliability and repeatability of the instrument were calculated using a statistical test, Cohen's Kappa (K).

## Results

### *Experts' Assessment of Video Scenarios*

In the first stage of this study 13 experts were sent a videotape of 10trainees, each of whom carried out the team restraint. One expert had dropped out at this stage due to ill health. The experts were asked to use their expertise to rate the performance of ten trainees. *Table 1 shows the collated responses for the restraint.*

This table shows a wide variation among trainers in their assessment of the performance of trainees. There is agreement about scenario 4, as all experts rated this as either *poor* or *very poor;* however, their ratings of other scenarios vary considerably. For example, for scenario 3, two experts rated it as *excellent,* four saw it as *very good,* and three saw it as *good,* while the other four experts rated it as either *poor* or *very poor.*

Some experts have a tendency to give low ratings for all scenarios. Expert 13 gives eight *poor* or *very poor* ratings for all restraints and, and he or she never awards an *excellent* or *very good* rating. Other experts have a tendency to give higher ratings. Expert 2 gives eight *excellent, very good*, or *good* ratings to all scenarios and rates only two scenarios as *poor* or *very poor*. Thus variations and contradictions among the experts occur with regard to what constitutes *good* and what constitutes *poor* performances of
the restraint.

A statistical test, the intra-class correlation coefficient (Fleiss, 1986) was used to calculate the agreement between trainers and the reasons for the differences between the scores awarded by trainers. This showed that the proportion of the difference attributable to differences between trainees was only 37.5% for the restraint and the remainder of the difference is caused by differences between trainers' judgements and random error.

### *Experts' Identification of Restraint Stages, Errors, and Error Ratings*

In the next stage of the study, the experts were asked to identify the stages of the restraint and the errors that a trainee could make in each stage. They then rated each error as *mild, moderate, severe* or *unacceptable.* As noted earlier, they were asked not to include minor technical errors.

**Table 1: 13 Trainers' Assessment of 10 Trainees Using Expertise**

Scores Awarded by Trainers (N= 13)

| Trainee | Excellent | Very Good | Good | Poor | Very Poor |
|---|---|---|---|---|---|
| 1 | 0 | 2 | 7 | 4 | 0 |
| 2 | 0 | 1 | 1 | 6 | 5 |
| 3 | 2 | 4 | 3 | 3 | 1 |
| 4 | 0 | 0 | 0 | 10 | 3 |
| 5 | 0 | 3 | 6 | 2 | 2 |
| 6 | 1 | 0 | 6 | 6 | 0 |
| 7 | 0 | 0 | 2 | 7 | 4 |
| 8 | 0 | 0 | 2 | 7 | 4 |
| 9 | 5 | 2 | 4 | 1 | 1 |
| 10 | 3 | 4 | 5 | 0 | 1 |

Where 75% of the experts agreed on a stage or error, these became part of the measuring instrument. In collating the rating of errors where 75% of experts were within one rating of each other, the arithmetical average was accepted as the judgement. When there was less than 75% agreement by the experts, then the lists of disputed errors and error ratings were sent back to the experts so that they could see the other experts' errors and error ratings. They were asked if they disagreed or agreed with the errors identified and were also asked to rate the errors again.

The collated responses to this first Delphi round were sent back to the experts and they were asked to reconsider their opinions in the light of their colleagues' judgement. When the responses to this second round were collated it was found that the experts agreed all the stages for the restraint. *(See Table 2.)*

**Table 2: Stages Agreed by Experts in First Delphi Round**

| Team | Restraint |
|---|---|
| 1 | Approach |
| 2 | Hold |
| 3 | Yoke |
| 4 | Take Down |
| 5 | Transfer Arms and Secure Legs |
| 6 | Roll Over |
| 7 | Straddle |

At this point they agreed on all but four of the errors, with agreement on 36 of 44 ratings of errors.

The disputed errors and error ratings were sent out to the experts for reassessment. Their responses showed that the experts agreed on all errors for the restraint but still disagreed on the rating of some of the errors.

The disputed error ratings were sent to the experts again; when these were returned there was agreement on all error ratings. The measuring instrument for the restraint was then produced by combining the agreed errors and error ratings from each of the three rounds in this stage of the project. *The results are in Table 3.*

**Table 3: Measuring Instrument**

| Error | Description | Score | Grade |
|---|---|---|---|
| 1a | Approach from front | 10 | Moderate |
| 2a | Grasps wrist or elbow | 20 | Unacceptable |
| 2b | Leader fails to immobilize child's arm | 15 | Severe |
| 2c | Leader wraps arms too low on child's body | 15 | Severe |
| 2d | Leader does not keep head low | 15 | Severe |
| 2e | Leader does not grasp their own wrist | 10 | Moderate |
| 3a | Pressure on elbow joint | 20 | Unacceptable |
| 3b | Holding child's arm at too high an angle | 15 | Severe |
| 3c | Staff members' hips not close enough to child | 15 | Severe |
| 3d | Arms not adequately secured | 15 | Severe |
| 4a | No nod from team leader before take down | 15 | Severe |
| 4b | Child's arms allowed to drift unsecured | 20 | Unacceptable |
| 4c | Shoulders of restrainers not close enough to the child | 15 | Severe |
| 4d | Staff step too far | 15 | Severe |
| 4e | One adult steps forward with inner leg | 15 | Severe |
| 4f | Adult kneels with outer leg | 15 | Severe |
| 4g | Child's head hits ground before the staff | 20 | Unacceptable |
| 4h | Put down too hard | 20 | Unacceptable |
| 4i | Failure to bring child's arm over their head | 15 | Severe |
| 4j | Failure to lean over child's body after the take-down | 15 | Severe |
| 5a | Team leader fails to move above child | 15 | Severe |

| | | | |
|---|---|---|---|
| 5b | Arms pulled into position | 20 | Unacceptable |
| 5c | Assistant lets go of child's arm before leader has a hold | 20 | Unacceptable |
| 5d | Leader presses on child's wrist | 20 | Unacceptable |
| 5e | Assistant does not lean over child as they secure child's legs | 15 | Severe |
| 5f | Assistant leans on child's knees | 20 | Unacceptable |
| 6a | Roll towards assistant | 10 | Moderate |
| 6b | Roll over not coordinated - 'corkscrew' child | 20 | Unacceptable |
| 6c | Assistant rolls child from legs rather thank hip | 20 | Unacceptable |
| 6d | Team leader leads roll over with child's arms | 20 | Unacceptable |
| 6e | Leader does not uncross their hold | 15 | Moderate |
| 7a | Leader straddles to face assistant | 20 | Unacceptable |
| 7c | Leader places weight on child | 20 | Unacceptable |
| 7d | Leader lies on top of child | 20 | Unacceptable |
| 7e | Assistant places weight on child's legs | 20 | Unacceptable |
| 7f | Held by wrist, elbow or shoulder during transfer of arms | 20 | Unacceptable |
| 7g | Passing hands down by lifting them off ground | 20 | Unacceptable |
| 7h | Sitting on child's arms | 20 | Unacceptable |
| 7i | Pressure on child's head | 15 | Severe |
| 7j | Too much pressure on child's head | 20 | Unacceptable |
| 7k | Failure to protect child's head | 20 | Unacceptable |
| 7l | Push child's face into floor | 20 | Unacceptable |
| 8a | Talk to child | 20 | Unacceptable |
| | Talk to colleague | 20 | Unacceptable |

During this Delphi process, 13 experts responded in the first round, 12 responded in the second round, and 9 responded in the third round. Sadly, during this project, one of the experts died and his co-trainer did not feel able to continue with the project. The Delphi process was time-consuming for the experts and the decrease in responses may be because their initial enthusiasm had waned by the end of the project.

### *Experts' Reassessment of Video Scenarios Using Measuring Instrument*

The video of the 10 trainees each using the team restraint was returned to the experts with the measuring instrument. They then were asked to reassess each scenario using the instrument. There was a distinct improvement in the consistency of experts' ratings but, because of the small numbers, it is not possible to calculate the statistical significance of the improvement.

The expert rated each error for each trainee as *mild* (5), *moderate* (10), *severe* (15), or *unacceptable* (20). The results appear in Table 4. There was still substantial disagreement among the experts even with the use of the measuring instrument. For example, the final score for trainee 4 for the team restraint ranges between 0 and 180. There appears to be a difference in the number of errors perceived by different experts; this difference is relatively consistent for every trainee. Expert 4 gives no score more than 30 in the restraint while expert 2 gives five trainees scores of more than 100 in the restraint.

Using 15 as the pass score Table 5 shows how many of the seven experts gave fail scores to each trainee for each restraint.

It would appear that while the experts are not consistent in their assessment of individual errors when using the measuring instrument, there is a greater consistency with regard to their evaluation of poor practice when they use the measuring instrument than when they rely only on their expertise as they did in Stage 1. This is illustrated in Table 5. In the final stage, using the measuring instrument, a greater proportion of experts agree that trainees' performance is poorer or worse than in the first stage.

### *Assessment of Scenarios by New Assessors*

One of the standard ways of assessing the reliability and the repeatability of measuring instruments is to compare the performance of two new assessors in their assessments of a large number of items or incidents. (Here "reliability" refers to inter-rater agreement and "repeatability" refers to intra-rater agreement).
A videotape was made of a large number of trainees practising the

**Table 4: Scores Given to Each Trainee by Each Expert Using the Measuring Instrument**

| **Trainee Expert** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 40 | 105 | 95 | 140 | 60 | 40 | 50 | 45 | 10 | 10 |
| 2 | 95 | 100 | 110 | 180 | 80 | 125 | 120 | 60 | 20 | 20 |
| 3 | 0 | 10 | 40 | 50 | 40 | 55 | 55 | 40 | 20 | 0 |
| 4 | 15 | 20 | 0 | 0 | 20 | 20 | 20 | 30 | 0 | 0 |
| 5 | 55 | 45 | 90 | 75 | 105 | 45 | 95 | 0 | 30 | 15 |
| 6 | 25 | 55 | 65 | 145 | 60 | 95 | 95 | 110 | 20 | 0 |
| 7 | 100 | 35 | 70 | 90 | 35 | 45 | 65 | 70 | 15 | 35 |

**Table 5: Difference Between Experts Scoring Using Expertise and Using Measuring Instrument**

| Trainee | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Experts giving fail score at final stage (N = 7) | 5 | 5 | 6 | 6 | 7 | 7 | 7 | 6 | 4 | 2 |
| Experts giving fail scores at Stage 1 (N = 13) | 4 | 11 | 4 | 13 | 4 | 6 | 11 | 11 | 2 | 1 |

**Table 6: Agreement Using Cohen's KAPPA**

| Technique | Observed Agreement | Chance Expected Agreement | KAPPA | Standard Error of KAPPA | One tailed p-value |
|---|---|---|---|---|---|
| Team Restraint | 0.92 | 0.89 | 0.27 | 0.087 | 0.00088 |

restraint. Some were videoed practising the restraint a number of times and were asked to deliberately make faults in their performances. There were 112 scenarios.

Two assessors, who were not experts in TCI techniques, observed the videos in different locations and assessed each scenario using the measuring instrument. Agreement between the two new assessors on whether a trainee passed or failed was high, i.e., 92%. Kappa was then used to calculate the agreement because K makes allowances for chance and gives credit only for agreement in excess of chance. (Streiner & Norman, 1991). Results are shown in Table 6.

Although agreement remained significantly better than would have been expected by chance, K values were low. This indicates that inter-observer variation was greater than would be acceptable in a reliable instrument. The results also suggest, however, that there was a systematic bias, with one observer consistently identifying more errors than the other.

## Conclusions

### *Key Findings*

**Wide Variation in Assessments.** One of the most surprising results of this study was the wide variation in the initial assessments of the performance of trainees on the video by the experts, relying on their own expertise to make the assessment. While it might have been expected that the difference between adjoining grades might be difficult to judge so that, for example, some experts would judge a performance as *good* while others would see it as *very good*, it is surprising that in some cases some experts judged a trainee's performance as *excellent* while others were judged it as *very poor*.

Such a variation in ratings among experts may be due to the speed of the performance of the restraint, disagreement on the correct way to perform restraints, disagreement on how seriously flaws in performance should be judged. Since we did not know what criteria or standard the experts were using to judge the trainees, it was not possible to ascertain the reason for this variation.

**Consistency While Using an Instrument.** There appeared to be a greater consistency among experts when they used the measuring instrument compared with when they used merely their expertise to assess a trainee's performance as pass or a fail. However, it is not possible to ascertain if this improvement is statistically significant due to the small sample size.

The two new assessors who used the measuring instrument to assess trainees achieved a greater level of agreement than would

have been predicted by chance. However, for the instrument to be entirely successful. an even greater level of agreement would be expected, with a perfect instrument achieving 100% agreement. Thus although percentage agreement between the two new assessors was high, the agreement in excess of chance was disappointing. However, the results suggest that reasonable intra-observer consistency could be achieved by improving the measuring instrument and by providing joint training sessions for assessors.

**TCI Application v. CPR Application.** TCI restraints involve a series of complex actions and are more complicated than CPR skills. A team restraint lasted approximately 16 seconds on the videotapes, and involved two restrainers moving their bodies, legs, and arms at the same time in a coordinated manner. The expert panel identified 44 possible errors a trainee could make in carrying out the restraint. That amounts to 44 errors in 16 seconds!. During the assessment of the scenarios for errors, it was necessary to stop the video constantly to get a frame-by-frame picture of each step, then rewind the video to view each step a couple of times. This had to do be done because if the assessor was looking to see what the leader's hand was doing, it was not possible to simultaneously observe what the assistant's feet were doing. It may be that the experts were more proficient in this respect, yet one expert suggested that he would have found the video exercise much easier if he had been provided with the picture in slow motion, as he found he had to keep stopping the video and rewinding it.

The scoring system upon which this project is based was developed by Berden (1992) to assess trainees' competence in CPR skills. However, there may be significant differences between resuscitation skills and restraint skills which make the latter more difficult to assess. Restraint skills are more complex and involve a more steps and happen over a briefer period of time. In Berden et al.'s study trainees were asked to perform cardiac massage and mouth-to-mouth breathing for at least two minutes, while in this restraint study the trainees' performance lasted for only a few seconds. According to the American Heart Association Standards and Guidelines for cardiopulmonary resuscitation, there are six manoeuvres which describe the quality of this technique. These six items, listed below, form the basis for assessing the techniques.

1. Correct placement of hands on the chest
2. A compression rate of 80-100 compressions/mins.
3. Compression/relaxation ratio of 1:0
4. Compression depth of 38-51 mm.

5. Ventilation volume of 0.8 - 1.21
6. Breathing interval of 1 to 1.5 s/min.

Only item 1 is assessed by means of instructor observation. The remaining five are assessed by a written print-out from the resuscitation mannequin, which is used by trainees to carry out resuscitation. This means that exact measurements are available, with little room for observer judgement. In assessing competence in restraint techniques, the experts were required to use their judgement, which could have accounted for the differences between their scoring.

***Implications***

This study has used the TCI programme as its subject. The focus has been on the process of the training, not on its content, and we are not able to make any judgement on the safety, and effectiveness of the techniques prescribed by Cornell University. Our findings relate specifically to TCI training, but we understand that other programmes which teach physical restraint skills adopt a similar training model; therefore we would surmise that our findings would be relevant to those programmes.

**Training Trainers.** In its Training Manual, Cornell University states that "the techniques and methods presented in this program represent what are believed to be the safest and the most effective methods for both staff and child. They are based on an extensive survey of the literature and current practice in the area of managing acute aggressive behaviour in children" (Family Life Development Center, 1993). The university expects its trainers to learn and then to teach a correct and accurate version of the TCI physical
restraint techniques.

However, the findings in this study suggest that the trainers, or experts, had in fact very different perceptions of what the restraint technique involved and had different perspectives on what was *good* and what was *poor* practice. This is evidenced in both the differences in their evaluations in Stage 1 and in their definitions of errors and their ratings of errors in Stage 2. Thus it seems likely that there is a distortion of the techniques as they are passed from Cornell University to its trainers, and again as they are passed from the trainers to the trainees. This means that what the trainees learn and adopt in practice may not be the techniques which Cornell originally developed and imagined that trainees would learn.

Organisations which purchase the programme because the restraints are, according to Cornell, the safest and most effective, will need to feel confident that their staff are acquiring an exact

version of the skills and not a modified version which has not been examined and approved by Cornell.

A considerable number of studies into CPR skills have assessed ways of improving the acquisition and retention of skills by practitioners and Moser and Coleman (1992) provide a useful summary of the recommendations from these studies. In applying these recommendations to physical restraint training it would appear that greater consistency among trainers could be achieved in a number of ways, which might include:

i. In-depth training of trainers so that they have a thorough understanding of the principles that underpin the restraint techniques.
ii. A systematic assessment of the trainers competence in the practice of the restraint techniques by Cornell University at the end of the Train the Trainers Programme and certification of that competence.
iii. Regular refresher training for trainers, which would require them to undergo re-assessment and re-certification of their competence in restraint techniques.

These measures would add significantly to the cost of training trainers.

**Lack of Consistency Among Trainers.** The finding that there was considerable variation among trainers has particularly important legal ramifications. Experienced trainers are sometimes used as expert witnesses to advise the court as to whether certain aspects of physical restraint are acceptable or unacceptable. This study has clearly demonstrated that further work is required to clarify what are the acceptable and unacceptable aspects of performance. This would ensure that trainers are relying on specific, clearly defined, and agreed targets, rather than on their general expertise as trainers, which is clearly unreliable.

**Improving Skills Acquisition and Retention.** Research into the acquisition and retention of CPR skills suggests that not all trainees achieve a uniformly high standard of practical skill during the training and, when re-tested months after training, the performance of many trainees declined. (O'Donnell & Skinner 1993).

The results of this study indicate that physical restraint skills are even more complex than CPR skills and have to be undertaken in a brief period of time. Thus it is likely that the problems of acquisition and retention noted for CPR skills will be compounded in the acquisition of physical restraint skills. This is illustrated both by the difficulty that trainers had in describing the restraint

technique and in the number of mistakes made by trainees in the first video.

However, research into the training of CPR skills has also identified ways of improving the acquisition and retention of skills which could be adopted by trainers in physical restraint skills. (See Moser & Coleman, 1992.) These include:

*Allowing sufficient time to practice restraint techniques.*

Research into the acquisition and retention of skills in CPR and in military settings suggest that retention of skills is enhanced when trainees are taught to saturation point so that they carry out the correct actions almost automatically.

Cornell University's TCI Trainers also recommend that there should be sufficient time allowed for practice of techniques during the training session; that refresher training should be built into the programme; and that the competence of trainees should be assessed at the completion of the training.

In the three-day TCI programme, trainees had the opportunity to practice the team restraint for one hour. As trainees were working in pairs during these practice sessions, there was a limited amount of time for them to practice the techniques during the programme and they were certainly not trained to saturation level. This shortage of time for practice has also been noted by Chadwick and Cooke (1995) in their evaluation of the TCI programme in Devon. They note that this issue "could only be resolved by extending the course, which would be our recommendation." We would concur. Cornell University states that the minimum recommended length of the TCI course in the direct training format is four days but that a five-day course is the ideal.

*Building in refresher training.*

Cornell University recommends that the initial programme should be followed by refresher training; yet, in a U.S. survey conducted in 1991 of 146 child care agencies that used TCI, the university found that only 69 of these agencies responded that they had refresher training and most of this was for four hours or less per year. (Cornell University, 1991). We do not have information on the extent of refresher training in Scotland but personal communication with trainers suggests that it is similar to the U.S. pattern.

In their evaluation of a short course in CPR training in a district general hospital, O'Donnell and Skinner divided the group of nurses who attended the initial training into three sub-groups. Group 1 underwent monthly refresher training, Group 2 had a single refresher after six months, and Group 3 had no refresher

training prior to re-testing of all trainees after six months. Over the six months of the study Group 1 showed a significant increase in skills. The other groups showed some increase in skill but this was of a much smaller magnitude and did not reach statistical significance.

Refresher training would appear to be an essential component in the training of skills and is an important issue to be addressed by trainers and managers when training in physical restraint is being planned.

*Assessment and accreditation of competence.*

In the current TCI programme, the competence of trainees to perform restraint skills is not systematically assessed and, at the end of training, they are not subjected to an examination of their ability nor do they receive certification of their ability. Thus trainees may complete the training but may not be practising the skill competently. Managers and trainers may wish therefore to consider introducing a more systematic evaluation of the performance of trainees and certification and re-certification of competence in physical restraint skills.

**Use of Video in Training and Assessment.** As noted earlier, physical restraint skills involve many complex actions in a short period of time and thus it is very difficult for trainers to make assessments of trainees' competence by simply watching them during training. Videotaping trainees makes this task much easier and trainers might wish to consider this during training so that they and the trainees can critically evaluate the performance of skills. They might also wish to consider using videotapes to assess the competence of trainees at the end of the training.

**Resource Implications.** The suggestions we have made for creating consistency among trainers and for enabling trainees to acquire and retain skills, such as extending the length of courses, providing refresher training, accrediting trainers and trainees, and using videos for training and assessment will have significant resource implications. These measures will make programme more costly to run and they will require residential child care staff to undertake more days of training.

**The Measuring Instrument.** There appears to be greater consistency among experts when they used the measuring instrument compared with when they merely used their expertise to assess whether the performance of a trainee should be regarded as a pass or a fail. The measuring instrument provides a means for systematically recording if and when trainees make errors when they practice the restraint. However, when the measuring

instrument was used to *score* performance, it was not particularly accurate and reliable. It would require further modification before it could be used for accreditation purposes. Despite the advice given to the experts not to include minor technical errors, it appears that some of these were included in the instrument. Therefore to improve the instrument, the number of errors needs to be reduced so that only those errors that could cause injury or would prohibit the satisfactory completion of the restraint are included. Some of the errors need to be defined more specifically and there needs to be more clarity about the definition of some errors.

### *Concluding Remarks*

Although we have identified a number of studies on the acquisition and retention of skills in CPR and in tank gunnery, we have been able to identify little research into the acquisition and retention of physical restraint skills. An unpublished study by staff at Ashworth Hospital in 1994 of "Care and Responsibility" training concluded that skill diminution becomes more marked as the interval between training and updating becomes more prolonged, and the ability to refresh skills seems to be less pronounced as the interval between training and updating is increased; however, we have not yet found any other research in the specific area of restraint skills in residential child care.

The lack of research is unfortunate, as there is a current concern that residential child care staff should be trained to be able to control the children in their care safely and effectively, and a number of programmes have been developed which purport to provide staff with those skills. At present, we do not know if staff acquires the required skills during training and if they do acquire them, whether they retain them for any significant length of time following the training. These questions need to be answered in order to assess whether or not this training is currently giving value for money and to determine whether or not the training needs to be altered to make it more effective. Currently organisations and residential child care workers themselves may believe that the training is properly equipping workers to cope with violent situations but at present there is no way of knowing if that is the case. Thus organisations and workers may be over-confident about the latter's ability to deal with aggressive young people and organisations may be placing responsibility on residential child care workers which is beyond their competence.

This also has serious legal implications. In the event of a charge of assault being brought against a member of staff who had

used a method of physical restraint, it is likely that the court would be interested in knowing whether or not the member of staff had implemented the technique accurately, as they had been trained; whether or not they had been trained correctly, according to the technique; and whether the technique itself could be considered as safe as possible

## References

Bell, L., & Million, A. (1995). An Evaluation of Therapeutic Crisis Intervention Training in Grampian Region: Final Report. University of Stirling, Stirling.

Bergen, H.J.J.M., Pijls, N.H.J., Willems, F.F., Hendrick, J.M.A., & Crul, J.F. (1992). A scoring system for basic cardiac life support skills in training situations. *Resuscitation, 23*(1992), 21-31.

Bond, S., & Bond, J. (1982). *A Delphi survey of clinical nursing research priorities. Journal of Advanced Nursing,* 7, 565-575.

Burns, T.J., Batavia, A.I., Smith, Q.W, & Dejong, G. (1990). Primary health care needs of persons with physical disabilities: What are the research and service priorities? *Archives of Physical Medicine and Rehabilitation,* 71, 138-143.

Chadwick, J., & Cooke, E. (1995). Managing Aggression and Violence : Review of Programme 1994/95. Devon County Council. Unpublished Report.

Chamberlain, L. (1993, November). Aycliffe closure call after SSI criticism. *Community Care, 19*(977).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20, 37-46.

Family Life Development Center. (1993). Therapeutic Crisis Intervention : Trainers Manual. Cornell University.

Fleiss, J.L. (1986). The Design and Analysis of Clinical Experiments. Wiley.

Gallagher, M., Bradshaw, C., & Nattress, H. (1996). Policy Priorities in diabetes care : A Delphi Study. *Quality in Health Care*, 5, 3-8.

Goodman, C.M. (1987).The Delphi technique : A critique. *Journal of Advanced Nursing*, 2, 729-734.

Harvey, J. (1992). A Review of Residential Child Care in Scotland. (HMSO). Edinburgh.

Hutchinson, A., & Fowler, P. (1992). Outcome measures for primary health care :What are the research priorities? *British Journal of General Practice,* 42, 227-231.

Levy, A., & Uchen, B. (1991). The Pindown Experience and the Protection of Children. Staffordshire County Council.

Marteau, T.M., Wynne, G., Kaye, W., & Evans, T.R. (1990). Resuscitation: Experience without feedback increases confidence but not skill. *British Medical Journal,* 300, 849-50.

Millham, S., Bullock, R., Hosie, K., & Haak, M. (1981). Issues of Control in Residential Child Care. (HMSO). London.

Moser, D.K., & Coleman, S. (1992, July/August). Recommendations for improving cardiopulmonary resuscitation skills retention. *Heart and Lung, (21)*4, 372-379.

O'Donnell, C.M., & Skinner, A.C. (1993). An evaluation of a short course in resuscitation training in a district general hospital. *Resuscitation, 26*(1993), 193-201.

Ruane, M., Eaton Y., McAuliffe, J., Tarbuck, P., & Thorpe, B. (1994). Care and Responsibility Training : Skills Diminution and Retention Survey. Ashworth Hospital. Liverpool. Unpublished Report.

Sackman, H. (1975). *A Delphi Critique.* Lexington, MA: Lexington Books.

Scheibe, M., Skutsch, M., & Schofer, J. (1975). Experiments in Delphi methodology. In H. Linestone & M. Turoff (Eds.), *The Delphi Method: Techniques and Application.* London: Addison-Wesley.

Skinner, A. (1992). Another Kind of Home: A Review of Residential Child Care. (HMSO). Edinburgh.

Stewart, D.W., & Shamdasani, P.N. (1990). *Focus Groups: Theory and Practice.* Newbury Park, CA: Sage Publications.

Streiner, D.L., & Norman, G.R. (1991). *Health Measurement Scales.* Oxford: Oxford University Press.

Utting, W. (1991). Children in the Public Care. (HMSO). London.

## Discussion

*Theme:* Developing Instruments to Measure Trainees' Acquisition of Specific Skills
*Title:* Physical Restraint Techniques
*Presenter:* Lorna Bell, *Ph.D.*
*Facilitator:* Jane Berdie, *M.S.W.*

---

Lorna Bell's presentation was based on her study aimed to design and test an instrument to measure the competence of residential child care workers in using physical restraint techniques. After describing the training, Dr. Bell shared how she developed and evaluated the instrument she used to measure competency in physical restraint skills, and included some of her findings throughout the various stages of the evaluation project. The discussion that followed centered on what was learned from conducting the study, soliciting client feedback on restraint techniques and the challenges of achieving inter-rater reliability.

### Topics of Discussion

**I. Lessons Learned**

- Using videos to train and assess trainees is effective; however, it was a time-consuming exercise to evaluate the videos due to having to stop and start it frequently.
- The decision to have only two non-experts review the scoring had to do with timing and recruitment. The project far surpassed the scheduled deadline. If there had been more time, we would have been able to learn more. It was difficult to recruit people to watch and rate 200 videos.
- Restraint skills are very complex, occur in a very short period of time and are difficult to learn and assess.
- Focusing on errors meant that if someone scored one "unacceptable" error, they failed on that particular restraint skill. There were many people who failed in terms of not reaching that standard. A question that remains is whether

it is useful to use an error-based model, rather than a positive- and negative-based model.

- Much in-depth training of these techniques is crucial. Trainees need to be taught to the saturation point, so they can do it automatically without needing to think about it.
- People were taught three techniques in two days and they tended to confuse them. Concentrating on one skill instead of bombarding trainees with too much information is probably best.
- It is important to assess and certify people as competent in doing these restraints and we have used this measuring instrument to make that assessment.
- While it is important to have refresher trainings and re-certification, a "boredom factor" can occur. One must find a way to maintain peoples' skills without boring them.
- There is great value in working alongside someone of a different discipline and applying some of the models they use to your own.
- This research had a major impact on how we revised our training curriculum and developed our certification system for trainers. It remains a challenge to train people so that they are faithful to a system, especially when the skill being trained is risky.

## II. Client/Resident Feedback of Restraint Techniques

- The purpose of this study was not to get feedback from residents, but to measure how well trainees had learned the restraint skills. What skills to learn should be informed by client feedback.
- There has been some literature on how residents feel about different techniques. Preferences vary among residents. Some of them find being restrained very safe. Perhaps we could learn from residents how best to do the restraints.
    1) The Buckeye Ranch in Columbus, Ohio, has developed a video that was informed by youth and provides their perspectives on being restrained.
    2) Cornell University is convening a symposium, and one sessions the research from the youth's point of view to address how they feel about the high-risk interventions that residential staff use.

**III. Challenges in Achieving Inter-Rater Reliability**

- Had there been more time, we would have recruited non-trainers to rate the videos. Non-trainers would not have had a vested interest in whether or not the trainees learned the skills, whereas potential trainer investment in the outcome may have led to biased results.
- The measuring instrument was a good checklist, but it needed more work to refine it before it could be used as a very reliable instrument.
- The interrater reliability was not as good as it could have been. Trainers were handed the instruments and asked to assess the videos. Providing them with training on how to rate trainees may have resulted in better inter-rater reliability. It would have helped to fully describe the meaning of each item measured.

# Developing Multiple Choice Tests for Social Work Trainings

Basil Qaqish, *Ph.D.*

## Abstract

This paper reviews 2 stages for development of multiple choice tests: question development and item analysis. Suggested steps are reviewed to develop and refine test questions. Common statistical trainings to analyze the performance of items are also explained. Recommendations for software and more detailed information on material analysis are found in the appendixes.

## Introduction

Multiple choice tests are best suited to test gains in cognitive knowledge domains. Testing other abilities, such as behavioral skills, may require other types of tests and evaluations. Multiple choice tests may be necessary in certain social work training environments because they can serve two important purposes:

1. Provide evidence for knowledge and/or skill acquisition.
2. Provide diagnostics for strengths and weaknesses in the teaching/learning process.

Scores on such instruments can help trainers improve the teaching/learning process and identify strengths and weaknesses of individual trainees. This process can help identify which trainees may benefit from additional training in specific areas.

In a sense, this paper can serve as a continuation to *Developing Multiple Choice Tests: Tips and Techniques,* a paper written by Richard J. McCowan and David P. Wegenast. Their research discusses issues like test instructions, examples of good items and bad items, how distracters should be presented, scoring, instructional taxonomies, validity, reliability, and pre-test/post-test analysis. The current paper concentrates on the stages of test development and gives particular attention to the psychometric

analysis on the item level as well as on the test level, using classical test theory. Because social work trainings include only a relatively small number of trainees, classical item and test analysis are appropriate.

The text development process, introduced below, is comprised of 2 steps. The first stage involves developing the items and the second involves analyzing the items after pilots.

## Stage 1: Developing and Piloting the Test

Stage 1 includes four steps:

1. Attending training;
2. Developing items for each competency;
3. Reviewing the developed items by content experts;
4. Piloting the items to improve the distracters. A distracter is one of the wrong choices in a question. For example, if a four-choice multiple choice question has only one correct answer, then the other three choices are called distracters;

### *Attending Training*

Attending training is extremely helpful for the instrument developer to attend a training session(s). This will help in understanding what the trainees go through and allow the instrument developer to assess which competencies are emphasized most during training. In addition, the test developer needs to read all manuals and documents (i.e., handouts) that pertain to the training. During this stage, the test developer can establish good contacts with the content experts (i.e., the trainers), who will prove very valuable in the instrument development process.

### *Developing Items for Each Competency*

Once all of this is established, the test developer should look at the competencies taught in the training. While consulting the training materials, the test developer should write test items for each competency in the training. These types of questions may be developed: (1) true/false or any other two-choice format; (2) multiple choice, containing three or more choices; and (3) matching, which is comprised of two columns of matched items, where the trainee has to correctly match items from each column. Figure 1 shows examples of each type of test items.

**Figure 1: Test Items by Type**

**True/False (two-choice item):**

Human development begins at birth and continues throughout life.

a) True
b) False

**Multiple-Choice question**:

Piaget's theory of child development stressed the importance of

a) Making correct moral decisions.
b) Play and friendship.
c) Learning to trust.
d) Learning by repetition in infancy.

Source: *North Carolina Child Development in Families at Risk* knowledge assessment. 2004)

**Matching (three questions):**

For items 1-3, select the ONE developmental stage that corresponds most closely to when each listed milestone normally first occurs. You may choose a developmental stage once, more than once, or not at all. Bubble in the corresponding circle on your answer sheet.

| Milestone | Developmental stage |
|---|---|
| 1. Begins pretend play. | A. Infant (1 year or younger) |
| 2. Cries, but is often soothed when held. | B. Toddler (1-3 years) |
| 3. Tracks people and objects with eyes. | C. Preschool (3-5 years) |

Source: *North Carolina Pre-Service Training Knowledge Assessment* (2003)

Once the items have been written, a table specifying competencies and which items test each competency should be prepared: Table 1 shows an example of this.

**Table 1: An Example of Item Competency Matrix**

| Competency | Test Items |
|---|---|
| 1. Recognize the physical, emotional, and behavioral indicators of placement-induced stress in the families of children being placed. | 3, 7, 12, 27 |
| 2. Understand the serious negative effects on children of changing and inconsistent living arrangements. | 2, 5, 11 |
| 3. Recognize the physical, emotional, and behavioral indicators of placement-induced stress in children of varying ages. | 4 |

Adapted from North Carolina's social work training, *The Effects of Separation and Loss on Attachment*. (2002)

Such a table helps identify which competencies have enough items and which ones need more items. There is no specific number of items required for each competency. Some competencies may be more important than others or may be stressed more during training. Some competencies cover a wider range of issues than others. One useful way to assure coverage of each competency is to initially read the training materials, marking all of the pages with important information related to the competencies. Once the first reading is done, the instrument developer rereads the material, this time writing two to four items for each marked page. After the items have been written, it is extremely important to have content experts review the item by competency matrix to make sure that there is agreement between content experts and the instrument developer regarding the breakdown of items into the different competencies.

***Reviewing of Items by Content Experts***

Once the test developer feels enough items that have been written for the assessment instrument as a whole and the underlying competencies, those items need to be sent to content experts (i.e., trainers or curriculum developers) for review. There is no minimum number of items that should be written. The total number depends on the amount of information acquisition to be assessed, the time allotted for piloting the instrument, and the size of the instrument that seems appropriate for the training.

Content experts will inform the test developer of any suggestions for language changes, additions of new items, and deletions of certain items that may be necessary to improve the instrument. This is a dynamic process that goes back and forth between the test developer and content experts. Content experts know which items are important to keep and which items are not stressed in the training, so they are the best advisors on improving the instrument. They can pinpoint weaknesses in the instrument and add their own items to improve the overall working of the test. Cooperative work between content experts and the test developer is very important in producing a final version of the instrument that works well and everyone is pleases all parties- the test developer, the content experts, and the training managers.

**Additional Factors in Test Development**

There are two additional issues that are important to recognize when developing such tests.

1) First, people are generally used to having the same number of choices for each item on the test. Such a "fixation" on the same number of choices is not necessary. One can have a combination of different numbers of choices for the items in the assessment. Existing software can handle different numbers of choices for different items in the instrument. Items with the same number of choices should be grouped together on the test.
2) Second, people tend to assume that there is only one correct answer for a multiple choice question. Such an assumption does not have to be the case for a test in social work training. For example, if one question has 4 choices, it is possible to have only one correct answer, and trainees get one point for choosing it. No points would be awarded to anyone who chooses any other choice for that question. It is also possible to have a four-choice question where *A* as the best possible answer but choice *B* also deserves some credit. In this case, answer choice *A* may get a value of 0.70, and choice *B* may get a value of 0.30. Existing software, such as Lertap 5 or Iteman, score such scenarios quickly and fairly easily

When the test developer and the content experts are in agreement regarding the items in the test, it is time to start piloting. Piloting during the first stage is primarily done for distracter

analysis and language changes. A sample size of 50 or more is needed here.

***Piloting the Items for the Purposes of Distracter Analysis and Language Changes***

After the test is administered to approximately 50 people, the item responses are analyzed through statistical descriptives. Frequency item analysis is important to help identify which distracters are working (chosen by at least one person), and which distracters are not working (not chosen by anyone). A distracter is working if some of the lower ability examinees choose it. It is not working if it is not chosen by anyone, or if it is fooling higher ability examinees while lower ability examinees are not choosing it. The distracters that are not working may need some language changes to improve them. Some distracters may need to be dropped and replaced by others. The following are three examples of distracters:

***Distractor Example 1***

| **Q11** | | |
|---|---|---|
| **Option** | **N** | **/38** |
| A | 3 | 7.9% |
| B | 1 | 2.6% |
| C | 4 | 10.5% |
| D | 30 | 78.9% |

The question in Example 1 seems to be working fine. There are 38 people who answered the question, and 30 of them chose the right answer (Choice D). One would expect that for most questions, most people will answer each question correctly. All three distracters A, B, and C were chosen by at least one examinee. This indicates that all three distracters are working. Even though one of the distracters was chosen by only one person, it is fine to keep it as is at this stage in the process. Another review of the distracters may be done later, after a few hundred examinees have taken the test. Therefore, such a question does not need any revision once the first stage pilot is done.

***Distractor Example 2***

| **Q30** | | |
|---|---|---|
| **Option** | **N** | **/38** |
| A | 2 | 5.3% |
| B | 0 | 0.0% |
| C | 1 | 2.6% |
| D | 35 | 92.1% |

For the question in example 2, 38 people answered the question, and 35 people chose the right answer (Choice D). It is a relatively easy item since 35 people out of 38 people answered it correctly. Distracters A and C are working because they were chosen by at least one examinee. However, Distracter B was not chosen by any examinee; it did not fool anyone. The job of a distracter is to fool lower ability students into thinking that it is the right answer choice for that specific question. If the distracter fails to do this, it has to be looked at for possible amendments or changes. It is also possible to get rid of this distracter completely and replace it with another. In summary, this question does need some revision and modification at the end of the first stage of the pilot.

***Distractor Example 3***

| **Q28** | | |
|---|---|---|
| **Option** | **N** | **/38** |
| A | 38 | 100.0% |
| B | 0 | 0.0% |
| C | 0 | 0.0% |

The correct choice, *A*, was chosen by all 38 examinees who took the test. This means that this question does not discriminate at all between those who know and those who do not know the information being tested. It does not discriminate between higher ability and lower ability students. Distracters *B* and *C* were not chosen by anyone. This question needs to be looked at for many reasons. It may be a very easy question, and regardless of what distracters the item has, trainees will always choose the correct answer. In this case, a question of substance arises: Should this question be kept in the test or taken out? This question is best answered by content experts. If content experts think that it is an important item, one may opt to keep it and change the distracters.

It may be a good idea, in this case, to ask content experts to write the distracters for this. They are in a position to know what kind of distracters may fool lower ability examinees.

Once the first stage of piloting is complete and all necessary changes to the questions are made, the test instrument is compiled again as a new instrument. The second stage of piloting begins here.

### Stage 2: Piloting the Items for the Purpose of Overall Analysis of the Test

The second stage of the pilot continues until at least 150 examinees have taken the final version of the instrument. Having at least 150 examinees who have taken the test will give stable estimates for the different measurement concepts that will be used in the analysis. Many variables, on the test level as well as on the item level, need to be examined. The following are the main psychometric variables the social work test developer needs to look at once all the data from the second stage of the pilot have been collected.

**Reliability**: This is a value between 0 and 1. The higher that the value is, the better the test instrument; (the test results are more reliable). Linn and Gronlund (1995) recommend a reliability value between 0.60 and 0.85 for tests similar to the ones discussed in this paper. Reliability is a measure of how "repeatable" the test results are. In classical test theory, this value is defined as the correlation between test scores on parallel forms of a test (Hambleton, Swaminathan, & Rogers, 1991). Table 2 provides an example of how reliability output values are given using the statistical software Lertap 5; (the statistical packages SPSS and JMP give similar output.)

**Table 2: Reliability Values Output**

| **alpha figures (alpha =0.6616)** | | |
|---|---|---|
| ***without*** | ***Alpha*** | ***change*** |
| Question 1 | 0.673 | **0.012** |
| Question 2 | 0.650 | -0.011 |
| Question 3 | 0.666 | **0.004** |
| Question 4 | 0.646 | -0.016 |
| Question 5 | 0.662 | **0.001** |
| Question 6 | 0.658 | -0.003 |
| Question 7 | 0.675 | **0.013** |
| Question 8 | 0.638 | -0.024 |
| Question 9 | 0.660 | -0.001 |
| Question 10 | 0.644 | -0.018 |
| Question 11 | 0.672 | **0.010** |
| Question 12 | 0.652 | -0.009 |
| Question 13 | 0.662 | **0.001** |
| Question 14 | 0.677 | **0.015** |
| Question 15 | 0.646 | -0.016 |

Cronbach's alpha, which is an expression of the average correlation values for all possible Flanagan's split halves of the test, is one of the most commonly used reliability values. All possible split halves means taking into consideration all possible combinations of dividing the test items into two halves and finding the total scores correlation values associated with the two halves (Flanagan's split half coefficient), then taking the average of all resultant correlation values. This average will be the same value as coefficient alpha. Flanagan's formula for internal consistency coefficient is provided below:

Reliability = $\frac{4S_{x1x2}}{S^2_x}$ where S and $S^2$ represent the standard deviation and the variance of the test, and $x_1$ and $x_2$ represent the two part-test scores, and x represents the full-test score ( Robert L. Linn, 1989. p.111 ).

Lertap 5 reports Cronbach's alpha value. In the table above, test reliability is 0.66. If question 1 is removed from the test, the reliability value increases by 0.012. The new reliability value, with item 1 removed is 0.673. In the above table it seems that items 1, 7, 11 and 14 are pulling the test reliability value down. Items like these are the ones that should be taken

out of the instrument. Other suspicious items like 3, 5, and 13 need to be looked at carefully to see if they can be salvaged or should be removed from the instrument.

**The Standard Error of Measurement (SE)**: This is a practical index of score precision. There are precision errors associated with any reported scores due to the fact that there are many variables involved in any individual performance on the test. Examinees may be having a good day or a bad day. They may guess correctly on many hard items or guess incorrectly on many easy items. They may have had a sleepless night before taking the test, which effected their performance on the test in a negative way. There are many variables involved that result in some inaccuracies in the reported scores. The standard error of measurement (SE) reflects these inaccuracies. For example, let's say that someone's raw score on an assessment is 31, and the SE for that test is 2. This means that a 68% confidence interval around a score of 31 is 29 to 33. A 95% confidence interval around the same score of 31 is 27 to 35. For a person who gets a score of 31 on this test, one is 68% confident that his/her "true score" is between 29 and 33. One is 95% confident that his/her "true score" is between 27 and 35. In general, a low SE value, less than 5%, is an acceptable value for diagnostic purposes for a test as a whole. If one is scoring a test on many subtest levels, for diagnostic purposes, then a SE value of 7.5% or less is realistic on the subtest level.

**Point Biserial and "z" Scores**: Tables, such as table 3 below, are used to see how every item is working in the test. Analysis of what is going on in each item is done to improve the items and delete problematic items.

**Table 3: Items Point Biserials, Biserials, Averages, and Z-Scores**

| **Question 24** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **option** | **wt.** | **n** | **p** | **Point biserial** | **biserial** | **avg.** | **z** |
| **A** | 1.00 | 32 | 0.41 | -0.14 | -0.17 | 32.41 | -0.03 |
| B | 0.00 | 15 | 0.19 | -0.16 | -0.23 | 31.13 | -0.32 |
| C | 0.00 | 32 | 0.41 | 0.15 | 0.19 | 33.34 | 0.18 |
| **Question 25** | | | | | | | |
| **option** | **wt.** | **n** | **p** | **Point biserial** | **biserial** | **avg.** | **z** |
| A | 0.00 | 3 | 0.04 | -0.22 | -0.52 | 27.67 | -1.12 |
| B | 0.00 | 7 | 0.09 | -0.37 | -0.65 | 27.43 | -1.18 |
| **C** | 1.00 | 69 | 0.87 | 0.38 | 0.60 | 33.28 | 0.17 |

Question 24 above has *A* as the correct answer. The proportion correct for that item is 0.41. It is a difficult item as only 32 people out of a total of 79 people (32+15+32 = 79) answered it correctly. The **point biserial** is the correlation value between the score on the particular item and the score on the test as a whole. For more information on the biserial and point biserial correlations, and the formulas used, see Allen & Yen, 1979. Correlation values can be between –1 and 1. One would expect a more positive value for the point biserial for those who answered the item correctly and a negative or less positive point biserial value for those who answered the item incorrectly. Question 24 is not giving good results regarding the point biserial. Choice A, which is the correct answer, has a point biserial value of –0.14, while choice C, which is a wrong choice, has a point biserial value of 0.15. The *avg* column shows the average raw test score for those who chose a certain answer on a specific question. For instance, the average raw score for those who chose choice *A* on item 24 is 32.41. Again, one would expect that those who answered an item correctly would have a higher average raw test score than those who answered the item incorrectly. This is not the case for item 24. Those who chose answer *C*, which is a wrong answer, have a higher average raw test score than those who chose *A*; the correct answer.

The last column is the z column. It transforms average raw scores of those who chose a certain choice on the item into a standardized score. One would expect that those who answered an item correctly should have a positive z score. This z value should be higher for the correct choice than the z values for any of the other choices on that item. This is not the case for item 24. The correct choice has a z value of -.0.03, while the wrong choice *C* has a z score of 0.18. It is clear that item 24 does not work well. With a relatively large sample size of 79, this item would probably be taken out of the instrument.

In contrast with question 24, question 25 works well. Most people got the answer right. The proportion correct is 0.87, so it is a relatively easy item. The point biserial for those who chose the correct answer *C* is 0.38, which is positive and higher than its counterparts for choices *A* and *B*. Those who chose answer *C* have a higher average raw score and z value than those who chose *A* or *B*. So, it seems that question 25 works well and would be kept in the final version of the instrument.

**Mastery Case Scenarios**: Sometimes there is a need to specify cut scores. Those who score above the set cut scores are considered "masters." Those who score below it are considered "nonmasters" or "others." Cut scores should be determined using a standard setting procedure, (i.e., Angoff, modified Angoff, etc.). Such a procedure can be lengthy and time consuming, and demands the collaboration of many key individuals. The reader is advised to consult suitable standard setting literature. In any case, cut score scenarios may prove very useful for certain social work trainings where there is a need to have some form of a diagnostic test that will guide decisions of who to send to which trainings. Mastery analysis results are given as part of Lertap 5 item analysis output if the software operator specifies, via specific code, a mastery scenario analysis. Table 4 shows part of a mastery analysis output; (the cut score of 70% was used).

**Table 4: Mastery Analysis Results**

| Mastery Analysis Results | n | avg. | s.d. |
|---|---|---|---|
| **Masters** | 51 | 21.9 | 1.8 |
| **Others** | 10 | 17.1 | 1.1 |
| **Everyone** | 61 | 21.1 | 2.5 |
| ========== | ======== | | |
| **Proportion Consistent Placings** | 0.847 | | |
| **Proportion Beyond Chance** | 0.346 | | |

The above summary group statistics gives the number of masters and nonmasters (others). The sum of the two groups (61) is the total number of people who took the test. The average raw scores and standard deviations are also given for each group and for everyone who took the test. The proportion of consistent placing is an estimate of the proportion of people who have been correctly classified as masters or others. In Table 4, approximately 84.7% of examinees have been correctly classified as masters or others. The other side of the story is that about 15.3% of examinees (100% - 84.7%) may have been misclassified. The proportion beyond chance is an estimate of how accurate our classification has been, above and beyond what one might expect by chance alone, (i.e., if we used a coin toss to decide whether each examinee is a master or other).

**Assembling the Final "Form" of the Test:** Item difficulty, as explained here, should be a major factor in guiding the instrument developer on how to order the items in the final version of the instrument. One of the main indicators of the difficulty of an item is its proportion correct. For instance if there are 2 items, and 70% of trainees got the first item right while only 40% of trainees got the second item right, then this indicates that the first item is easier than the second. As an example, Table 5 shows the proportion correct for a five-item instrument.

**Table 5: Item Proportion Correct**

| Item | Proportion correct |
|---|---|
| 1 | 0.30 |
| 2 | 0.40 |
| 3 | 0.70 |
| 4 | 0.50 |
| 5 | 0.60 |

Item 3 is the easiest item; 70% of trainees got it right. Item 1 is the hardest; only 30% of trainees got it right. When assembling the final version of the instrument, easier items should come first and harder items should come last. The final item order version of the above test is shown in Table 6:

**Table 6: Reordered Items Based on Item Difficulty Analysis**

| Item | Proportion correct |
|---|---|
| 3 | 0.7 |
| 5 | 0.6 |
| 4 | 0.5 |
| 2 | 0.4 |
| 1 | 0.3 |

Ordering the items from the easiest to the hardest is natural and fits the human psyche. Imagine you are taking a test. You start trying to answer the first question and find it to be very hard. It is a natural reaction to say to oneself, "I won't be able to complete this test. I can't even get the first question right." On the other hand,

imagine you are taking a test and find the first couple of questions to be fairly easy. Then, it is a natural reaction to say to oneself, "This is good. I know the answers. I can do this test successfully. "

## Reporting Scores

The simplest, clearest, and most direct method of reporting trainees' scores is to report the percentage correct value that examinees achieve. Diagnostic scores on the subtest level can also be reported as percentages in the score report.

It is worth mentioning that reporting percentages is just one method of reporting scores. One can choose any average score and any standard deviation and report scores accordingly. This matter can be clarified via an example: A person's raw score is 30 on an instrument that has a mean score of 36 and a standard deviation of 3. If one wanted to transform this person's score to a score with an instrument mean of 50 and a standard deviation of 5, a linear transformation is applied. The result of this transformation is the following formula:

New Score = 5 * ((x-36)/3) + 50 = 40 where x is the person's raw score.

Thus, the reported score will be 40 for a person with a raw score of 30 on the test for more information on transforming scores, see Allen and Yen, 1979.

## Appendix A

### Useful Software for Test Development in Social Work Training

**Lertap 5**: This is a classical item analysis software. It is Excel based and easy to operate. The user will need to read chapter 2 in the manual to be able to start using the software. The user will also need to learn how to write simple code to make the software do specific item analysis requests. This software is also useful in analyzing surveys. Most of the output used in this paper came from Lertap 5 output.

**Iteman**: This is another classical item analysis software. Using this software, instead of Lertap 5 is up to the test developer. The output is given in flat text files.

**Statistical Package for Social Sciences (SPSS)**: This is a general statistical analysis software that can be used to obtain most of the analysis results that mentioned above.

**JMP:** This is a statistical software developed by Statistical Analysis System (SAS). It is menu driven and can perform general statistical analysis, with an added bonus: beginning in 2005, JMP can do item analysis using item response theory. It can do one-, two-, and three-parameter logistic model analysis. The program will output ability estimates (thetas), and item difficulties, item discriminations. It will also output item characteristic curves and item information functions.

The following web address contains many testing and assessment software that can be purchased: http://www.assess.com

## Appendix B

## Formula Background for Statistical and Measurement Concepts

**Reliability**: The formula used to calculate coefficient alpha is:

$$\alpha = \left[\frac{N}{N-1}\right]\left[\frac{\sigma_X^2 - \sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right]$$

where $\alpha$ equals coefficient alpha, $N$ is the number of items in a test, $X$ equals the obtained score on a test by summing all individual ($Y_i$) items on the test, $\sigma^2_X$ is the variance of test scores, and $\sigma^2_Y$ is the variance of an individual test item (Allen & Yen, 1979).

It is worth mentioning that one concept of reliability (Spearman-Brown formula) can be used to estimate how many more items need to be added to a test to increase the reliability of the test to a certain value. For example, a 20-item test that has a reliability of 0.6 and wants to see a reliability value of 0.7 for that test. The question is: How many more items need to be added to this test to achieve the desired reliability value of 0.7? The problem is solved through the following computation (N represents the number of lengthening times for the test):

$$N = \frac{(0.7) * (1 - 0.6)}{(0.6) * (1 - 0.7)} = 1.56$$

One must multiply 1.56 * the number of items on the test = 1.56 * 20 = 31.2 items.

One should round up to 32 items needed on that test to raise the reliability to 0.70. This requires adding at least 12 more items. Off course, if one is piloting new items, one should probably pilot 20 items to be able to end up with 12 useful items to add to that test. For more on this topic, see Allen and Yen (1979).

**Standard Error of Measurement:** The formula for the standard error of measurement is:

SEM = s.d. $\sqrt{1 - Alpha}$

Where s.d. is the standard deviation of the test scores and Alpha Cronbach's alpha reliability value.

**The Point-Biserial Correlation Coefficient**: The following example clarifies how this index of item discrimination works in a test. *See Table 7.*

There are eight examinees. The first four examinees have higher scores than the last four, and got the first item correct. The last four examinees got the first item wrong. Those four examinees also have lower test score. Thus, one expects the first item to have a positive point biserial since those who got it right have higher total test scores too. Everyone got item 2 correct; therefore, that item has no discrimination power. One expects its point biserial value to be zero. Item 3 should have a negative value for the point biserial, since only the lower ability people got it right. Some of the expressions used in Table 7 are defined below:

Total Test mean is the mean total raw score on the test.

St is the standard deviation of the total test raw scores.

M*p* is the mean total raw score for those who got a certain item correct.

M*q* is the mean total raw score for those who got a certain item false.

P is the proportion correct on an item.

q is the proportion incorrect on an item.

$\mathbf{r_{pbi}}$ is the point-biserial value

**Table 7: Clarifying the Point-Biserial Correlation**

| Examine ID | Question 1 | Question 2 | Question 3 | | Total Test Score |
|---|---|---|---|---|---|
| **1** | 1 | 1 | 0 | | 30 |
| **2** | 1 | 1 | 0 | | 25 |
| **3** | 1 | 1 | 0 | | 25 |
| **4** | 1 | 1 | 0 | | 20 |
| **5** | 0 | 1 | 1 | | 15 |
| **6** | 0 | 1 | 1 | | 10 |
| **7** | 0 | 1 | 1 | | 10 |
| **8** | 0 | 1 | 1 | | 5 |
| | | | | **Total Test Mean** | **17.5** |
| **Mean Correct (*Mp*)** | 25 | 17.5 | 10 | **St** | **8.86** |
| **Mean False (*Mq*)** | 10 | 0 | 25 | | |
| **Proportion Correct (*P*)** | 0.5 | 1 | 0.5 | | |
| **Proportion incorrect (*q*)** | 0.5 | 0 | 0.5 | | |
| **Point-biserial ($r_{pbi}$)** | 0.85 | 0.0 | -0.85 | | |

The point-biserial formula is:

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq}$$

For question 1, the point biserial is calculated as follows:

$\mathbf{r_{pbi}} = \frac{25-10}{8.86} * \sqrt{0.5 * 0.5} = 0.85$

For question 2, the point biserial is calculated as follows:

$\mathbf{r_{pbi}} = \frac{17.5-0}{8.86} * \sqrt{1 * 0} = 0$

For question 3, the point biserial is calculated as follows:

$\mathbf{r_{pbi}} = \frac{10-25}{8.86} * \sqrt{0.5 * 0.5} = -0.85$

In conclusion, the item that discriminates well between higher ability examinees and lower ability examinees will have a positive discrimination (point-biserial) value. If the item has a negative point biserial value, this indicates that lower ability examinees got the item correct while higher ability examinees got it wrong. Such an item is faulty and is in need of further analysis to see if it should be taken out of the test. An item that has a zero point-biserial value fails to discriminate amongst examinees and was either answered correctly by all examinees or answered falsely by all examinees. In either case, it is not a good item if left in the test as is. The test developer will need to investigate such an item further to see if item amendments are possible or if the item needs to be taken out of the test.

## References

Allen, M.J., & Yen, W.M. (1979). *Introduction to Measurement Theory*. Belmont, California: Wadsworth.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational Measurement*. New York: Macmillan.

Hambleton, R.K., Swaminathan, H., & Rogers, H. Jane. (1991). *Fundamentals of Item Response Theory*. Newbury Park, California: SAGE.

Linn, R.L., & Gronlund, N.E. (1995). Measurement and Assessment in Teaching (7th edition). Englewood Cliffs, NJ: Prentice-Hall.

Nelson, L.R. (2001*). Item Analysis for Tests and Surveys Using lertap 5.* Perth: Australia: Curtin University of Technology.

## Acknowledgments

I would like to thank professors Dale Curry of Kent State University and Betsy Lindsey of the University of North Carolina at Greensboro for reviewing an earlier version of this paper and giving me valuable advice that helped enhance it.

## Discussion

*Theme:* Technical Assistance Forum
*Title:* Developing Multiple Choice Tests for Social Work Trainings
*Presenter:* Basil Qaqish, *Ph.D.*

---

Basil Qaqish presented his work on developing five different multiple choice tests that assess the knowledge of newly hired North Carolina child welfare workers. He described the stages of test development, including how to pilot the test items, and the best ways to analyze the data. The discussion covered technical, theoretical and ethical issues in developing and applying testing instruments in social work.

### Topics of Discussion

**I. Point Biserial Correlations Measuring One or More Than One Construct**

- Point biserial coefficients are representative of item total correlation.
- Pont biserial values are sample dependent.
- The point biserial coefficient tries to answer the question: How did the people who selected an item option do on the criterion score measure? If they did well on the criterion measure, the point biserial value will be high (closer to 1.0). In social science, "high" may be taken as any value above 0.30 or 0.40.
- The point biserial is an index of item discrimination.

**II. Item Response Theory Versus Classical Test Theory in Scoring**

- Item response theory is better than classical test theory, particularly if one is working with large numbers.
- In classical test theory, one reports the percentage correct without taking into consideration how hard or easy the items are.

- In item response theory, ability estimates depend on the difficulty of the question. The ability estimate of someone who answers a difficult question will be different than the ability estimate for someone who answers an easy question.
- The items that participants miss and those they answer correctly are very important. In item response theory, the score, or ability estimate, is an interaction between the items and the examinees.

### III. Involving Trainers in Question Development

- Even if the trainers have the test questions, it does not mean that they will train to the questions rather than to the curriculum, thereby creating biased results. For example, the knowledge assessment test in North Carolina includes 181 questions. With such a large number of items, in addition to what they must cover in their training, it would be very challenging to train based on items on the test.
- Even if evaluators provide assurances, and are conscientious and professional, the may be a perceived (or real) bias if the evaluations are not independent from the trainers. Using independent evaluators provides assurance that the evaluation is objective, and the trainers have not "trained to the test."
- Testing companies use experienced freelance teachers and pay them by item to do item banking. While they develop items, they pilot others all the time to create a wealth of items with different difficulty and discrimination parameters. The number of test questions in social work training is limited. The resources that testing companies have to develop questions are not generally available to those who develop tests for social work trainings.
- In social work, a trainer may teach to the test questions because the test items tap into important information and important constructs. This is fine. One wouldn't want to teach something totally different. How much of an issue that is depends on the purpose of the evaluation. If the purpose is maximizing instruction, it should not be an issue.

### IV. High Stakes Testing

- In North Carolina, testing cognitive domains of knowledge in social work trainings is not considered "high stakes" testing. The responses are used strictly for diagnostic purposes to give feedback to supervisors because they ask for this information. There are no set cut scores. There is no such thing as a passing or failing score. If this type of evaluation were to move to a place of high stakes testing, there would need to be more rigorous criteria.
- If there is high stakes testing, unscrupulous things can happen. For example, one state had a situation where their training evaluation pre- and post-tests were the same. The pre-test had to be changed because it appeared that unethical behavior was taking place. The post-test was kept. The post-test scores fell such that 40% of the trainees failed who had previously passed. There were such high stakes riding on this test that people were memorizing the pre-test, even taking pictures of the test with their cell phones, but they were not studying beyond the material that was on the pre-test.

### V. Using Alternate Forms in Testing

- One can also use alternate forms of the test to avoid the issue of training to the test.
- However, having alternate forms of the test is difficult, if not impossible, if there are hardly enough questions in its original form and the items are placed in order from easiest to hardest. One must really develop an item bank, then assess and pilot the items to ensure that items among the different tests match regarding their content and difficulty. Afterwards, one will have to do test equating, to make the two test forms comparable, before determining the scores because it would not be fair to use the percentage correct.
- For trainings with small sample sizes and a limited number of items, it is best to have one good test form with accurate measures regarding reliability and standard error that covers the whole domain.
- Testing and assessment software can be found at www.assess.com. This web address is a rich source of testing and test development software.

# Strength-Based Family-Centered Training Evaluation in the Los Angeles County Department of Children's and Family Services

Todd Franke, *Ph.D.*, Walter Furman, *M. Phil*., and William Donnelly, *LC..W, M.P.A,* .
Department of Social Welfare, University of California, Los Angeles and The Inter-University Consortium

## Abstract

In 2004, the Los Angeles County Department of Children's and Family Services (DCFS) and the Inter-University Consortium IUC) developed and delivered a one day training on Strength-Based Family-Centered Practice (SBFC) in Child Welfare Services. SBFC was deemed an important tool for all line staff of DCFS to utilize in their daily interactions with clients, as it promised to help the Agency meet its goals of ensuring the safety of children, strengthening families, and preventing detention.

Over the past year, the IUC and the current authors implemented several methods to assess the impact of the SBFC training, including surveying participant reaction, testing knowledge and evaluating transfer of learning (TOL). Supervisors and line workers were surveyed as part of a pilot study of TOL. The paper reviews the results of the pilot study, which show a reasonably thorough adoption of SBFC practices. It also discusses the challenges of communicating about such an evaluation in a large system, and the difficulties in obtaining a sufficient response rate.

## Introduction

In 2004, the Los Angeles County Department of Children's and Family Services (DCFS) and the Inter-University Consortium (IUC) developed and delivered a one-day training on Strength-Based Family-Centered Practice (SBFC) in Child Welfare Services. The principal designers of the training were consultants Rose Wentz and Nora Gerber, with significant input on content and methods from staff from the DCFS training unit and the IUC.(SBFC) practice was deemed an important tool for all line staff of DCFS to utilize in their daily interactions with clients, as it promised to help the Agency meet its goals of ensuring the safety of children, strengthening families, and preventing detention.

Over the past year, the IUC and the current authors implemented several methods to assess the impact of the SBFC training. A December 2004 report transmitted to DCFS noted that:

- The SBFC one-day training was delivered in 59 presentations to 1,526 staff, or 58.7% of DCFS' approximately 2,600 Children's Social Workers (CSWs).
- There are participant reaction evaluations for 815 staff; the ratings range from 4.23 to 4.52 (Scale: 1 = Very Poor; 5 = Very Good). Staff rated the training well across the twelve focused and general domains, as shown in Table 1.

**Table 1: Participant Reactions—Focused Results**

| Item | Average |
|---|---|
| Organization | 4.45 |
| Personal Learning | 4.23 |
| Applicability | 4.4 |
| Instructor Responsiveness | 4.52 |
| Overall | 4.47 |
| SBFCP Use Language | 4.33 |
| SBFCP Value Contributions | 4.36 |
| SBFCP Understand Choices | 4.33 |
| SBFCP Have Participation | 4.31 |
| SBFCP Write Goals | 4.27 |
| SBFCP Engage Family | 4.29 |
| SBFCP Use Tools | 4.23 |

- Knowledge assessments (24 multiple-choice items) were administered to 528 staff at the conclusion of training; pre-tests were administered to 70 staff. The average score on the pre-test was 58.3% correct responses, and the average score on the post-test was 72.1%.

This paper pertains to a pilot test of evaluating the transfer of learning from training to practice. Using an assessment survey tool based upon practice elements that was designed by the consultants, the IUC attempted to survey CSWs and SCSWs (Supervisors) in one DCFS office, using an Internet-based survey approach. The twin goals of the pilot were: 1) to learn if this approach to studying transfer of learning in the DCFS environment could be successful, and 2) to ascertain, in this pilot, the extent of adoption of SBFC practice by CSWs and SCSWs, and to learn what factors are perceived as promoting and inhibiting the use of the practice.

## Study Methods

### *Instruments*

Two surveys were developed for this study—one for CSWs and one for SCSWs. The CSW survey includes self-rating of practice in 19 areas that were suggested by Gerber and Wentz as reflecting adoption in daily practice of the tenets of the SBFC approach. Table 2 shows these nineteen social work practices that are at the core of the training.

**Table 2: 19 Strength Based CSW Practices in Survey**

| Item # | Description of Strength Based Practice |
|---|---|
| 1 | I spend at least one hour observing the family in a home environment to identify strengths as well as deficits. |
| 2 | I give clients specific strength-based feedback, using strength-based assessment and decision making tools. |
| 3 | I demonstrate confidence in the ability of family members/caregivers to make good choices and let them do the job in the way they think it should be done |
| 4 | When mistakes occur, I avoid blame and instead work with family members/caregivers to find ways to prevent similar problems in the future. |
| 5 | I acknowledge family members (verbally and in writing, i.e. in case plans, court orders) who are doing good work. |

| | |
|---|---|
| **6** | **I state clear expectations for families in strength-based language.** |
| 7 | I back the family members/caregivers when I think they are right and help them resolve problems. |
| 8 | If a family member/caregiver does not meet expectations, I meet with that person in private to find out what happened, provide honest feedback and to prevent unacceptable behavior from becoming the norm. |
| 9 | I ask family members/caregivers for ideas about how to improve the service/case plan. |
| 10 | I consider "family" to be defined as blood and legal ties, PLUS any others the family considers an essential resource. |
| 11 | I ask for family/caregiver feedback to learn how good my service is and how it could improve |
| 12 | I work to create a positive environment for everyone involved in a case. |
| 13 | I work with family members to develop case plans and outcomes. |
| 14 | I celebrate with family members/caregivers when they meet case outcomes or goals. |
| 15 | My recommendations to court and at case planning meetings recognize the successes of family members. |
| 16 | I communicate with out of home providers (foster care or residential) regarding case plan objectives. |
| 17 | I take time to hear and know the family's "stories" (their accomplishments, attributes, skills, goals, values, etc.) to better understand their unique family culture and strengths. |
| 18 | I write up case plans in a family-friendly, behaviorally specific way: each item reflects the positive outcome (change) that is expected/desired. |
| 19 | When I speak with family members face to face, on the phone or in writing, I use terms that are comprehensible (no jargon), positive in tone and promote respect, rapport and trust. |

In addition to the 19 SBFC practices identified above, the survey queried the CWSs about certain aspects of their views of Strength-Based practice, the organizational support they get for its use, and the barriers they perceive to using it. The survey also

asked if they believed further training could assist them in effectively using SBFC practice. Table 3 presents those items.

**Table 3: CSW Survey Items on Views of SBFC Practice, Organizational Support, and Training**

| Item and Area | Content |
|---|---|
| Views of Strength Based Practice | |
| 1 | The SBFC approach is an important new technique for improving child welfare practice and outcomes |
| 2 | The SBFC approach just won't work with involuntary clients |
| Views of Organizational Supports/Barriers | |
| 3 | My supervisor is able to help me use SBFC techniques |
| 4 | There is just not enough time in the day to do what the SBFC approach would require |
| 5 | The computer systems (e.g. CWS) do not let me document my SBFC work. |
| 6 | DCFS now supports SBFC practice at all levels of the organization. |
| 7 | My supervisor supports my use of the SBFC approach. |
| Views on Training | |
| 8 | More training could help me use SBFC effectively |

The supervisors' survey instrument mirrored, to a large extent, the practice elements and views on the CSW survey. In the interest of time, and hoping to get as high a response rate as possible, supervisors were asked to rate the SBFC practice of their CSWs in general, and 10 of the 19 practice elements from the CSW survey were selected for supervisors to rate "their" CSWs. Each of those items was re-written and stated in terms such as, "My CSWs state clear expectations for families in strength-based language." The five items from the table above on views of organizational supports/barriers were repeated with SCSWs, as were the two

items on views of Strength-Based practice, and the item about desiring more training. A new section pertaining to supervision of Strength-Based practice was added, with the questions as shown in Table 4, below.

**Table 4: Survey Items on Strength Based Supervision**

| Item # | Content |
|---|---|
| 1 | I feel full prepared to supervise CSWs in SBFC practice |
| 2 | I am able to make effective suggestions about how CSWs can engage families |
| 3 | I have thorough knowledge of my CSWs SBFC practice from observation, supervision and documentation |
| 4 | I am able to use positive, strength based feedback with my CSWs. |
| 5 | I fully understand and value culture-specific approaches used by my CSWs. |

***Survey Methods***

As noted, this survey was an Internet survey administered at one cooperating DCFS office. DCFS administration approved the survey design and instruments. The Regional Administrator (RA) was both cooperative and helpful, picking the time when the survey should be sent to staff, providing lists of e-mail addresses to the IUC evaluators, and notifying and urging staff participation in the project.

The initial invitation to participate, from the Regional Administrator (RA) and Division Director, was e-mailed to 20 SCSWs and 119 CSWs on February 22, 2005. Two e-mail reminders from the RA were sent to those who had not responded, and the survey closed on March 19, 2005. The survey was designed and submitted to staff using the services of the Zoomerang online survey company. The study asks about aspects of SBFC practice, but no inference can be drawn that the observed practice resulted from the training. Assessing the causal link between training and practice would require a more elaborate study design.

### *Focus Group on Survey Method and Results*

Several weeks after the survey closed, one of the investigators held an informal feedback meeting with four staff from the office to query them about the effectiveness of the survey methods.

## Results

### *Results Concerning the Survey Method*

The CSW survey yielded 39 complete responses from 121 e-mail addresses supplied by the RA, a 32% completion rate. Of the CSWs who responded to the survey, 38 (or 97%) reported having attended the Strength-Based training session. In terms of job title, 14, or 36%, were classified as CSW-IIs, 21 (54%) as CSW III, and the other 10% in other job titles. The CSW respondents were generally experienced workers, with 59% reporting 6 or more years of experience, 31% with 3-5 years experience, and only 10% with less than two years experience.

The SCSW survey yielded 14 completed responses from the 20 e-mail contacts, a 70% completion rate. Twelve of these fourteen (86%) said they had attended the SBFC training. The supervisors were a highly experienced group of staff—all had over 6 years experience, and 86% had over 10 or more years experience with DCFS or in Child Protective Services.

The CSW response rate is problematic. The number of respondents is enough to present descriptive results, but there is no way of knowing if those who chose to respond to the survey have the same approach to SBFC practice as those who did not respond. In general, one suspects that voluntary respondents to such a survey might indeed have somewhat different attitudes and practice, but there is no certainty in this regard. Thus, the generalizability of the results is limited.

Three CSWs e-mailed the survey coordinator that they had a problem with accessing the Internet and submitting the survey. For one respondent, this was fixed by using an external e-mail address. Two others were unable to participate. A further problem was noted in retrospect by the Regional Administrator, who stated in an e-mail:

> "The months of February and March were peak referral and detention months for our office. As a result, my social workers were not as available as we hoped."

In the midst of the survey, the survey coordinators suggested to IUC and DCFS administration that some sort of extra incentive for participation was required, and the idea of a lottery with a prize was put forward. The administrators felt this approach was problematic, and no incentive was offered.

The focus group with regional staff after the survey closed yielded valuable information about the survey process. The deputy regional office managers stated that the communication within the office about the survey had been poor; they had not been briefed on the purpose and methods of the survey. One supervisor decided that the survey was not for her group, and another stated that when she learned it was voluntary, she did not follow up with her workers. Apparently, the manager's active involvement and communication with the evaluators was insufficient to assure adequate communication up and down the hierarchy within the office.

### *CSW Strength-Based Practice in the Pilot Office*

Table 5 *(see page 123)* presents the basic results of how CSWs rated their own Strength-Based practice. The first column shows the percent of the 39 respondents who said they *always* or *usually* do what the item indicates. The order of items in the table is ranked from those that are done most frequently at the top, to those that the CSWs report doing least frequently at the bottom. The second column, which is discussed later, shows how the 14 Supervisors rated their CSWs on SBFC practice, showing the percent of SCSWs that said their CSWs *always* or *usually* do what the item indicates. Note that the SCSWs were asked only 10 of the 19 items, therefore the N/A (not available) notation is used meaning no data from supervisors on this practice by CSWs is available.

The CSW respondents generally report doing most of the SBFC practice items with clients *always* or *usually*. Over 90% of CSWs say they always or usually talk to families without jargon and in positive terms, that they back family members when they think they are right, take time to hear their stories, and work with everyone to create a positive environment. Between 80% to 90% say that in their own practice they always or usually recognize the success of family members in court recommendations, acknowledge family members doing good work, state clear expectations in strength-based language, consider family to be blood ties plus other resources, meet in private with a family

**Table 5: Social Workers' and Supervisors' Views of CSW SBFC Practices, Percent Reporting *"Always" and "Usually"***

| ITEM | CSW Responses % "*Always*" & "*Usually*" N = 39 | SCSW Responses % "*Always*" & "*Usually*" N = 14 |
|---|---|---|
| When I speak with family members face to face, on the phone or in writing, I use terms that are comprehensible (no jargon), positive in tone and promote rapport and trust. | 98 | 78 |
| I back family members/caregivers when I think they are right and help them resolve problems. | 98 | NA |
| I take time to hear and know the family's "stories" (their accomplishments, skills, values, etc.) to better understand their unique family culture and strengths. | 95 | 92 |
| I work to create a positive environment for everyone involved in a case. | 93 | NA |
| My recommendations to court and at case planning meetings recognize the successes of family members. | 89 | NA |
| I acknowledge family members (verbally and in writing, i.e. in case plans, court orders) who are doing good work. | 87 | 78 |
| I state clear expectations for families in strength-based language. | 87 | 50 |
| I consider "family" to be defined as blood and legal ties, PLUS any others the family considers an essential resource. | 86 | 85 |

| | | |
|---|---|---|
| **If a family member/caregiver does not meet expectations, I meet with that person in private to find out what happened and provide honest feedback to prevent unacceptable behavior from becoming the norm.** | **85** | **NA** |
| I work with family members to develop case plans and outcomes. | 85 | 64 |
| I write up case plans in a family-friendly, behaviorally specific way: each item reflects the positive outcome (change) that is expected/desired. | 79 | NA |
| I demonstrate confidence in the ability of family members/caregivers to make good choices and let them do the job in the way they think it should be done. | 77 | NA |
| I communicate with out of home providers (foster care or residential) regarding case plan objectives. | 77 | NA |
| When mistakes occur, I avoid blame and instead work with family members/caregivers to find ways to prevent similar problems in the future. | 72 | 71 |
| I ask family members/caregivers for ideas about how to improve the service/case plan. | 64 | NA |
| I spend at least one hour observing the family in a home environment to identify strengths as well as deficits. | 54 | NA |
| I give clients specific strength-based feedback, using strength-based assessment and decision making tools. | 54 | 71 |
| I celebrate with family members/caregivers when they meet case outcomes or goals. | 47 | 43 |
| I ask for family/caregiver feedback to learn how good my service is and how it could improve. | 31 | 21 |

member not meeting expectations, and work with family members to develop case plans and outcomes.

There are a few items that the CSW respondents report doing less frequently than those at the top of the list. Asking for feedback from clients—whether on the service plan or on their own services—are among the SBFC practices reported as being used less frequently. The other three items done least frequently are spending an hour observing the family at home, giving strength based feedback using the assessment tools, and celebrating with family members when they reach case plan goals.

Table 5 suggests, that on 9 of the 10 items rated by both supervisors and CSWs, supervisors see less Strength-Based practice than CSWs self-report. Supervisors and CSWs more or less agree that CSWs always or usually take time to hear family stories, consider family to be broadly defined, and they also have similar ratings (about 70%) for CSWs working with family members to avoid blame and finds way to avoid similar problems in the future. CSWs and Supervisors both rate the practices of celebrating with families, and asking for feedback on services, as not usually occurring. Two items pertaining to case planning show a difference in CSW and SCSW perceptions. Supervisors, as compared to the self-ratings of CSWs, do not as frequently see CSWs stating clear expectations for families in Strength-Based language, or as frequently working with family members to develop case plan outcomes. However, supervisors see their social workers as more often giving Strength-Based feedback to clients than do the CSWs themselves. Despite these differences in perceptions of SBFC practice, the data in Table 5 show that CSWs and SCSWs share a common view of how SBFC practice is being implemented. For the 10 items rated by both CSWs and SCSWs, there is a statistically significant correlation of .76 ($p<.05$).

### *Views of the SBFC Approach*

The survey asked two questions about how respondents viewed the Strength Based approach, the assumption being that despite training, it was possible that staff might not fully endorse the SBFC precepts, and that this lack of faith in the approach might, in turn, affect adoption of SBFC practices. Table 6 *(see page 126)* shows that both CSW respondents and SCSW respondents endorse the approach in the main, with 72% agreeing that the SBFC

**Table 6: Social Workers' and Supervisors' Views of the Strength-Based Approach, Percent Reporting *"Strongly Agree" or "Agree"***

| ITEM | CSW Responses % "*Strongly Agree" or "Agree*" N = 39 | SCSW Responses % "*Strongly Agree" or "Agree*" N = 14 |
|---|---|---|
| The SBFC approach is an important new technique for improving child welfare practice and outcomes. | 72 | 72 |
| The SBFC approach just won't work with involuntary clients. | 19 | 21 |

approach is an important innovation for improving practice and outcomes, and approximately 80% of both groups disagreeing that the approach is not appropriate for involuntary clients.

Table 7 *(see page 128)* reports on respondents' views of organizational features that were deemed potentially important in helping or hindering SBFC practice on the line. Significant proportions of CSW respondents do perceive barriers to adoption of SBFC practice. Only about half "agree" or "strongly agree" that their supervisor supports their use of the SBFC approach, and about half agree that there is not enough time in the day to do what the approach requires. While only a quarter say (strongly agree or agree) that the computer systems do not let them document their SBFC work with clients, only about one quarter agree that SBFC practice is now supported by DCFS at all levels of the organization.

The SCSW responses to the organizational barriers are less extreme than those of the CSW respondents, but still recognize such barriers in sizable proportion. Supervisors may be less likely than CSWs to see time pressure as a barrier to SBFC practice, but still about one-third tend to agree that this is always or usually a barrier. And, like the CSWs, only a minority of the SCSW respondents sees organizational support for SBFC practice at all levels of DCFS.

Table 8 *(see page 129)* suggests that both the Social Workers and Supervisors who answered the survey think that more training in SBFC would be useful.

Table 9 *(see page 129)* shows that supervisors are quite comfortable about some areas of SBFC supervision but they feel less confident about others. Half of the responding supervisors felt fully prepared to supervise SBFC practice. Also of interest is the recognition by a majority of supervisors that they do not always or usually have thorough knowledge of their CSWs SBFC practice.

## Discussion

The pilot test of the Internet-based method of assessing SBFC practice has provided a useful test of the survey method, and offers a glimpse into the kinds of findings that Transfer of Learning Evaluation can offer. The cooperation of the evaluators and administrators was exemplary and the web based software and

**Table 7: Social Workers' and Supervisors' Views of Organizational Support And Barriers, Percent Reporting *"Strongly Agree" or "Agree"***

| ITEM | CSW Responses % *"Strongly Agree" or "Agree"* N = 39 | SCSW Responses % *"Strongly Agree" or "Agree"* N = 14 |
|---|---|---|
| My supervisor is able to help me use SBFC techniques. | 64 | NA |
| My supervisor supports my use of the SBFC approach. | 51 | NA |
| There is just not enough time in the day to do what SBFC approach would require. | 46 | 36 |
| DCFS now supports SBFC practice at all levels of the organization. | 26 | 43 |
| The computer systems (e.g. CWS) do not let me document my SBFC work. | 23 | 28 |

**Table 8: Social Workers' and Supervisors' Views of SBFC Training Need, Percent Reporting *"Strongly Agree" or "Agree"***

| ITEM | CSW Responses % *"Strongly Agree" or "Agree"* N = 39 | SCSW Responses % *"Strongly Agree" or "Agree"* N = 14 |
|---|---|---|
| More training could help me use SBFC effectively. | 66 | 72 |

**Table 9: Supervisors' Views of Their SBFC Supervision, Percent Reporting *"Always" or "Usually "***

| ITEM | Percent *"Always"* & *"Usually"* |
|---|---|
| I am able to use positive, strength based feedback with my CSWs. | 85 |
| I fully understand and value culture-specific approaches used by my CSWs. | 85 |
| I am able to make effective suggestions about how CSWs can engage families. | 57 |
| I feel fully prepared to supervise CSWs in SBFC practice. | 50 |
| I have thorough knowledge of my CSWs' SBFC practice from observation, supervision and documentation. | 43 |

services proved to be efficient and inexpensive. With just a few exceptions, it was possible to communicate with staff via e-mail. However, the CSW response rate was low enough to inhibit generalizing from the results, even within one office. While office administration reports some unusual caseload circumstances contributing to the low response rate, operational contingencies often occur in the DCFS/CPS environment, so new ideas need to be tested on how to achieve greater response rates. A focus group on the survey process revealed communication gaps in the office.

The results show reasonably thorough adoption of SBFC practices as reported by the responding CSWs. Areas where improvement could be shown involve getting feedback from clients on agency and CSW services, and spending time with families to observe, celebrate, and give specific strength-based feedback. Supervisors also think that CSW practice embraces many aspects of SBFC practice but they in general see less SBFC practice than the CSWs report. In particular, supervisors see less involvement by CSWs in strength-based case planning with families.

Both groups of respondents endorse SBFC practice as useful and appropriate with their caseloads but both also perceive some organizational barriers to implementation. Supervisors express some hesitancy about knowing the details of their CSWs practice and about being able to be fully effective in SBFC casework supervision. Neither group perceives DCFS as supporting adoption of SBFC practice at all levels of the organization. About half of the CSWs express some reservation about their supervisor's support for the use of SBFC practice. Further training in SBFC practice is deemed useful by strong majorities of CSW and SCSW respondents.

## Conclusion

### *Recommendations*

The evaluators suggested the following preliminary recommendations to Agency management from the experience and results of this pilot test.

- Survey other offices to allow comparison of results, and confirmation and refinement of findings.
- Incorporate incentives for survey completion in future surveys.

- Management should consider certain initiatives based on preliminary findings:
  1) Provision of advanced training to Supervisors on SBFC supervision.
  2) Stressing a) the utility of client feedback to workers as part of SBFC practice, and b) SBFC case planning methods, both in training and routine operations.
  3) Assess Agency support for SBFC practice at all levels, with communication to line staff about corrective actions undertaken.

## Discussion

*Theme:* Issues of Practice Assessment in a Large System
*Title:* Mainstream Training Evaluation: Multi-Levels, Multi-Benefits
*Presenters:* Bill Donnelly, *LCSW, M.P.A.,* Todd Franke, *Ph.D.,* and Walter Furman, *M.Phil.*
*Facilitator:* Sherrill Clark, *LCSW, Ph.D.*

---

Bill Donnelly, Todd Franke, and Walter Furman used their Strength-Based Family-Centered (SBFC) Training Evaluation at the Los Angeles (LA) County Department of Children's and Family Services (DCFS) as a platform from which to discuss the many aspects of conducting practice assessment in a large training system. The discussion that followed their presentation included diverse topic areas such as identifying and using internal advocates in large systems, and soliciting feedback from clients and families.

### Topics of Discussion

**I. Use of Self-Reporting in Surveys**

- Most researchers are leery of using self-report when the topics can cause cognitive dissonance—for example, with family-centered practice, fairness and equity, and cultural competence. These topics are more vulnerable to people not answering than what one would find with another method.
- Wherever there is potential for much cognitive dissonance, one needs to be careful about what they ask participants to self-report.
- The method of asking supervisors about workers' practice, as a sort of check (and vice versa), may provide further support for self-reported data.
- Another way to try and collect less biased self-reported data is to administer an anonymous survey, where workers report on feedback they received from clients in various areas.

## II. Internal Advocates—Who Are They and How Do You Use Them?

- Identify those administrators who have been in the system for a long time and with whom regional administrators have worked.
- Los Angeles's success depended upon managers' and administrators' interest in the effectiveness of training in promoting practice fidelity. They have come to see this as something that is important for their own organization, practice, and outcomes.
- One strategy is to hold a training evaluation summit to educate and orient management to the value of training evaluation and evaluating practice fidelity. Los Angeles made a series of presentations to the Children's Commission and Senior Management. This built a process of consensus-building throughout the agency.
- In many places, evaluation is not part of the culture, given all of the other competing interests. Los Angeles's working towards changing the culture so that staff starts seeing evaluation as providing some useful information that supports and builds their practice.
- One has to treat a large system as if it were a small system and develop relationships with people in it.

## III. Using Identifying Information in Evaluation

- The strength of the unions is one barrier to identifying individuals in evaluation.
- One of the problems with identifying and evaluating individual workers is that training evaluation may result in "high stakes" evaluations. To create a culture change, one must reinforce the ability of supervisors to supervise effectively and support the transfer of learning. This type of evaluation has the potential to turn into a quick way for county administrators to evaluate staff, rather than a method to support the learning process.
- As one moves into higher levels of evaluation, there is no way to avoid using identifying information.
- The original design of the Los Angeles study had supervisors individually rating the workers they supervise. As the evaluation unfolded, time constraints prevented the completion of this part of the design.

### IV. Obtaining Institutional Review Board (IRB) Approval to Share Findings

- In Kentucky, one must go through the IRB process and get approval to share findings. Knowledge testing may have presented a problem, however, because it was stated in the IRB that the University made this testing mandatory. We explained that in pre-service training, supervisors are given the trainees' scores.
- If framed correctly, the University will view knowledge testing within its mandate of teaching and public service. Somehow it takes it a little farther out of the research area and more into the traditional services and functions domain.
- It is important to make the distinction between evaluation and research.
- Two ways to facilitate approval by the IRB:
  1) If you have a mandatory system where you are evaluating for the state because it is an evaluation of your training system and you find the data is useful enough to present or publish outside the agency, you can go back and do a chart reviews study.
  2) Always get informed consents signed any time the evaluation is voluntary.
- One organization put a survey on the intranet in a child welfare agency. The IRB initially was not willing to approve it because they did not understand how it worked. It took awhile to identify someone in the University who was willing to go and tell the IRB it was okay.
- In large systems, approval must come from both the IRB and the agency with regard to publishing. Before beginning the evaluation, get in writing that you will be able to publish your data with their approval. Otherwise, if you proceed, you can get stuck and possibly have to enter a legal battle.
- Every IRB is different. The Feds have guidelines but there is much individuality among the different IRBs. Some consider evaluation as research.
- IRBs are changing. If you talked to them two years ago about a study, talk to them again. Every day there is a headline about leaking confidential information onto the Internet, which is causing a knee-jerk reaction.

- There are a lot of IRB training programs through universities that you can take via the Internet.

**V. Research Ethics**

- Working with indigenous people, and other oppressed groups, one has to not only think in terms of standard research ethics about individuals, but how to do this in a respectful manner with communities that have been exploited by researchers.
- One organization conducted a review not only with the University, but with tribal leaders. The review helped to rethink the methodology used. It also created some tension with the university about intellectual property and with the community about what it "owns."
- The same kind of issues come up with other groups, such as domestic violence workers, in terms of always checking with them about the methodology being used.

**VI. Soliciting Feedback from Families**

- Conflict and awkwardness often arise when workers ask families directly for feedback or when they receive critical feedback from families. This practice is not as highly regarded in the field as it is in training.
- Alternatively, supervisors may ask clients about their workers' performance, as one would do with any other product or service area where it is not necessarily the provider that asks. Having another part of the hierarchy of the agency actually solicit feedback may yield better results.
- Kentucky has an anonymous customer service survey that the state administers every year to clients, foster parents and foster children, particularly to older foster children aging out of care. The survey is administered by region to know how each region is doing. This consumer evaluation is conducted at a higher level to get around the power imbalance between the worker and the client and yields a 60% response rate by mail.

### VII. Use of Relationships in Training Evaluation

- Part of relationship-building is starting where the client is and that happens at the system level, too. In a pure professional model, people would request feedback on their practice because they aspire to be the very best that they can be. Yet, in the political context of child welfare, it is not necessarily safe to have that conversation.
- Technology changes the way relationships are developed and what is focused on. One feels less obtrusive when sending emails but one cannot form a relationship over e-mail. People will respond to an e-mail once they know you.

# Development of Multi-Level Methodologies to Support Mainstreaming Training Evaluation

E. Douglas Pratt, *D.S.W., LCSW*

## Abstract

Since the early 1990s, human services training evaluation has experienced a thrust of developmental acceleration. In this context, trainers, evaluators and planners have been discussing the meaning of "mainstreaming evaluation" and the related quality of "transparency." This paper explores the potential for one field-tested multi-level methodology to promote mainstreaming, and makes a functional connection between mainstreaming and transparency. Development of new methodologies such as this, which evaluate the impact of training on families and communities, is recommended.

## Mainstreaming Training Evaluation

Traditionally, training evaluation in human services has functioned within fairly clear limits. The scope of training evaluation has been limited, mostly to six important functions: to measure participants' satisfaction, determine pass-fail for participants, monitor trainers' performance, provide accountability documentation, improve curricula, and conduct research. Traditionally, training units have been somewhat distant from services and other units within human service systems. In this context, distance between evaluation and curriculum design or training may be a tolerable norm. Enhanced collaboration that extends the scope of training evaluation to include post-training performance as well as outcomes for families, the organization or the community is exceptional, but envisioning this gives promise for our future in the mainstream of human services.

For the purpose of this paper, mainstreaming training evaluation is defined as the process of helping all other units of a system develop partnership with the evaluation unit so each routinely derives maximum benefits from evaluation of training. The outcome of this process, mainstreamed evaluation, is enculturated mutuality among evaluators, clinicians, managers and others who collaboratively use results as essential to staff and organizational development, interagency collaboration, and attainment of long-term community outcomes. Reciprocally, evaluators in the mainstream seek feedback and understanding of others' priorities to design evaluation that strategically monitors and promotes the long-term vision of success.

### *Transparency in Training Evaluation*

One result of a traditionally isolated position is that evaluation methods have often been poorly explained or understood. Misunderstood methods limit the capability of system decision-makers to gain maximum benefits from evaluation. The power of an evaluation unit has sometimes been exaggerated, in the expectations of administrators as well as colleagues with minimal power in the organization. On one hand, some administrators may use training as the main or only way to reform practice and then rely on evaluation as an accountability measure, to document the diligence of their efforts regardless of outcome. On the other hand, trainees or the trainers themselves sometimes fear that evaluation alone could damage one's career development.

Too often evaluation has been sequestered. Training itself has been conducted primarily in classrooms, apart from the practice setting, and evaluation may be located even further from practice arenas. Research design, instrumentation, testing and other data collection, analysis and interpretation of findings have often been done by a few scientists who may have little or no practice experience with the abilities being taught. They may work in locations that afford rare contact with the practice setting. Bringing evaluation from misperception and sequester into the mainstream is a systemic challenge.

Transparency is a system quality that can help colleagues see evaluators as individuals and partners in a useful win-win process. For the purpose of this paper, transparency is defined as the inclusion of all who might benefit, in the process and utilization of training evaluation. At the very least, transparency is openness with all who might benefit, in language that is useful for them,

about the strengths and needs identified by evaluation, with candor about what remains unknown.

***Is It Evaluation or Is It Intervention?***

People tend to do what they know is being evaluated. Exploiting this Hawthorn Effect needs to be strategic and regarded as an intervention of sorts. To extract all benefits from training evaluation, our systems need relationships and routines that planfully exploit a broader array of uses and methods for evaluation, including the use of evaluation as an organizational development intervention.

Training evaluators usually assume that after we evaluate training, somebody is going to change something as a result. When this assumption is explicit, it is expressed in terms that come close to Kirkpatrick's language for Four Levels of training evaluation (Kirkpatrick, 1994).

For example, at Level I, evaluation results might be used to help make the training feel more positive to participants. At Level II, results may be used to help trainers enhance their focus on some learning objectives that participants are not demonstrating at an acceptable level. At Level III, results may be used to help supervisors coach the transfer of learning. At Level IV, results might be used to help resource families and case managers apply successfully-transferred skills to enhance their teamwork so the community demonstrates timely permanence.

Using a systems perspective, we can identify most of the concentric ripples radiating from any change related to training evaluation. We may observe, for example, how a clinician's newly acquired family-centered knowledge and skills affect a family, but also how the clinician's learning affects their informal and formal support persons. More specifically, the effect of Level II evaluation could be to reinforce the application of clinicians' new knowledge, or it could be to increase clinicians' defensiveness and constrict their ability to individualize each family and its support team. But when evaluation units partner with stakeholders to anticipate and strategically exploit the "concentric ripples" around training evaluation, then organizational development and even community systems can be enhanced.

Intervention based on Level III evaluation is less traditional and comes closer to a mainstreamed role for training evaluation. Intervention based on Level IV evaluation, considered by many an extraordinary use of training evaluation, is proximal to the

mainstreamed function. Simultaneous interventions based on all four levels of evaluation are conceptually most consistent with mainstreamed evaluation. When a system develops to the point that people at all levels routinely expect to craft, support and benefit from interventions based on any of the four levels of evaluation, that system may be a functional example of mainstreamed training evaluation.

### *Developing Methodology Useful for Development*

Under-developed capability for Level III evaluation has enabled some misuses of Levels I and II evaluation. For example, a tacit expectation of many decision-makers is that Level I shows whether morale has been boosted by popular or entertaining trainers who seem to know what they're doing. Decision makers may celebrate this result so quickly, that they under-utilize Level II findings. That is, any immediate aid to retaining staff and boosting morale is a priority for systems with resource allocations that are chronically inadequate for their mission. This shortsighted use of Level I enables under-utilization of the Level II results. Together, these weak uses of Levels I and II functionally reward entertaining trainers regardless of their effectiveness with curriculum objectives, and enable inadequate TOL strategies. Eventually, informal Level III evaluation ("grapevine" and "parking lot" methodologies) concludes that practice and outcomes do not seem any better. Training that may have potential for enhancing practice (Level III) and outcomes (Level IV) will wither.

In the early 1990s, the State of Alabama collaborated with external consultants to design formative and summative evaluations of Alabama's new family-centered, strengths-needs, community-based model of child welfare practice. Practice development was driven from the top down and enabled by major changes in funding and policy. Local practice champions, state managers, and external consultants designed the model of practice and implementation plan collaboratively. This added input from the bottom up.

### *Outcome Based Trainings*

The model of practice was launched with training as one of the primary implementation strategies. Supervisory Effectiveness Training (SET) helped supervisors become practice-skills coaches. Alabama Certification Training (ACT) helped experienced case managers unlearn social control modes and develop skills for family-centered engagement and family-team building. Some

Alabama trainers were trained to facilitate ACT for new case managers (four weeks in class and four weeks on-the-job-training (OJT)). Other Alabama trainers were trained to facilitate a program preparing foster and adoptive parents as professional colleagues with family-team competencies. In addition to OJT and supervisory coaching, external consultants were deployed to county offices to support transfer of learning (Bazelon Center for Mental Health Law, 1998).

### *Evaluating Practice Development*

The evaluation methodology that the external consultants developed, Qualitative Case Review (QCR)[1], was intended to evaluate practice development in the broad sense, including supervisory skills-coaching and managerial supports for inter-agency, multi-disciplinary, family-team strategies. More specifically, QCR was designed to evaluate the subtle qualities of practice, how core skills are strategically timed and combined to engage each unique family or community partner.

A spin-off from the development of QCR was the federal Child and Family Service Review (CFSR). QCR is the prototype for CFSR, and a number of systems use QCR as rehearsal for CFSR. CFSR is distinctive in that it has more of an accountability focus than a quality of practice focus. For example, CFSR tends to emphasize the correct number and frequency of visits more than what happens during a visit that contributes to a case outcome. However, the different weights each methodology gives to accountability measures versus quality measures make QCR useful for training evaluation at Levels III and IV. Thus, while QCR was not initially conceptualized in terms of Levels III and IV evaluation, this potential became clearer when training and outcome evaluators started examining QCR methodology through the Kirkpatrick lens.

## Design

Satisfaction with Alabama's trainings and knowledge of the state's model of practice were evaluated with traditional Levels I and II instruments. Because it is difficult to imagine a direct connection between classroom training and client population outcomes, the planners, who had great faith in training and

[1] QCR is proprietary with Human Systems and Outcomes, Inc., Tallahasse, Florida, copyright 2003.

development in the 1990s, did not fully appreciate the potential QCR held for Levels III and IV evaluation of Alabama's trainings.

QCR did capture the transfer of ACT case management skills and strategies, in terms of child and family status. QCR examined the transfer of ACT family-team strategies, in terms of system performance with specific families. QCR also captured the use of SET and ACT skills, GPS-MAPP[2] abilities, and local managers' use of the practice model in building collaboration among their community partners, in terms of system performance outcomes.

By 2005, QCR had been individualized for many states and communities, but the design remained essentially the same as in the early 1990s[3]. To encourage and enable exploitation of QCR potential for Levels III and IV evaluation of training, some of the more relevant components of the methodology are explicated in the sections that follow.

### *Preparation for QCR*

To prepare for QCR, all managers, supervisors, and clinicians or case managers are briefed on the steps of QCR, with emphasis on what's in it for them. They learn that they will discover more about what they and their community partners are doing effectively, and that they will receive concrete suggestions to enhance progress with individual families' visions of success and permanency goals, for example. Encouraging staff to list what they might gain from a reviewer who can be seen as their personal case consultant for a few days, is an example of transparency designed into the methodology.

Six to twelve cases are randomly selected. The family members are contacted to explain the review process and gain their consent in a signed release, permitting reviewers and case collaterals to communicate.

### *Prep for Days One to Four*

Interviews are scheduled for the first half of the cases on days one and two of the five-day review and for the second half of the cases on days three and four. Parents, children, kin, informal and formal supports are scheduled at locations most convenient for

---

[2] Designed by The Child Welfare Institute, Atlanta, Georgia, which holds 1990s copyrights on SET and MAPP-GPS.

[3] *Qualitative Case Review* methodology has evolved as it's been individualized for human services and child welfare systems in New Jersey, Utah, Florida, Georgia, New York City, Iowa, New Hampshire and other systems. Currently, QCR is also known as *Quality Case Review* and *Quality Service Review (QSR).*

them, usually at their residence or their office. An important strategy is interviewing all formal and informal supports, ranging from a physician to a birth father who has been labeled "uninvolved" for years. When time or distance present barriers, phone interviews are scheduled. Case records are pulled; case managers assemble packets containing a case summary, the signed releases, and the interviews schedule with phone numbers and driving directions to each meeting or interview.

Closed debriefings on each case are scheduled for the reviewer, case manager and supervisor for the end of days two and four. Larger group debriefings are scheduled for all supervisors, managers and reviewers as the last activity on days two and four. Stakeholder focus groups with community partners such as mental health, education, and Juvenile/Family Court, are scheduled for simultaneously with the case reviews for days one through four.

***Prep for Day Five***

Large group exit debriefings are scheduled for day five. The first, for all supervisors and managers, is scheduled for the morning. The second, for all managers and the community stakeholders, is scheduled for early afternoon. The process for these groups, called parallel process, is explained below.

## Instrumentation and Data Collection

The "QCR Protocol" is a handbook for structuring interviews and rating data. The first section organizes data about case management, child, and family status in 12 parameters. The second section organizes data about family-team building and community performance in thirteen parameters. Each of these 25 parameters has six lists of rating criteria, one for each anchor on a Likert scale (from 1, which represents worsening family status or system performance, through 6 which represents optimal status or performance).

Each parameter page is paired with a worksheet page for organizing brief notes from each interview. The Protocol becomes the raw data for each case, to be retained along with a score sheet of parameter ratings, a prognosis rating, and a summary score.

Interviews begin with the case manager, who is asked first how the reviewer's two days could be helpful to her or him. This exchange builds mutuality, decreases natural anxiety, and is an example of transparency designed into the methodology.

Data collection interviews are conducted using parallel process. That is, reviewers consistently model strategic use of the skills that are (or will be) taught in training, coached in supervision, and modeled by administrators. For example, interviews as well as the debriefings consistently begin with demonstrations of respect, reflections of strengths and emotions, and questions that help the person who is interviewed describe what is important to him or her. Problems are discussed second, and are reframed by the reviewer in terms of challenges, needs, or opportunities.

***Developing Internal Capacity***

Experienced reviewers are paired with agency or community "practice champions," who observe and assist the reviewers as indigenous experts. These "shadows" study the reviewer's parallel process skills and learn the QCR procedures. As part of the system being studied, the shadows would be considered research "subjects" in a traditional evaluation. QCR methodology intentionally softens the scientist-subject roles and boundary, however, to gain four specific benefits.

Actively involving these "subjects" in data collection, in rating and in reporting, builds mutuality, positive expectations for enhanced success, and diminishes anxiety. Most importantly, the way reviewers engage and involve shadows ultimately empowers system ownership for the QSR process.

Several systems have included elected legislators as shadows. Knowing their legislators would see "the good and the bad", these systems' planners trusted the parallel process. That is, the planners knew QSR reviewers would help legislators reframe their experiences so everything they learned would be solution-focused and would empower them to be system advocates. This approach to using shadows is an example of transparency designed into the methodology.

Astute reviewers and shadows bring to QCR a detailed understanding of the competencies from the trainings that support (or will support) the model of practice. They model these with conscious competence, and for some systems the reviewers coach shadows who have been selected to become reviewers. As reviewer candidates, these shadows take the lead with the second case and complete the week by writing their Individualized Reviewer Candidate Development Plans. Thus, reviewer development is parallel process with the model of practice.

On day five, a remarkable number of shadows spontaneously tell debriefing groups that their capability for strategic use of core skills has grown beyond their expectations. These practice champions go on to energize their colleagues and promote the use of QCR results in practice, organizational and system development (Bazelon Center for Mental Health Law).

## Analysis and Report of Findings

Each reviewer-shadow pair presents findings, as strengths and needs or suggestions, to the case manager and supervisor in private. The pair then present these findings in the supervisors-managers debriefing groups, typically scheduled at the end of days two and four. In each presentation, they invite the listeners to ask questions and add insights or corrections. This is another example of transparency designed into the methodology.

A strength of the methodology is that qualitative and quantitative data are each necessary, but separately are insufficient for useful findings. Following all the case debriefings, aggregate findings are prepared. Qualitative strengths and suggestions, quantitative ratings of parameters, and final scores from all the cases are summarized as system scores. At this time, the findings of the community stakeholder focus groups are also summarized as organizational and community strengths and needs.

On day five, in the managers-stakeholders exit debriefing, the aggregate case review findings are presented. Strengths are discussed first, needs or opportunities are discussed second, and great care is taken to keep these separate until all have been listed. These are listed on newsprint and kept visible throughout day five, to enhance stakeholders' trust and motivation.

Using parallel process, reviewers ask participants what this summary means to them. Almost invariably, stakeholders are enthusiastic about seeing the strengths of their performance with families. They embrace and add to the list of suggestions that follows strengths, as an exciting opportunity to build on their own momentum. The reviewers' strategic use of parallel process to facilitate the group is an example of transparency designed into the methodology.

After a break, the managers-stakeholders exit debriefing is reconvened for the summary of stakeholder focus group findings.

These strengths and needs are listed on newsprint and posted to energize the focus on solutions related to their vision of success.

Again, reviewers' use of parallel process to facilitate this group enables participants to craft and list what they see as important next steps for enhanced training, TOL, community partnerships and outcomes.

***Utilization of Findings***

QCR can be used in at least three ways: to establish a useful baseline; to monitor and shape the progress of practice development; and to demonstrate attainment of a standard.

QCR is often used to establish a baseline for practice development. The research-based assumption driving QSR methodology is that certain core skills, used strategically for engagement of families and community partners in trusting working relationships, promote an evidence-based model of practice (Miller, 1998, Hubble, 1999, Egan, 2004, Hawaii Department of Health, 2003). Therefore QCR establishes a particular type of baseline, comparing the quality of current practice to a generic model of evidence-based family practice.

Planners choose QCR knowing this in advance, and then use their system's baseline strengths and needs to craft their own individualized version of the general model. Then they will articulate performance expectations and managerial supports, and commission training designs to support implementation of their model.

QCR will then be scheduled as a series of formative evaluations. QCR can be used for Levels III and IV training evaluation; for evaluating the effectiveness of administrators' steps and community stakeholders' steps for growing the model; and evaluating the quality of partnership among administrators and stakeholders.

Circumstances, such as the need to qualify for funding or to satisfy a settlement decree, may require summative evaluation. QCR is a useful summative evaluation methodology because it combines accountability measures and sensitive quality measures.

***Tuning QCR for Training Evaluation***

QCR can be fine tuned for training evaluation when that is one of the system's explicit purposes. The two most important adjustments for training evaluation are 1) the process of data collection and 2) instrument calibration. For data collection, reviewers are selected who know the curricula and objectives, and

who have also practiced using those skills and strategies in the model of practice. Sensitivity of the instrumentation can be focused by matching curricula objectives with the appropriate QCR parameters, then adding the objectives to the lists of criteria that guide rating on the Likert-type scales.

### *Limitations of the Method*

The long-term benefits of training shadows and system managers to steward their own ongoing QCR process appear to outweigh any limitations of the method. Ironically, one limitation is degradation of the method when a system uses it independently. Until a system develops its internal capacity to steward QCR, there is financial cost for the method. In any case, the chief limitation of the method is that it requires detailed planning, reliable teamwork, and deployment of many to this temporary duty.

### *Relaxing the Boundaries*

The process of QCR integrates, almost seamlessly, the evaluation of training with the ongoing development of practice, of the organization and community partnerships. Boundaries between evaluators and the evaluated, between child welfare and other specializations, become relaxed and permeable.

Remember the folksy caution to be careful what you pray for? At this exciting stage in the development of human service systems, any Level IV training evaluation that identifies changes needed in the organization or community will likely raise some anxieties. Resistance is a natural response to change, so it is not hard to imagine that when training evaluation identifies changes in organizational norms, evaluation might be subject to backlash. Reactions like this may send evaluators back toward the traditional limits for training evaluation, somewhere between Levels II and III.

### *The Heart and Soul of Change*

Since we can't control the anxieties of others, it seems prudent for training evaluators, who willingly share stewardship for the development of our human service systems, to be aware of our own anxieties. Some evaluators may experience anxiety about the political risks if we advocate for mainstreaming. Others may be more anxious about taking on the span of responsibility attendant with mainstreaming.

Tacit anxiety can obscure evaluators' best intentions but may become empowering when evaluators risk giving it voice. Most will choose to channel their anxieties positively, allowing them to

be transparent, and will reach out to each other for support. Mutually supported, they will ease the slowest along.

Transparency in relationships, Rogerian genuineness, is a useful quality for managing the anxieties of rapid methodological and practice culture development. For training evaluators, transparency is a parallel process; what evaluators expect of each other is needed of trainers also, of coaches, clinicians and case managers, and certainly of administrators. Envisioning our individual transparency within a transparent multi-level training evaluation process gives promise for our future in the mainstream of human services.

## References

Bazelon Center for Mental Health Law. (1998). Making child welfare work: How the R.C. lawsuit forged new partnerships to protect children and sustain families. Washington, D.C.: Judge David L. Bazelon Center for Mental Health Law.

Egan, G. (2002). *Exercises in helping skill: A manual to accompany the skilled helper (7th Edition).* Pacific Grove, CA: Brooks/Cole, Wadsworth Group.

Hawaii Department of Health (2003). Summary of effective interventions with youth. Evidence Based Practices Committee Biennial Report.

Hubble, M.A., Duncan B.L., & Miller, S.D. (Eds.) (1999). The heart and soul of change: What works in therapy. Washington, D.C.: American Psychological Association.

Kirkpatrick, Donald L. (1994). *Evaluating training programs: The four levels.* San Francisco: Barrett-Koehler.

Miller, S.D., Duncan, B.L., & Hubble, M.A. (1996). *Escape from Babel: Toward a unifying language for psychotherapy practice.* New York: Norton Professional Books.

## Discussion

*Title:* A Mainstream Training Evaluation Methodology: Transparent Use of Practice Performance Measures
*Presenter:* E. Douglas Pratt, *D.S.W., LCSW*
*Facilitator:* Sherrill Clark, *Ph.D., LCSW*

---

E. Douglas Pratt began by explaining that the training evaluator's task involved the two extremes of training evaluation, the hard science, statistics, instrumentation, measurement on one end and the values, philosophy, model of practice on the other, and integrating the two. He asserts that the goal is to create this integration at both an individual and organizational level, both of which may be addressed using the Quality Service Review (QSR) method. Discussion of this method centered on the use of client feedback and an example of how Quality Case Review (QCR) was implemented in Utah.

### Topics of Discussion

**I. Use of Client Feedback**

- Client feedback is vital. One must ask individual family members what their goals are and use that for evaluation purposes, for general training and to do good practice.
- Agencies that use family-team conferencing set forth plans that indicate goals. Even if staff have certain goals in their conferencing plans that are not in their individual work, this does not mean that their goals are in conflict.
- Sometimes, in family-team conferencing, there are family goals out on the table that are authentic; yet, sometimes there are goals that cannot be put on that table for everybody to be aware of or participate in because of risks, like domestic violence, that may be present. How does one use that when evaluating practice development? The dilemma is presented to the system and requires that they articulate the meaning of it.

- As reviewers, we cannot be definitive about what is going on in the family but we can look at the situation and work together to decide what it means in terms of practice, procedures, and safety protocols, when needed.

**II. Quality Case Review (QCR) in Utah**

- Under settlement agreement, since 1994, QCR has been part of the annual review in each of the 5 regions.
- Before the use of QCR, workers misperceived their own practice and were not able to self-evaluate well. They wanted to do better work and recognized that the strengths-based, family-centered (SBFC) practice gave them something they were missing.
- QCR has been effective in helping implement SBFC practice because every year, every region knows they will have 24 or more cases randomly pulled and reviewed.
- The QCR has not led to a compliance mode; rather, it has allowed us to get out of a compliance mode. There is consensus in Utah that this is the best way to serve families.
- Part of the concern in doing training evaluation is considering the outcomes of the families, so that one creates an evaluation that is ecological and contextual. Focusing only on the training is short-sighted as it won't get the workers or the system where they need to be.
- For workers to provide optimal practice, training evaluation must be tied to the outcomes of children and families.
- The value of the QCR is that it keeps everyone focused on outcomes. It allows for a random review of the outcomes of 24 children and families in a particular region to see how well case workers are doing at engaging, assessing, teaming, planning, and intervening.
- The QCR provides a single direction that everyone can agree on and to which they can adhere. It is more stringent than the CFSR. It captures what is missing in the system. It is precise enough to reveal where the issues are in safety, well-being, and permanency. For example, it may inform managers that their workers don't have the assessment skills they need. They may have the information-gathering skills but there are other more sophisticated skills they lack.

- Legislators are involved in the QCR. It is simple enough so that someone shadowing the process for a day or two can understand the model.

# Discussion: Brainstorming an Agenda for the National Staff Development and Training Association (NSDTA) and Evaluation Committee and Certification Committee

*Theme:* Collaboration on National Issues in Training Evaluation
*Title:* Brainstorming an Agenda for the National Staff Development and Training Association (NSDTA) Evaluation Committee and Certification Committee
*Facilitators:* Dale Curry, *LCSW, Ph.D.,* and Cynthia Parry, *Ph.D.*

---

## Topics of Discussion

Dale Curry and Cynthia Parry held a brainstorming session to discuss the direction of NSDTA, generate interest in joining the evaluation and certification committees and get participant ideas for future initiatives and collaborative projects.

### I. What is NSDTA?

- The National Staff Development and Training Association (NSDTA), an affiliate of the American Public Human Services Association (APHSA), has existed for 20 years.
- NSDTA was established to create a bank of the types of evaluation projects that were happening and the new tools that have emerged.
    1) The latest initiative is to continue to work with CalSWEC on NHSTES.
    2) Possible future initiatives may include:
        a) inter-state collaborative projects,
        b) submitting proposals for child welfare training grants, and
        c) focusing on certification for training and development professionals.

**II. Training and Development**

- Training and Development in human services is being viewed as a field of study and an emerging profession.
- Establishing the *Training and Development in Human Services* journal and other publications contributes to this movement.
- A code of ethics for training and development staff was adopted two years ago.

**III. Committee's Goals and Collaborative Projects**

- Work on creating a training certificate program for training and development professionals focused on the NSDTA competency model of nine roles and competencies.
- Focus on higher skill development for evaluators.
- Access resources and create linkages from the American Evaluation Association at their national conference.
- Organize small workshops at next year's NHSTES for participant skill development.
- Update current publications or develop additional publications/newsletters to share information; submit articles to the journal, *Training and Development in Human Services.*
- Identify other sources of revenue besides the annual national conference, such as national grants.
- Collaborate on a national or multi-state study to build knowledge on outcomes and their relationship between education (MSW, BSW, IV-E) and training.
- Other ways of sharing ideas and information in the interim between gathering annually at NHSTES include:
  1) Collaborating to develop curriculum,
  2) Developing a national item bank, and
  3) Replicating studies to build upon evidence-based practice body of knowledge.
- Encourage NHSTES participants to present at NSDTA.
- International linkages or contacts with other groups or universities.

**IV. Next Steps**

- Mr. Curry and Ms. Parry will contact interested people for participation on the committee.

- An initial conference call will convene prior to the next NSDTA conference in November 2005.
- Mr. Curry will set up a meeting at the NSDTA conference.
- Will establish an initial conference call.
- Will meet at NSDTA conference.

# Closing Remarks

Leslie W. Zeitler, *M.S.W., LCSW*

For the past three days, we've been talking about several topics related to training and training evaluation: ethics, values, transparency, honesty, what we are actually training people to do, linking training to outcomes, and assessing the way training evaluators evaluate skills and knowledge acquisition and transfer. Based on these discussions, it is clear that one of our major goals is to improve child welfare services via an integrated evaluation mechanism in the training class.

I was particularly struck by Katherine Cahn's presentation on the challenges of measuring attitudes and values in training. Katherine encouraged us to take on the task of figuring out how to incorporate measurement of attitudes and values into training evaluations—especially as the field of child welfare struggles to address the disproportionate numbers of African-American and Native American children in the child welfare system.

Utilizing the Child and Family Service Reviews, the Administration for Children and Families has made it clear that they are interested in the outcomes of safety, permanency, and well-being for all children in care. The disproportionate representation of Children of Color in America's child welfare system is clearly an undesirable outcome. We cannot talk about outcomes of safety, permanency, and well-being, without dealing with one of the most obvious outcomes that has a negative impact on hundreds of thousands of children in this country.

The most recent data available indicates that 35.4% of California's public child welfare population consists of African-American children, and yet this group comprises only 7% of the entire child population in the state. Additionally, 13.8% of the entire public child welfare population consists of Native American/Indian children, yet this group comprises less than 1% of the entire child population in California. Taken together, African-American and Native American children represent almost half

(49.2%) of all children involved in California's child welfare system, even though their combined representation in California's total child population is approximately 8%. California data is consistent with national statistics regarding the disproportionate representation of African-American and Native American children in the child welfare system.

If members of communities of Color abused or neglected their children at significantly higher rates than White parents, then the higher numbers of African-American and Native American/Indian children involved with public child welfare might make sense. However, results from the National Incidence Studies of Child Abuse and Neglect (NIS) reflect no significant differences in maltreatment with relation to race. The Third National Incidence Study (NIS-3) found "no race differences in maltreatment incidence. The NIS-3 reiterates the findings of the earlier national incidence studies in this regard. That is, the NIS-1 and the NIS-2 also found no significant race differences in the incidence of maltreatment or maltreatment-related injuries". These studies magnify the significant discrepancy between who maltreats and who eventually is over-represented in the public child welfare system.

The current rates of child welfare removals of African-American and Native American children are reminiscent of earlier times when U.S. government programs either targeted or turned a blind eye to large segments of the population. Specifically, during the four centuries in which slavery was legal in the United States, Americans of African descent were routinely abused, tortured, killed, and separated from their husbands, wives, children, extended family members, and communities. During these same four centuries, indigenous peoples native to the North American continent (all Indian nations within U.S. borders) were subject to mass genocide and were routinely separated from their families and children as Europeans settled in North America. Even as late as the 1950s, Indian children were separated from their families and national/cultural groups, and forced to attend boarding schools to eliminate their cultural identities and familial bonds.

Given the historical context, and given that separation of children from their families (regardless of circumstances) is traumatic and has long-lasting impact, might professionals working in the field of public child welfare be unwittingly participating in

the re-traumatization of historically targeted populations? What is our ethical obligation in mitigating this trauma for families of any racial/ethnic/cultural groups who have historically lost their children to government or institutional mandates?

What the child welfare system, and we training professionals who are part of this system, are left to contend with is that institutionalized racism is one of the most likely reasons that we see the disproportionate numbers of children of color in the child welfare system. Each time I see the statistics regarding the disproportionate representation of children of color in the child welfare system, I reflect upon the fact that even with the best of intentions, there has been some level at which I have participated (albeit unconsciously) in institutionalized racism during my tenure as a social worker affiliated with the field of public child welfare in the United States.

This difficult piece of self-reflection requires me to ask myself: How am I unconsciously complicit in institutionalized racism, and how do I end that complicity? As a training evaluator, what role can I play in helping people become more aware of this dynamic? What are my ethical obligations in addressing this subject matter as a training and training evaluation issue?"

Trainers and training evaluators affiliated with public child welfare draw from several professional backgrounds, and are guided by their respective codes of ethics. My own professional conduct is guided by the NASW Code of Ethics. Several sections of the NASW Code of Ethics can apply to our discussion of addressing issues of Fairness & Equity in training evaluation: Section 4.02, Section 6.04(b), Section 6.04(c), and Section 6.04(d) (NASW Code of Ethics, 1996). However, I want to specifically pay attention to two sections of the NASW Code that seem particularly relevant to the overlap between child welfare workers, child welfare trainers, and child welfare training evaluators:

- "Social workers should promote conditions that encourage respect for cultural and social diversity within the United States and globally. Social workers should promote policies and practices that demonstrate respect for difference, support the expansion of cultural knowledge and resources, advocate for programs and institutions that demonstrate cultural competence, and promote policies that safeguard

the rights of and confirm equity and social justice for all people."

- "Social workers should act to prevent and eliminate domination of, exploitation of, and discrimination against any person, group, or class on the basis of race, ethnicity, national origin, color, sex, sexual orientation, age, marital status, political belief, religion, or mental or physical disability."

The challenge for training evaluators is to operationalize incorporation of knowledge (NIS studies in comparison with AFCARS data), ethics, and direct practice to assist in changing the outcomes for all children and families involved with public child welfare. Right now, we have more questions than answers. Here are some questions to start with as we support and challenge one another:[1]

- How would one write a knowledge item on a bias, attitude, or value?
- How would one assess skill acquisition regarding biases, attitudes, or values - and what would that look like?
- How would one assess transfer of learning regarding a bias, attitude, or value?
- How would one link training about biases, attitudes, or values to statistical outcomes?
- How can we encourage child welfare workers to work with their own biases, attitudes, and values that affect the provision of services, especially if trainers don't incorporate this into their trainings at some level?
- How can we encourage trainers to incorporate discussion, self-reflection, and assessment of such biases, attitudes, and values into trainings if they don't have support from training evaluators and administrators (especially now, as we are at the first stages of linking training to outcomes)?
- When we address attitudes and values in training, what are we actually training people to do? How does this fit in with supporting evidence-based practices?

[1] Dr. Katherine Cahn discussed many of these questions during her presentation at this (2005) NHSTES regarding the challenges of training evaluation around issues of fairness and equity. In addition, she contributed additional questions via personal communication 27 May 2005.

- How do we define or measure attitudes and values in the context of supporting or promoting evidence-based practices?
- How can we support trainers and child welfare administrators to address the effects of biases, attitudes, and values on the disproportionate representation of children of color in the system?

We training evaluators find ourselves at a crossroads. With the hard work of training evaluators who have come before us, the availability of statewide and national child welfare statistics, the benefit of historical perspective, prodding from the Feds, the NSDTA in the process of drafting a code of ethics for training professionals, and improved ability to link outcomes to training[2], the field of training evaluation is poised to take a giant leap forward.

Given that the field of child welfare training evaluation is about to make such a leap, what is our collective ethical obligation in addressing issues of disproportionality, fairness, and equity? I would like to hold myself accountable to all of you as I find my way in addressing biases, attitudes, and values that directly affect service provision to children and families involved in the public child welfare system. In fact, I invite all of us to hold ourselves accountable. Would you join me in doing so?

[2] Becky Antle's and Anita Barbie's presentation at this (2005) NHSTES nicely demonstrated the progress that Kentucky has made in this area.

## References

National Association of Social Workers. (1996). NASW Code of Ethics. Retrieved May 2005 from http://www.socialworkers.org/pubs/code/code.asp

Needell, B., Webster, D., Armijo, M., Lee, S., Cuccaro-Alamin, S., Shaw, T., Dawson, W., Piccus, W., Magruder, J., Exel, M., Conley, A., Zaman, J., Smith, J. , Dunn, A., Frerer, K., Putnam Hornstein, E., & Kaczorowski, M.R., (2005). Child Welfare Supervised Foster Care Highlights from CWS/CMS. Retrieved May 2005 from University of California, Berkeley Center for Social Services Research website, http://cssr.berkeley.edu/CWSCMSreports/highlights

Sedlak, A.J. & Broadhurst, D.D. (1996). Third national incidence study of child abuse and neglect: Executive summary. (Prepared under contract to the National Center on Child Abuse and Neglect, U.S. Department of Health and Human Services.) Retrieved November 30, 2005, from http://nccanch.acf.hhs.gov/pubs/statsinfo/nis3.cfm

U.S. Department of Health and Human Services, Administration for Children and Families, Administration on Children, Youth, and Families, Children's Bureau. (2005). The AFCARS Report: Preliminary FY 2003 Estimates as of April 2005. Retrieved May 2005 from http://www.acf.hhs.gov/programs/cb/stats_research/afcars/tar/report10.htm

# Wrap-up

Friday, May 27, 2005, 12:15 p.m.
*Facilitators:* Barrett Johnson, *M.S.W., LCSW,* and Leslie W. Zeitler, *M.S.W., LCSW*

---

Each year, a wrap-up discussion helps to identify the strengths of the symposium and suggested changes for next year. Coupled with the evaluation forms, this wrap-up serves as a course-level evaluation for the symposium.

## Topics of Discussion

### I. What Worked:

- Networking:
  - √ Networking/ time to interact with others.
  - √ Energy level was great, consistent.
  - √ Development of evaluator skills.
  - √ Great for university partners to get support on training evaluation.
  - √ Sharing professional resources.
  - √ Talking with other participants about specific issues, e.g., skills evaluation.
  - √ Opportunities to check about projects back home & how to adopt ideas learned here to them (e.g., evaluating legal training).

- Presentations, General:
  - √ Interactive exercise—it tricked us into saying what we were working on and helped us meet new people.
  - √ Eileen Gambrill's presentation.
  - √ Particularly the Kentucky (x7) and New York (x3) presentations.
  - √ Interactions on Thursday were great, especially presentations related to linking training to CFSR outcomes/family outcomes (x3).
  - √ I learned a lot from Basil's presentation.

- √ LA County presentation.
- √ Presentations from Lorna, Michelle, and Megan.
- √ Presentation from Barry and Cindy.
- √ All the presentations stimulated thinking.
- √ Good learning experience.
- √ Succinct presentations.
- √ Enjoyed the initial exercise.
- √ Showcasing of excellent programs.
- √ Hearing from many states.
- √ Listing of current projects others are working on.
- √ Handouts of presentations were appreciated.
- √ It seems to get more focused every year.
- √ Poster presentations—got additional information and other materials of what others are doing (x2).

- Presentations, Topics/Content:
  - √ How to develop a knowledge test.
  - √ The presentations were focused on processes and methods—what worked and what didn't. The specific info was wonderful.
  - √ Importance of seamlessness and strength-based approaches.
  - √ Topics and speakers spoke more directly to system needs than last year.
  - √ Going home with "tools": evaluation of testifying skills and knowledge, Kentucky tools, etc.
  - √ Learning about how large and smaller systems are working at being evaluated.
  - √ Learning the "errors."
  - √ Tips from colleagues about ways to do something I am doing right now (e.g., evaluate the fidelity of trainer delivery).

- Group Discussion:
  - √ Discussions about cultural competence issues, power, influence, and racism (x3).
  - √ Discussions about process surrounding evaluations.
  - √ Discussions on how to do evaluation in big systems.
  - √ IRB discussion.

- √ Respectful discussions addressing current issues.
- √ Small group discussions on first day.

- Atmosphere/Format:
  - √ Open dialogue.
  - √ The openness of the presenters and the discussions that follow (x3), especially Bob Highsmith and Henry Ilian's openness about challenges.
  - √ Very friendly group—do not always get this among evaluators.
  - √ The relaxed, informal format—the willingness of other participants to provide feedback. I really enjoyed this format- and look forward to it; you do a great job organizing this conference!
  - √ It was good to have a person (Basil) who is working on something and looking for feedback.

**II. Suggested Changes:**

- Discussions:
  - √ More time to discuss issues (x3).
  - √ Small group discussions identified by the group throughout the three days.
  - √ Consider limiting comments—a few participants tended to dominate conversations.

- Conference Room/Facilities:
  - √ Improve microphones.
  - √ The room was not conducive to optimal learning and participation. Not everyone could see overheads, hear the presenters, and the room was very crowded. (I had to keep moving my chair to allow people to navigate through the room, which was not only distracting for me, but also for others.) (x3).
  - √ A larger room (Clark Kerr campus!).
  - √ Softer chairs.
  - √ Sound system (super technology) could be improved (x2).
  - √ Microphones for questions from the participants.
  - √ The location of the poster room was useless. If you are going to have posters again, please have them directly in the rooms where people will congregate.

- Presentations
  - √ Opening activity tried to pack in too much. Great idea to start off with an interactive activity, but need to give enough time. The task isn't really the point; it is the interaction to get us connected.
  - √ Opening was weak; activity is good for helping to orient new attendees, but I think we need a good speaker to fire our imaginations.
  - √ Visioning exercise did not seem very useful except for questions that emerged.
  - √ Some presentations seemed only tangentially related to training evaluation. Perhaps screen proposals more closely to ensure relevance to purpose of symposium.
  - √ Make sure every presentation is specifically tied to training evaluation.
  - √ Have new presenters/presentations (x2); even consider inviting states that have not attended, but you are probably already doing that.
  - √ More sharing of instruments (e.g., Kentucky questionnaires).
  - √ Too much time spent on projects still theoretical (i.e., CalSWEC). I would like specifics on design, methodology, and outcomes.
  - √ Perhaps some higher-level theoretical analyses.

- Networking:
  - √ Regional sessions—opportunity for those from like regions of the U.S. to spend some formal time together sharing.
  - √ Continue to encourage networking and brief presentations.
  - √ More time to network with participants.

- Information to provide in the future:
  - √ Provide basic information about each state's child welfare training system, structure, staff, and populations served. This would help presenters by allowing them to provide an overview of their state at the beginning of the presentation; help participants to understand the broader/bigger picture in that state, and

determine if an evaluation activity could be applicable in their state.
- √ Handout in packet highlighting what others are doing.
- √ Consider a session in which participants strategize about evaluating a particular piece of curriculum. Use problem-centered learning approach.

### III. Application of Knowledge: What Participants Will Apply to Their Job

- How to begin to link training to CFSR outcomes (x2):
  - √ Will look into collaborating with agency QA unit to explore ways to link training to agency outcomes.
  - √ Develop a model that links training to outcomes.

- Programmatic:
  - √ Address development of test bank.
  - √ Continue to explore affordable TOL techniques.
  - √ Include ethical dimension in discussion of and teaching of EBP. Provide concrete examples of EBP.
  - √ Cultural Competence/Fairness and Equity in training evaluation.
    - Talk with agency about revamping cultural competence curriculum.
    - Explore our new cultural competence workshop for supervisors.
    - Write proposal on fairness.
    - Consider ways to support Family Group Decision Making (FGDM) and impact on organization.
    - Develop an overall training evaluation plan and proposal.
    - As best I can, download my brain to co-workers who are very interested/eager to continue on training evaluation activities.
    - Plan to follow up (re: resources) with at least one other attendee. Anticipate continued contact and conversation initiated from this year's symposium.
    - Coordinate with a particular person/ organization to design evaluation.
    - E-mail inquiries to colleagues.
- Methodologies/Applications:
    - Evaluation of testifying in court very helpful.

- KY team training, supervisors with caseworkers, very helpful.
  - Exploring use of case review tool such as used in KY that can provide outcome data for training evaluation.
- Spend more time looking at issues relating to reliability calculations and read up on multiple choice questions.
- Checking the items analysis of data where applicable.
- Re-do a current knowledge test which we use on a particular program to make it more effective.
- Views on working with large agencies (i.e., create smaller modules); QSR very informative.
- Use of multiple choice tests for knowledge base (Basil's test question development presentation very helpful.)
- Use state data system support evaluation.
- I will be developing a citywide evaluation of learning systems using embedded techniques and linking them to agency and city-wide outcomes.
- Methodologies for data collection.
- Find a way to utilize mock trial training in case worker training.
- Ideas for skill evaluation—stats for validating questions.
- We are designing a follow-up survey for a supervisor training, and there were several ideas we could use (e.g., web-based forms, way of designing scales, creating questions that reflect the learning objectives). Colleagues who are redesigning the case worker pre-service curriculum can use a lot of the info (e.g., creating multiple choice questions).
- Consider applying some of Kentucky's protocols.
- Apply Basil's statistical tests to our tests.

- Use of software
  - Check into "Zoomerang."
  - Exploring on-line surveys.
  - Explore Internet technology for surveys.

- Relationship-building:
  - √ Engage all stakeholders in conversations about their needs in evaluation.
  - √ Strategize with stakeholders in how training, competencies, practice and research will interface.
  - √ Need to form relationships in evaluation process.

- Review:
  - √ Resource material from other states.
  - √ Revisit decision-making and critical thinking literature to look for ideas.
  - √ Review the training materials and attend trainings I will be evaluating.

**IV. Future Topics:**

- CFSR/ PIP
  - √ More on how to link training to CFSR outcomes (x3).
  - √ Specifics and training evaluation samples.
  - √ Collect basic data on states prior to symposium for matrix.
  - √ Each state that attends should provide a three-minute update on how that state is using training evaluation activities to measure progress in meeting the PIP.
  - √ More on family outcomes.
  - √ Linking client outcomes to training (AHA Level 7).

- Remember the focus on evaluation:
  - √ Have a rigorous measurement sessions every year, like Basil's, but at a higher level for our professional development (x2).
  - √ Hands-on/skill building session for evaluators (x2).
  - √ More technical aspects of training evaluation.
  - √ Qualitative methods.
  - √ Small scale evaluations and evaluation ideas.
  - √ More depth on measuring attitudes.
  - √ Professional self-assessments evaluation.
  - √ Language/unintended consequence of focus on evaluation (post-CFSR).
  - √ Possible panel or workgroup re: evaluation strategies on Fairness and Equity (F &E) issues. Possibly connect with F &E Symposium in some way.

- Concrete examples of evidence-based practice (x2):
  - √ Follow-up on Chris's idea about integrating evidence-based practice with training evaluation level methods.

- Ethics:
  - √ Research ethics of carrying out program and training evaluation with non-mainstream communities.
  - √ Format is more important, but keep ethical issues at the center, with shared models and techniques supporting that discussion.

- Vertical integration of training and training evaluation-staff, caregivers, and providers.
  - √ Transform from an agency not being ready for training evaluation to doing training evaluation

- Discussions on cultural competency.
  - √ Issues of culture/cultural exchange working with child welfare agencies vs. partner agencies.

- Using mentoring in evaluation.

- Solution-focused casework.

- Interviewing skill.

- Update on federal efforts to block grant IV-E and eliminate training component.

- Reasonably affordable evaluation strategies.
  - √ Evaluations that are "quick" and affordable.
  - √ No-cost evaluations.

- How foster parents and other caregivers are evaluated by training system and how training is implemented.

- IRB, union, political issues.

- Let's hear from Joan Pennel in upcoming years on evaluation of FGDM training. She is working on evaluation of training on conferencing in N.C. and D.C., and she is a specialist at involving people who are affected by training and programs in their design.

- Lorna Bell's presentation piqued interest in hearing what is happening in other countries ad what we can learn from other disciplines.
  - √ International focus—similarities and differences

- Training evaluation approaches outside child welfare.

**V. How Can the Symposium Be Enhanced?**

- Great/ good job/satisfied (x14).
- Reaching out: Especially liked that those of us who have attended in the past needed to reach out to newcomers.

- Format:
  - √ Keep the Wednesday afternoon session.
  - √ More structured time to learn directly from each other.
  - √ Shorten interactive activity on first day.
  - √ 15 minutes of discussion appeared to be too short and left the audience with many unanswered questions—maybe 30 minutes.
  - √ Issue talking points to tables and report out at times.
  - √ Find ways to deliberately mix tables up (people sit in the same place all the time).
  - √ More small group work, if on a worthwhile topic.

- Presenters/Presentations:
  - √ Get more people on the planning committee who are not presenters.
  - √ Do more to screen presentation proposals (x2). A couple that seemed promising were very disappointing when presented (overly simplistic content).
  - √ Make sure all of the presentation forms focus on training evaluation rather than program evaluation.
  - √ Set presentation standards for presenters. For example, PowerPoint slides should not have more than eight lines, flip chart notes must be legible for participants.

- √ Provide basic training system data in packet so participants have baseline information for each state presenting.
- √ Get additional projects others are working on for handout.
- √ More international perspectives.

- Facilitators:
  - √ Spark broad questions instead of drilling down.
  - √ Make more observations; be more active.
  - √ Make sure people get called on.
  - √ Possibly build in table discussions before longer Q&A.

- Posters:
  - √ I actually don't think that the posters were a good fit for the symposium.
  - √ Poster session was a nice addition, but room was too far removed from the rest of the symposium.
  - √ Move to a roving resource/poster session.
  - √ Consider putting poster sessions in the program.

- Facilities/Environment
  - √ A different room layout so everyone can see better (x3).
  - √ Different facility. Hold symposium in a larger space (x4).
  - √ We need a larger room or a smaller group, but I am in favor of keeping the group at the current size and no larger and keeping a critical mass of evaluators working on current projects.

- Finding out interests ahead of time in order to structure groups at symposium

## Program

***Eighth Annual***
**National Human Services**
**Training Evaluation Symposium**
**2005**

### Wednesday, May 25

**2:00–3:00 p.m.**
**Registration**

---

**3:00–3:15 p.m.**
**Convene, Welcome, and Introduction**
Barrett Johnson, LCSW, *Regional Training Academy Coordinator, California Social Work Education Center (CalSWEC), University of California, Berkeley*

---

**3:15–4:45 p.m.**
**Theme: Trends in Child Welfare Training Evaluation**
*Title:* Future Search
*Presenters/Facilitators:*
- Teresa Hubley, M.P.A., Ph.D., *Muskie School of Public Service, University of Southern Maine*
- E. Douglas Pratt, D.S.W., LCSW, *Policy-Practice Resources, Inc.*

---

**4:45–5:45 p.m.**
**Theme: Addressing Fairness & Equity in Training Evaluation**
*Title:* How Do You Measure an Attitude or Value?
*Presenter/Facilitator:*

- Katharine Cahn, M.S.W., Ph.D., *Child Welfare Partnership, Portland State University*

***5:45–6:30 p.m.—Reception***

**6:30–6:45 p.m.**

**Welcome from the Dean**

James Midgley, Ph.D., *Dean and Specht Professor, School of Social Welfare, University of California, Berkeley*

***6:30–8:00 p.m.—Dinner***

## Thursday, May 26

***8:00–9:00 a.m.—Breakfast***

**9:00–9:10 a.m.**

**Reconvene, Introduction**

Chris Mathias, M.S.W., *Director, California Social Work Education Center (CalSWEC), University of California, Berkeley*

---

**9:10–9:30 a.m.**

**Opening Remarks**

Eileen Gambrill, Ph.D., *Hutto Patterson Professor, School of Social Welfare, University of California, Berkeley*

---

**9:30–11:00 a.m.**

**Theme: Training Evaluation in Large Systems**

*Title:* The California Framework

*Co-Presenters:*

- Cindy Parry, Ph.D., *CF Parry Associates*
- Barrett Johnson, LCSW, *Regional Training Academy Coordinator, CalSWEC, University of California, Berkeley*

*Title:* Aiming at a Moving Target: A Multi-Dimensional Evaluation Two Years Later

*Co-Presenters:*

- Bob Highsmith, Ph.D., *and*
- Henry Ilian, D.S.W.
  *Administration for Children's Services, James Satterwhite Academy for Child Welfare Training*

*Facilitator:*

- Norma Harris, Ph.D., *Social Research Institute, University of Utah College of Social Work*

***11:00–11:10 a.m.—Break***

**11:10 a.m.–12:10 p.m.**

**Theme: Linking Training to Outcomes for Children and Families**

*Title:* Measuring Training Practice, Child, and Organizational Outcomes

*Co-Presenters:*

- Anita Barbee, M.S.S.W., Ph.D., *and*
- Becky Antle, M.S.S.W., Ph.D., *Kent School of Social Work, University of Louisville;*
- Susan Kanak, M.B.A., *Muskie School of Public Service, University of Southern Maine*

*Facilitator:*

- Kyle Hoffman, *Institute for Human Services*

***12:10–1:15 p.m.—Lunch***

**1:15–2:45 p.m.**

**Theme: Developing Instruments to Measure Trainees' Acquisition of Specific Skills**

*Title:* Testifying Behaviors

*Presenters:*

- Michelle Graef, Ph.D., *and*
- Megan Potter, Ph.D.
  *Center on Children, Families, and the Law, University of Nebraska—Lincoln*

*Title:* Physical Restraint Techniques
*Presenter:*
- Lorna Bell, CQSW, Ph.D., *St. George's Medical School/Kingston University, School of Social Work, U.K.*

*Facilitator:*
- Jane Berdie, M.S.W., *Jane Berdie & Associates*

***2:45–3:00 p.m.—Break***

**3:00–4:00 p.m.**
**Technical Assistance Forum**
*Title:* Developing Multiple Choice Tests for Social Work Trainings
*Presenter:*
- Basil Qaqish, M.A., *Department of Social Work, University of North Carolina—Greensboro*

---

**4:00–4:30 p.m.**
**Logistics for evening and morning**
Leslie W. Zeitler, LCSW, *Training and Evaluation Specialist, CalSWEC, University of California, Berkeley*

***4:30 p.m.—Break for the evening***

## Friday, May 27

***8:00–8:45 a.m.—Breakfast***

**8:45–9:00 a.m.**
**Reconvene**
- Barrett Johnson, LCSW, *Regional Training Academy Coordinator,* and
- Leslie W. Zeitler, LCSW, *Training and Evaluation Specialist CalSWEC, University of California, Berkeley*

**9:00–10:30 a.m.**
**Theme: Mainstream Training Evaluation: Multi-Levels, Multi-Benefits**

*Title:* Issues of Practice Assessment in a Large System
*Presenters:*
- Bill Donnelly, LCSW, M.P.A., *Inter-University Consortium, University of California, Los Angeles;*
- Todd Franke, Ph.D., *and*
- Walter Furman, M.Phil.
  *Department of Social Welfare, University of California, Los Angeles*

*Title:* A Mainstream Training Evaluation Methodology: Transparent Use of Practice Performance Measures
*Presenter:*
- E. Douglas Pratt, D.S.W., LCSW, *Policy-Practice Resources, Inc.*

*Facilitator:*
- Sherrill Clark, Ph.D., LCSW, *Research Specialist, CalSWEC, University of California, Berkeley*

***10:30–11:40 a.m.—Break***

**10:40–11:40 a.m.**

**Theme: Collaboration on National Issues in Training Evaluation**

*Title:* Brainstorming an Agenda for the National Staff Development and Training Association (NSDTA) Evaluation Committee
*Facilitators:*
- Dale Curry, Ph.D., LCSW., *School of Family and Consumer Studies, Kent State University*
- Cindy Parry, Ph.D., *CF Parry Associates*

**11:40 a.m.–12:00 noon**

**Closing Remarks**

Eileen Gambrill, Ph.D., *Hutto Patterson Professor, School of Social Welfare, University of California, Berkeley*

***12:00 noon–12:15***
***Pick up box lunches for a working lunch.***

**12:15–1:00 p.m.**

**Wrap-up and Strategize for 2006**

- Barrett Johnson, LCSW, *Regional Training Academy Coordinator,* and
- Leslie W. Zeitler, LCSW, *Training and Evaluation Specialist, CalSWEC, University of California, Berkeley*

## Acknowledgements

*CalSWEC extends its gratitude to the Steering Committee of the 8th Annual National Human Services Training Evaluation Symposium, who made this event possible:*

| | |
|---|---|
| Anita Barbee | Barrett Johnson |
| Jane Berdie | Michel Lahti |
| Dale Curry | Chris Mathias |
| Midge Delavan | Michael Nunno |
| Bill Donnelly | Cindy Parry |
| David Foster | Debra Peel |
| Shaaron Gilson | E. Douglas Pratt |
| Michelle I. Graef | Shradha Tibrewal |
| Henry Ilian | Naomi Lynch White |
| Mindy Ing | Leslie W. Zeitler |

The *California Social Work Education Center (CalSWEC)* is a partnership between the state's schools of social work, public human service agencies, and other related professional organizations that facilitate the integration of education and practice to assure effective, culturally competent service delivery and leadership to the people of California. CalSWEC is the nation's largest coalition of social work educators and practitioners.

California Social Work Education Center (CalSWEC)
University of California, Berkeley
School of Social Welfare
Marchant Building, Suite 420
6701 San Pablo
Berkeley, CA 94720-7420
Phone: 510-642-9272
FAX: 510-642-8573
*http://calswec.berkeley.edu*

## Directory of Presenters and Participants

| | |
|---|---|
| Becky Antle, *MSSW, Ph.D.*<br>Assistant Research Professor<br>Kent School of Social Work<br>University of Louisville<br>Oppenheimer Hall<br>Louisville, KY 40292<br>502-852-2917<br>Fax: 502-852-0422<br>Becky.antle@louisville.edu | Becky F. Antle is the principal investigator on two grants and co-principal investigator on three grants. Her dissertation work involved evaluating a Children's Bureau 426 Grant focused on the development, delivery, and evaluation of ASFA Supervisory Team Training, particularly on engagement, assessment, case planning, and case work. She also evaluates child welfare training, the Casey Family-to-Family Initiative in Kentucky, and other training and intervention initiatives. |
| Anita Barbee, *MSSW, Ph.D.*<br>Professor, Distinguished Scholar<br>Kent School of Social Work<br>University of Louisville<br>Oppenheimer Hall<br>Louisville, KY 40292<br>502-852-0416<br>502-852-0422<br>anita.barbee@louisville.edu | Anita Barbee is the principal investigator or co-principal investigator on seven grants totaling $1.2 million in 2004–2005. She has spent the past 13 years evaluating child welfare training and outcomes for Kentucky's Cabinet for Health and Family Services. She is serving as the evaluator for the Children's Bureau Training and Technical Assistance Network of National Resource Centers for October 2004–September 2009. |

| | |
|---|---|
| Lorna Bell, *CQSW, Ph.D.*<br>Professor<br>St. George's Hospital Medical School<br>Kingston University School of Social Work<br>Kenry House, Kingston Hill<br>Kingston, Surrey, UKKT2 7LB<br>+44 0 208-547-8669<br>Fax: +44 0 208-547-8675<br>lbell@hscs.sghms.ac.uk | Although Lorna Bell also teaches, her primary role is conducting research covering a range of areas in child welfare, including multi-agency work in child protection and child welfare. She has undertaken an evaluation of child protection training across Scotland and a case study of an unusual practice placement undertaken by a social work student. She is currently preparing proposals to evaluate child protection training and to evaluate the extent social work students use their learning in practice on a one-day training program in a hospice. |
| Jane Berdie, *M.S.W.*<br>Child Welfare Consultant<br>435 South Gaylord St.<br>Denver, CO 80209<br>303-733-9532<br>Fax: 303-733-9532<br>jberdie@msn.com | Jane Berdie has been providing strategic planning and technical assistance to CalSWEC and the California Statewide Training & Education Committee's Macro Evaluation Team. She and Cindy Parry facilitated the development of the recently drafted Training Evaluation Framework for California. They, along with Dale Curry, are currently conducting a two-phase evaluation of the Pennsylvania Child Welfare Training program, under the auspices of the American Humane Association. |

| | |
|---|---|
| Kathleen Ja Sook Bergquist, *LCSW, Ph.D.*<br>Assistant Professor<br>School of Social Work<br>University of Nevada, Las Vegas<br>4505 Maryland Parkway<br>Box 455032<br>Las Vegas, NV 89154<br>707-895-2449<br>Fax: 707-895-4079<br>kathleen.bergquist@ccmail.nevada.edu | Kathleen Bergquist is a co-evaluator of the State of Nevada's Title IV-funded Child Welfare Training Academy, a curriculum with a mission to develop the capacity of workers' ability to provide solution-based, relationship-focused services. |
| Susan Brooks, *M.S.W.*<br>Program Director<br>Northern California Training Academy<br>UC Davis Extension<br>1632 Da Vinci Ct.<br>Davis, CA 95616-4860<br>530-757-8643<br>Fax: 530-752-6910<br>sbrooks@unexmail.ucdavis.edu | Susan Brooks has nearly 20 years of experience in social services, with expertise in substance abuse, collaboration, team-building, and supervision. For seven years, Ms. Brooks supervised the multi-disciplinary team for children's services in San Mateo County. She was also the founder and executive director of the San Mateo Perinatal Council, a nonprofit community collaborative. |
| Katharine Cahn, *MSW, Ph.D.*<br>Executive Director<br>Child Welfare Partnership<br>Portland State University<br>POB 751<br>Portland, OR 97207<br>503-725-8122<br>Fax: 503-725-8030<br>cahnk@pdx.edu | Katharine Cahn is the executive director of Oregon's Child Welfare Partnership, which offers training, graduate education, and research to advance child welfare in Oregon and beyond. Prior to assuming this position in October 2004, Dr. Cahn directed the Northwest |

| | |
|---|---|
| | Institute for Children and Families at the University of Washington School of Social Work for 19 years, where she developed both training and training and program evaluation programs. Her special interests are dynamics of systems change and the adoption of innovation in child welfare, addressing disproportionality in child welfare, and the social work practice of leadership, administration, and community organizing. |
| Soledad Caldera-Gammage, *M.S.W.*<br>Curriculum & Evaluation Specialist<br>Central California Child WelfareTraining Academy<br>112 Ron Ct.<br>Vallejo, CA 94591<br>cell: 559-246-5581<br>Fax: 707-647-1655<br>scaldera@csufresno.edu | Soledad Caldera-Gammage's responsibilities include the oversight and coordination of curriculum development, evaluation of curriculum and mentoring activities, and trainer development. She is currently working on the evaluation of the regional mentoring program of new workers in three Central Valley counties. |
| Sherrill Clark, *LCSW, Ph.D.*<br>Research Specialist<br>CalSWEC<br>University of California, Berkeley<br>School of Social Welfare<br>Marchant Building, Suite 420<br>6701 San Pablo<br>Berkeley, CA 94720-7420 | Dr. Sherrill Clark, former curriculum specialist and executive director of CalSWEC, has taught policy, practice, and research methods to master's students in the School of Social Welfare at UC Berkeley. Her primary research areas include work- |

| | |
|---|---|
| 510-642-4480<br>Fax: 510-642-8573<br>sjclark@berkeley.edu | force preparation and retention of specially educated MSW child welfare workers, and child welfare education. From 1998–2001, she was the project coordinator for the Child Welfare Fellows Project, a grant to enhance applied child welfare research and curriculum development. Dr. Clark participated in the California Department of Social Services Stakeholders' Group for three years on workforce preparation/support, fairness and equity, and the early intervention and differential response groups. She recently completed the 2004 statewide workforce study of California's public child welfare human resources. |
| James Coloma, *M.S.W.*<br>Research & Development Specialist<br>Public Child Welfare Training Academy<br>Academy for Professional Excellence<br>6505 Alvarado Rd., Suite 107<br>San Diego, CA 92120<br>619-594-3219<br>Fax: 619-594-1118<br>jcoloma@projects.sdsu.edu | James Coloma, a recent graduate of San Diego State University, is currently working on research and evaluation projects for the Public Child Welfare Training Academy and the Southern Area Consortium of Human Services. |

| | |
|---|---|
| John B. Cook, *M.A.*<br>Research Coordinator<br>Education & Training Partnership<br>Granite State College<br>117 Pleasant St.<br>Dolloff Building, 3rd Floor<br>Concord, NH 03301<br>603-271-6625<br>Fax: 603-271-4947<br>john.cook@cll.edu | John Cook is the research coordinator for a Title IV-E university/agency partnership that trains foster parents, adoptive parents, and residential providers statewide in New Hampshire. Mr. Cook is responsible for the development and implementation of foster care training evaluation activities. He recently concluded an evaluation of pre-license training for prospective foster parents. |
| Dale Curry, *LCSW, Ph.D.*<br>Assistant Professor<br>School of Family & Consumer Studies<br>Kent State University<br>Nixson Hall, POB 5190<br>Kent, OH 44242-0001<br>330-672-2998<br>Fax: 330-672-2194<br>dcurry@kent.edu | Dale Curry coordinates evaluation for a trainer development certificate program in collaboration with the Northeast Ohio Regional Training Center. He is the principal investigator for two statewide training evaluation projects in Ohio, and served as the consultant to the American Humane Association on its comprehensive evaluation of the Pennsylvania Competency-Based Child Welfare Training and Certification Program. Dr. Curry co-chairs the National Staff Development and Training Association Evaluation Committee and Assessment Committee of the North American Certification Project for child and youth workers. |

| | |
|---|---|
| Midge Delavan, *Ph.D.*<br>Training Coordinator<br>Utah Division of Child & Family Services<br>120 North 200 West, #225<br>Salt Lake City, UT 84103<br>801-538-4404<br>Fax: 801-538-3993<br>mdelavan@utah.gov | |
| Christy Derrick, *M.P.H.*<br>Training Evaluator<br>The Center for Child & Family Studies<br>The College of Social Work, Benson Building<br>University of North Carolina<br>Columbia, SC 29208<br>803-777-8494<br>Fax: 803-777-1366<br>cderrick@sc.edu | Christy Derrick is responsible for evaluating child welfare training offered by The Center for Child and Family Studies. She has participated in the evaluation of an interactive training program for independent living and has conducted level two evaluations on several training programs offered by The Center. She is working on helping develop a "basic" casework training for new workers of the South Carolina Department of Social Services and plans to evaluate the training at all four levels. |
| Bill Donnelly, *LCSW, M.P.A.*<br>Director<br>Inter-University Consortium<br>University of California, Los Angeles<br>Department of Social Welfare<br>3250 Public Policy Building<br>Box 951656<br>Los Angeles, CA 90095-1656<br>310-825-2811 | The Inter-University Consortium (IUC) and Department of Children and Family Services Training Project is a collaborative endeavor between the Los Angeles County Department of Children and Family Services (DCFS) and the graduate programs of social work at |

| | |
|---|---|
| Fax: 310-206-2716<br>donnelly@sppsr.ucla.edu | California State University, Long Beach; California State University, Los Angeles; the University of California, Los Angeles; and the University of Southern California. The IUC/DCFS Training Project is currently in its 11th year. During 2000–2001 the Training Project trained 6,971 staff and generated 155,174 training hours. |
| David Foster, *LCSW*<br>Project Director<br>Central California Child Welfare Training Academy<br>California State University, Fresno<br>5310 N. Campus Dr., M/S PH102<br>Fresno, CA 93740-8019<br>559-278-2258<br>Fax: 559-278-7229<br>david_j_foster@csufresno.edu | David Foster has 19 years' experience in public child welfare as a social worker, supervisor, and manager. He was the Title IV-E coordinator for the CSU, Fresno Social Work Program during1993–1998, founding member of the Central California Regional Training Academy Development and Implementation Project, and Academy director since 1998. He is married to Barbara and they have five children, (Adrian, Amanda, Nicole, Matt, Stephen) and one grandchild (Ryan). They are avid Ice Hockey fans (in Fresno, of all places…go figure!). |

| | |
|---|---|
| Todd M. Franke, *Ph.D.*<br>Professor<br>University of California,<br>Los Angeles<br>1100 Glendon Ave., Suite 850<br>Los Angeles, CA 90024<br>310-794-2583<br>Fax: 310-794-2728<br>tfranke@ucla.edu | |
| Walter Furman, *M. Phil.*<br>Academic Coordinator<br>University of California,<br>Los Angeles<br>Department of Social Welfare<br>3250 Public Policy Building<br>POB 951656<br>Los Angeles, CA 90095-1652<br>310-825-0852<br>Fax: 310-206-2227<br>wfurman@ucla.edu | Walter Furman's work focuses on research and evaluation of child welfare services, including: training evaluation for the Inter-University Consortium, recently including studies of Strength-Based and Concurrent Planning training, and evaluation of Prevention Initiative in California's small counties. |
| Elizabeth Gilman, *M.A., J.D.*<br>Curriculum Specialist<br>CalSWEC<br>University of California,<br>Berkeley<br>School of Social Welfare<br>Marchant Building, Suite 420<br>6701 San Pablo<br>Berkeley, CA 94720-7420<br>510-642-9273<br>Fax: 510-642-8573<br>egilman@berkeley.edu | Elizabeth Gilman joined CalSWEC in March 2002. Her responsibilities include curriculum development, evaluation, and planning associated with the statewide Title IV-E MSW Program and the newly launched BSW Program. She also serves as staff to the CalSWEC Board Curriculum Committee. Prior to joining CalSWEC, Ms. Gilman was an associate research scientist in the Yale University Psychology Department, where she served as an instructor and policy |

| | |
|---|---|
| | analyst specializing in child development and social policy issues. |
| Eileen Gambrill, *Ph.D.*<br>Professor<br>School of Social Welfare<br>University of California, Berkeley<br>120 Haviland Hall MC 7400<br>Berkeley, CA 94720<br>510-642-4450<br>Fax: 510-643-6126<br>gambrill@berkeley.edu | Eileen Gambrill is the Hutto Patterson Professor of Child and Family Studies at UC Berkeley's School of Social Welfare. Her research expertise includes professional decision-making, evidence-based practice, and the role of critical thinking. She has published extensively, including Critical Thinking in Clinical Practice, Critical Thinking for Social Workers: Exercises for the Helping Professions (with Len Gibbs), and Controversial Issues in Social Work Ethics, Values and Obligations (with Robert Pruger). She received the Pro Humanitate Literary Award from the North American Resource Center for Child Welfare in 2001 and 2003, in 2001 for the editorial "Honest Brokering of Knowledge and Ignorance" in the *Journal of Social Work Education* and in 2003 for two articles, "Evidence-based practice: Sea change or the emperor's new clothes?" in the *Journal of Social Work Education* and "A client-focused definition of social work practice" in the |

| | |
|---|---|
| | *Journal of Research on Social Work Practice.* |
| Shaaron Gilson, *LCSW*<br>Title IV-E Project Coordinator<br>School of Social Welfare<br>University of California, Berkeley<br>120 Haviland Hall MC 7400<br>Berkeley, CA 94720-7400<br>510-642-2424<br>Fax: 510-643-6126<br>shaaron@berkeley.edu | Shaaron Gilson has served as the Title IV-E Project Coordinator for UC Berkeley's School of Social Welfare for over 10 years. Her prior experience includes a career in public and private agencies in mental health and child welfare services, as well as academic appointments. Her most recent evaluation projects include the ongoing assessments of the Title IV-E training curriculum. |
| Michelle I. Graef, *Ph.D.*<br>Research Assistant Professor<br>The Center on Children, Families and the Law<br>University of Nebraska–Lincoln<br>121 South 13th St., Suite 302<br>Lincoln, NE 68588-0227<br>402-472-3741<br>Fax: 402-472-8412<br>mgraef1@unl.edu | Michelle I. Graef is an industrial/organizational psychologist on the faculty of The Center on Children, Families and the Law at the University of Nebraska–Lincoln. Under a contract with the Nebraska Health and Human Services System, Dr. Graef designs and implements the training evaluation procedures for the pre-service training for CPS workers in Nebraska. Her current work also includes the design and evaluation of new supervisor training, consultation with human services agencies on staff recruitment, selection and retention issues, and a study of CPS case decision-making. |

| | |
|---|---|
| Bart Grossman, *Ph.D.*<br>Adjunct Professor<br>School of Social Welfare<br>University of California, Berkeley<br>120 Haviland Hall, #7400<br>Berkeley, CA 94720-7400<br>510-642-0722<br>Fax: 510-643-6126<br>bfg@berkeley.edu | Bart Grossman is CalSWEC's founding director. He is adjunct professor and director of Field Education at UC Berkeley's School of Social Welfare. He provides consultation to Rutgers University regarding training and evaluation of New Jersey DYFS staff. He also provides extensive consultations regarding the use of Title IV-E funds for child welfare training and field education. |
| Norma Harris, *Ph.D.*<br>Director<br>Social Research Institute<br>College of Social Work<br>University of Utah<br>395 S. 1500 E, Rm. 130<br>Salt Lake City, UT 84112<br>801-581-3822<br>Fax: 801-585-6865<br>nharris@socwk.utah.edu | Norma Harris has extensive experience in the evaluation of child welfare programs. She is currently the principal investigator for the Title IV-E Grant with the Division of Child and Family Services. This grant includes a research/evaluation component. |
| Robert Highsmith, *Ph.D.*<br>Director of Assessment & Evaluation<br>NYC Administration for Children's Services<br>James Satterwhite Academy<br>492 First Ave., 5th Floor<br>New York, NY 10016<br>646-935-1484<br>Fax: 646-935-7282<br>robert.highsmith@dfa.state.ny.us | Robert Highsmith is responsible for the design and implementation of instruments and systems used by the Satterwhite Academy, the training arm of New York City's Child Welfare Administration, for assessing the readiness of case workers to deliver child welfare services. He is also responsible for evaluating the effectiveness of the academy's training. |

| | |
|---|---|
| Claudia “Kyle” Hoffman<br>Manager<br>OCWTP Evaluation Project<br>Institute for Human Services<br>Gwinn House<br>1706 E. Broad St.<br>Columbus, OH 43203-2039<br>614-251-6000<br>Fax: 614-251-6005<br>khoffman@ihs-trainet.com | As a training coordinator with the Institute for Human Services, Kyle Hoffman chairs the Ohio Child Welfare Training Program (OCWTP) Evaluation Work Team, including county child welfare staff, Regional Training Center coordinators, the Ohio Department of Job and Family Services staff, and research consultants. The work team is in the last year of a four-year effort to develop a comprehensive system to evaluate knowledge acquisition, knowledge comprehension, skill demonstration, and skill transfer that occurred as a result of OCWTP Training. |
| Teresa Hubley, *M.P.A., Ph.D.*<br>Research Associate II<br>Institute for Public Sector Innovation<br>Edmund S. Muskie School of Public Service<br>University of Southern Maine<br>295 Water St.<br>Augusta, ME 04330<br>207-626-5292<br>Fax: 207-626-5210<br>teresa.a.hubley@maine.gov | Teresa Hubley specializes in child welfare projects and general training evaluation. She has a doctorate in cultural anthropology, specializing in medical anthropology, and a Master’s in Public Administration focusing on health policy, both from Syracuse University. |
| Henry Ilian, *D.S.W.*<br>Training Evaluator<br>NYC Administration for Children’s Services | Henry Ilian has been involved since 1987 in evaluation and testing at the James Satterwhite Academy for Child Welfare |

| | |
|---|---|
| James Satterwhite Academy<br>for Child Welfare Training<br>492 First Ave., 5th Floor<br>New York, NY 10016<br>646-935-1410<br>Fax: 646-935-1782<br>henry.ilian@dfa.state.ny.us | Training. He has been developing competency-based measures to accompany New York City's adaptation of the New York State Common Core training system for child welfare workers. He also teaches research at the Columbia University School of Social Work. |
| Susan E. Jacquet, *Ph.D.*<br>Research Specialist<br>CalSWEC<br>University of California, Berkeley<br>School of Social Welfare<br>Marchant Building, Suite 420<br>6701 San Pablo<br>Berkeley, CA 94720-7420<br>510-643-9846<br>Fax: 510-642-8573<br>sjacquet@berkeley.edu | Susan E. Jacquet works with Dr. Sherrill Clark on CalSWEC's research component, including ongoing surveys of California's MSW students and Title IV-E MSW graduates who have completed payback work in child welfare, and the development of new research initiatives on outcomes for child welfare and the efficacy of the Title IV-E program. Dr. Jacquet is also responsible for coordinating CalSWEC's funded research process from RFP through review of proposals. |
| Phyllis Jeroslow, *M.F.A., MFT*<br>Training & Curriculum Specialist<br>CalSWEC<br>University of California, Berkeley<br>School of Social Welfare<br>Marchant Building, Suite 420<br>6701 San Pablo<br>Berkeley, CA 94720-7420 | Phyllis Jeroslow assists in the development and dissemination of statewide standards and curriculum for core and ongoing child welfare training. Previously, she co-developed the family conferencing program and the protocols for team decision-making for San Francisco |

| | |
|---|---|
| 510-643-5440<br>Fax: 510-642-8573<br>pjero@berkeley.edu | County's Family & Children's Services. |
| Barrett Johnson, *M.S.W., LCSW*<br>RTA Coordinator<br>CalSWEC<br>University of California, Berkeley<br>School of Social Welfare<br>Marchant Building, Suite 420<br>6701 San Pablo<br>Berkeley, CA 94720-7420<br>510-643-5484<br>Fax: 510-642-8573<br>barrettj@berkeley.edu | Barrett Johnson oversees CalSWEC's coordination of statewide training efforts, including development and implementation of a common core curriculum and strategic planning for a statewide training evaluation system. He is involved in planning the training efforts for California's Outcomes and Accountability System. He has worked for many years with urban children and families, with an emphasis on intervention in cases of child sexual abuse. |
| Susan N. Kanak, *M.B.A.*<br>Policy Associate II<br>National Child Welfare Resource Center for Organizational Improvement<br>Edmund S. Muskie School of Public Service<br>University of Southern Maine<br>400 Congress St.<br>POB 15010<br>Portland, ME 04112<br>207-780-5840<br>Fax: 207-780-5817<br>skanak@usm.maine.edu | Susan Kanak has 28 years of government and non-profit agency management, performance auditing, curriculum development, and training experience. Currently serving as a policy associate with the National Child Welfare Resource Center for Organizational Improvement and the Institute for Child and Family Policy, Ms. Kanak has expertise in curriculum design and delivery, system reviews, adult education, and implementing change in the public sector. She has provided technical assistance on training |

| | |
|---|---|
| | to public child welfare agencies in several states, including California, Iowa, Rhode Island, Kentucky, Colorado, New Mexico, Wisconsin, and Arkansas, and Cuyahoga County, Ohio. With several public child welfare agencies, she recently developed, published, and trained agency trainers to implement three curricula. Her most recent publication is the *Training System Assessment Guide for Child Welfare Agencies.* |
| Eileen Lally, *LCSW, Ed.D.*<br>Manager<br>Family & Youth Services Training Academy<br>University of Alaska, Anchorage<br>4500 Diplomacy Dr., Suite 430<br>Anchorage, AK 99518<br>907-786-6720<br>Fax: 907-786-6735<br>emlally@uaa.alaska.edu | Eileen Lally is in her seventh year of a partnership at the State of Alaska's Division of Family and Youth Services, providing training to statewide social work staff that utilizes Title IV-E funding. |
| Judith Lefler, *RN, PHN, CLNC*<br>Assistant Director<br>Bay Area Child Welfare Training Academy<br>2201 Broadway, Suite 100<br>Oakland, CA 94612<br>707-751-0419<br>Fax: 707-751-1439<br>judithlefler@aol.com | Judith Lefler is the assistant director of the Bay Area Child Welfare Training Academy with oversight of all deliverables and capacity building activities in 12 Bay Area counties. |

| | |
|---|---|
| Barbara Legier<br>Clinical Program Planner III<br>Division of Child & Family Services<br>State of Nevada<br>711 E. 5th St.<br>Carson City, NV 98701<br>775-684-4407<br>Fax: 775-684-4457<br>blegier@dcfs.state.nv.us | |
| Elizabeth W. Lindsey, *M.S.W., Ph.D.*<br>Associate Professor<br>Department of Social Work<br>University of North Carolina–Greensboro<br>POB 26170<br>Greensboro, NC 27402<br>336-334-5225<br>Fax: 336-334-5210<br>betsy_lindsey@uncg.edu | Elizabeth Lindsey's ongoing work has involved the use of embedded evaluation strategies to assess transfer of learning of supervisory knowledge and skills associated with implementation of North Carolina's Multiple Response System (MRS), providing technical assistance to training vendors in development of evaluation plans for their deliverables, and revision of the Participant Satisfaction Form to include items that have been shown to predict transfer of learning on the job. Her project has also been responsible for the ongoing development and validation of five knowledge assessments for new child welfare workers. This spring Dr. Lindsey will begin to develop a process for curriculum analysis, to compare various curricula for consistencies and inconsistencies as well as |

| | |
|---|---|
| | possible recommendations for policy and practice standards, especially with respect to MRS issues. |
| Marty Lowrey, *M.S.W., LCSW*<br>Director of Training<br>Child Welfare Partnership<br>Portland State University<br>4061 Winema Place NE<br>Salem, OR 97305<br>503-315-4373<br>Fax: 503-399-6439<br>lowreym@pdx.edu | Marty Lowrey is the director of Training for the Child Welfare Partnership at Portland State University's School of Social Work. Prior to her appointment to this position, Ms. Lowrey served as lead CPS worker and CPS worker in Oregon's Benton County Office. Her practice history also includes directing a community youth center in Chico, California, as well as serving as director of Training for a psychiatric residential care center in Tennessee. |
| Brad Lundahl, *Ph.D.*<br>Professor<br>College of Social Work<br>University of Utah<br>395 S. 1500 E., #310<br>Salt Lake City, UT 84112<br>801-581-4570<br>Fax: 801-585-3219<br>brad.lundahl@socwk.utah.edu | New faculty member Brad Lundahl is interested in training and evaluation related to delivery of evidence-based practice models. He is particularly interested in training and evaluation related to parent-training and improving worker-client relationships. |
| Susan Mason, *LCSW, Ph.D.*<br>Professor & Senior Education Consultant<br>Social Work Education Consortium<br>Yeshiva University, SSW | This is year three of the NY State Social Work Child Welfare Consortium's evaluation of incentive course work at the state's schools of social work to gauge the |

| | |
|---|---|
| 2495 Amsterdam Ave., Suite 926<br>New York, NY 10033<br>212-960-0806<br>Fax: 212-960-0821<br>masonse@yu.edu | transfer of learning; year two of its supervisor study to learn how supervisors value and/or observe workers' transfer of learning; and year two of its student unit evaluation of a designated child welfare unit at a children's sex abuse clinic. |
| Chris Mathias, *M.S.W.*<br>Director<br>CalSWEC<br>School of Social Welfare<br>University of California, Berkeley<br>Marchant Building, Suite 420<br>6701 San Pablo<br>Berkeley, CA 94720-7420<br>510-642-7490<br>Fax: 510-642-8573<br>cmathias@berkeley.edu | Chris Mathias began her work with CalSWEC in March2000, when she headed the Regional Training Academy Coordination Project. As CalSWEC director, she leads the development and evaluation of the Title IV-E Program for public child welfare and the Regional Training Academy Coordination Project. She heads a consortium that includes 17 universities, the County Welfare Directors Association, the Mental Health Directors Association, the four Regional Training Academies, the Inter-University Consortium in Los Angeles and the California Department of Social Services. For 14 years prior to joining CalSWEC, Ms. Mathias worked primarily in the private, non-profit sector with children in out-of-home care. During that period, she developed curriculum, training, and quality assurance methods |

| | |
|---|---|
| | for practice for direct care workers, clinicians, and administrators. |
| Timothy McCarragher, *LISW, Ph.D.*<br>Assistant Professor<br>School of Social Work<br>University of Akron<br>Akron, OH 44325<br>330-972-5976<br>Fax: 330-972-5739<br>mccarra@uakron.edu | Timothy McCarragher teaches graduate research courses. For the past three years, he has collaborated with the Institute for Human Services and the Ohio Child Welfare Training Program (OCWTP) Evaluation Work Team, which includes county child welfare staff, Regional Training Center coordinators, the Ohio Department of Job and Family services staff, and research consultants. The work team is in the last year of a four-year effort to develop a comprehensive system to evaluate knowledge acquisition, knowledge comprehension, skill demonstration, and skill transfer that occurred as a result of OCWTP Training. |
| James Midgley, *Ph.D.*<br>Dean and Specht Professor<br>School of Social Welfare<br>University of California, Berkeley<br>120 Haviland Hall MC 7400<br>Berkeley, CA 94720-7400 | James Midgley has served as the Harry and Riva Specht Professor of Public Social Services and dean of the School of Social Welfare at UC Berkeley since 1997. He will complete his term in June 2005. He previously was dean of the School of Social Work at Louisiana State University and subsequently as associate |

| | |
|---|---|
| | vice-chancellor. Prior to LSU, he taught at the London School of Economics and the University of Cape Town. He has published widely on issues of social policy, international social work and welfare, and social development. Recent books include: *Controversial Issues in Social Policy* (2003), *Social Policy for Development* (2004), and *Lessons from Abroad: Adapting International Social Welfare Innovations* (2004). |
| Salvador Montana<br>Faculty<br>Department of Social Work<br>California State University, Fresno<br>5310 N. Campus Dr. M/S PH102<br>Fresno,CA93775<br>559-278-8581<br>559-278-7191<br>smontana@csufresno.edu | |
| Michael A. Nunno, *D.S.W.*<br>Senior Extension Associate<br>Family Life Development Center<br>Cornell University<br>Beebe Hall<br>Ithaca, NY 14853-4401<br>607-254-5127<br>Fax: 607-255-4837<br>man2@cornell.edu | Michael Nunno is the principal investigator of Cornell University's Residential Child Care Project, and is responsible for the evaluation of its training and technical assistance programs. The project trains residential child care workers in non-confrontational limit-setting strategies to reduce the levels |

| | |
|---|---|
| | of aggression and critical incidents in residential child care, juvenile correction, mental health, and mental retardation facilities, as well as training institutional abuse prevention, investigation, and remediation. These training and technical assistance programs for governmental and nongovernmental facilities are conducted throughout North America, the United Kingdom, Ireland, Bermuda, and Australia. |
| Kirstin O'Dell, *M.S.W.*<br>Researcher<br>Child Welfare Partnership<br>Portland State University<br>POB 751<br>Portland, OR 97207-0751<br>503-725-8071<br>Fax: 503-725-8030<br>odellk@pdx.edu | Kirstin O'Dell is the training evaluator for the Child Welfare Partnership at Portland State University. She has seven years of research experience, primarily in the fields of child welfare and early childhood. |
| Sandra Owens-Kane, *LCSW, Ph.D.*<br>Assistant Professor/Trainer/ Evaluator<br>School of Social Work<br>University of Nevada, Las Vegas<br>4505 Maryland Parkway<br>Box 455032<br>Las Vegas, NV 89154-5032<br>702-895-2898<br>Fax: 702-895-4079 | Sandra Owens-Kane has a B.A. and M.S.W. from the University of Nevada, Las Vegas and a Ph.D. from UC Berkeley. During the past 15 years, she has utilized her skills as a licensed clinical social worker, researcher, trainer, and educator to provide and evaluate outcomes related to topics including, but not limited to, cultural competence of child welfare workers, child |

| | |
|---|---|
| sandra.owens@ccmail.nevada.edu | welfare organizational culture, and child abuse and neglect. |
| Cynthia Parry, *Ph.D.*<br>Consultant<br>C.F. Parry Associates<br>520 Monroe Ave.<br>Helena, MT 59601<br>406-449-3263<br>Fax: 406-449-3263<br>cfparry@msn.com | Cynthia Parry specializes in program evaluation in human services. She has over 20 years of experience in evaluation in child welfare, juvenile justice, and education. Dr. Parry is a former director of Program Analysis and Research at American Humane Association. She is currently involved in evaluations of a trauma treatment project in Colorado, and evaluations of training effectiveness in California, Pennsylvania, and Colorado. She is a co-facilitator for a strategic planning process for training evaluation in California. Dr. Parry is a co-author of the National Staff Development and Training Association (NSDTA) publication *Training Evaluation in the Human Services* and has presented at numerous NSDTA conferences beginning in 1994. |
| Bruce Parsons, *Ph.D.*<br>Associate Professor, Research<br>Social Research Institute<br>College of Social Work<br>University of Utah<br>395 S. 1500 E., #111<br>Salt Lake City, UT 84112<br>801-581-4858 | Bruce Parsons is currently training child welfare professionals in foster care, family preservation, and post-adoption services. He teaches the graduate-level practice classes for Title IV-E students and students focused on the |

| | |
|---|---|
| Fax: 801-585-6865<br>bruce.parsons@socwk.utah.edu | public service domain. Dr. Parsons is implementing a practice model change in post-adoption services and preparing to evaluate its efficacy. |
| Joan Pennell, *M.S.W., Ph.D.*<br>Professor & Department Head<br>Department of Social Work<br>North Carolina State University<br>C.B. 7639<br>Raleigh, NC 27695<br>919-513-0008<br>Fax: 919-515-4403<br>jpennell@ncsu.edu | Joan Pennell is the principal investigator of the North Carolina Family-Centered Meetings Project and previously directed the North Carolina Family Group Conferencing Project. Before returning to the U.S., she served as a principal investigator for a Newfoundland & Labrador (Canada) demonstration of family group conferencing in situations of child maltreatment and domestic violence. She co-authored *Community Research as Empowerment* (Oxford University Press), *Family Group Conferencing: Evaluation Guidelines* (American Humane Association), and *Widening the Circle: The Practice and Evaluation of Family Group Conferencing with Children, Young Persons, and Their Families* (NASW Press). |

| | |
|---|---|
| Robin Perry, *Ph.D.*<br>Associate Director<br>Institute for Health & Human Services Research<br>School of Social Work<br>Florida State University<br>2300 University Center<br>Building C<br>Tallahassee, FL 32306-2702<br>850-645-5769<br>Fax: 850-644-8331<br>reperry@fsu.edu | Robin Perry is the principal investigator for Florida State University's Title IV-E stipend program and associated evaluation efforts. A former CalSWEC research associate, he has published on training and training evaluation topics. Dr. Perry's recent research within the child welfare field has focused upon the following areas: the examination of the influence of educational background upon CPI and CPS workers' job performance in Florida; the development of a task-analysis tool/model for CPS workers; a cost analysis of integrating select information technology into child protection investigations in Florida; and the development of a statistical/ actuarial model for the equitable distribution of child welfare service money in Florida. |
| Megan E. Potter, *Ph.D.*<br>Research Assistant Professor of Psychology<br>Center on Children, Families and the Law<br>University of Nebraska–Lincoln<br>121 S. 13th St., Suite 302<br>Lincoln, NE 68588<br>402-472-9812<br>Fax: 402-472-8412 | Megan E. Potter is jointly responsible for conducting training evaluation and competency assessment projects designed to improve the selection, training, performance, and retention of protective service workers in Nebraska. Her recent training evaluation projects include: continual development and |

| | |
|---|---|
| mpotter2@unl.edu | updating of written knowledge tests; development of skills evaluations, such as for testifying and computer documentation; and development of a reporting system to regularly provide trainee performance feedback to trainees and their supervisors. |
| E. Douglas Pratt, *D.S.W., LCSW*<br>Staff Development Specialist<br>Policy–Practice Resources Inc.<br>3786 Evans Rd.<br>Atlanta, GA 30340<br>770-723-9105<br>Fax: 770-723-9105<br>ppr-edp@juno.com | E. Douglas Pratt has been using Skills Progress Assessment, Quality Services Review, and other specialized methods to help six states evaluate outcomes of training and development in Family-Centered Engagement, Family Team Meeting Process, Supervisory Coaching, and Community-Based Practice. He has also conducted formative and summative evaluations of Training Academies in two states. He was a trainer of trainers and a performance evaluator for Alabama's model child welfare reform from 1991 through 2002 (Making Child Welfare Work, Bazelon, 1998). Dr. Pratt co-designed the MAPP curricula (GPS-MAPP and CSA-MAPP) used by Alabama and 20 other states, Israel, and the Netherlands. He has also trained numerous MAPP trainers. |

| | |
|---|---|
| Basil Qaqish, *Ph.D.*<br>Research Scientist<br>University of North Carolina–Greensboro<br>1673 Thompson Dr.<br>Winston-Salem, NC 27127<br>336-315-7044<br>Fax: 336-334-5210<br>bfqaqish@uncg.edu | Basil Qaqish is working on the transfer of learning survey in North Carolina and assessment instrument development for social work training for the following core topics: (1) the effects of separation, loss, and attachment, (2) medical aspects of child abuse, (3) child development and families at risk, and (4) legal aspects of child welfare in North Carolina. |
| Greg Rose, *M.S.W.*<br>Chief<br>Office of Child Abuse Prevention<br>California Department of Social Services<br>744 P St., MS 11-82<br>Sacramento, CA 95814<br>916-651-6100<br>Fax: 916-651-6328<br>greg.rose@dss.ca.gov | Greg Rose recently became chief of California's Office of Child Abuse Prevention. He formerly served as bureau chief of the Resources Development and Training Support Bureau at the California Department of Social Services (CDSS), and represented CDSS as the co-chair of California's Statewide Training and Education Committee. |
| Marcia Sanderson, *LMSW*<br>Director<br>Protective Services Training Institute<br>School of Social Work<br>University of Texas at Austin<br>1 University Station D3500<br>Austin, TX 78712-0358<br>512-471-0521<br>Fax: 512-232-9585<br>MSanderson@mail.utexas.edu | Marcia Sanderson has been the director of the Protective Services Training Institute (PSTI) since 1993. From 1999 to 2002, she was also the director of the Child Welfare Education Project, a Title IV-E stipend program at the University of Houston Graduate School of Social Work. In 2002, she became |

| | |
|---|---|
| | full-time PSTI director and relocated to the University of Texas at Austin. While at the University of Houston, she taught grant writing in the social work master's program and program evaluation as part of continuing education workshop in program planning and proposal writing. |
| Melanie Silveria<br>Evaluation Coordinator<br>Northern California Training Academy<br>University of California, Davis<br>1632 Da Vinci Ct.<br>Davis, CA 95616-4860<br>530-757-8643<br>Fax: 530-752-6910<br>msilveria@unexmail.ucdavis.edu | |
| Becky L. Thomas, *M.S.W., LSW*<br>Senior Training Officer<br>Cuyahoga County Department of Children & Family Services<br>3955 Euclid Ave., #340E<br>Streetsboro, OH 44115<br>216-881-5316<br>Fax: 216-432-3516<br>bthomas@www.cuyahoga.oh.us | Becky L. Thomas has been with the Cuyahoga County Department of Children & Family Services in Cleveland, Ohio, for over three years. She works on several committees addressing various methods of evaluating child welfare training systems. Ms. Thomas has over 10 years of experience working with children and families in need. |

| | |
|---|---|
| Charlene Urwin, *LCSW, Ph.D.*<br>Curriculum Director & Site Manager<br>Protective Services Training Institute<br>School of Social Work<br>University of Texas at Austin<br>1 University Station D3500<br>Austin, TX 78712<br>512-471-0560<br>Fax: 512-232-9585<br>curwinn@mail.utexas.edu | Charlene Urwin has directed training needs assessments and competency projects in child and adult protection and recently completed a performance dimension approach in child care licensing. She has extensive experience in teaching and directing social work education programs. |
| Quinn Wilder, *M.S.W.*<br>Youth Worker Training Program Manager<br>Youth Work Learning Center<br>University of Wisconsin–Milwaukee<br>161 W. Wisconsin Ave., Suite 6000<br>Milwaukee, WI 93203<br>414-227-3172<br>Fax: 414-227-3224<br>qwild@uwm.edu | Quinn Wilder is in a doctoral program for Urban Education with an Educational Psychology Concentration at the University of Wisconsin–Milwaukee. He has helped manage an ongoing multi-grant training initiative as senior outreach specialist with the Child and Youth Care Learning Center, part of the Division of Outreach and Continuing Education at the University of Wisconsin–Milwaukee. |
| Leslie W. Zeitler, *LCSW*<br>Training & Evaluation Specialist<br>CalSWEC<br>University of California, Berkeley<br>School of Social Welfare<br>Marchant Building, Suite 420<br>6701 San Pablo<br>Berkeley, CA 94720-7420 | Leslie W. Zeitler coordinates and implements CalSWEC's statewide training and evaluation efforts, and provides technical assistance to the California Regional Training Academies and counties. Prior to joining CalSWEC, Ms. Zeitler provided direct social work |

| | |
|---|---|
| 510-643-6400<br>Fax: 510-642-8573<br>lzeitler@berkeley.edu | services for six years to low-income children and families through the San Francisco-based Legal Services for Children. She is a graduate of the Coro Fellowship Program in Public Affairs, and also teaches family assessment and risk management skills to Head-Start direct services staff at annual Head-Start conferences. |