

HELPING CHILD WELFARE WORKERS LEARN INTERVIEWING SKILLS

A Research Report

**Colleen Friend, PhD, LCSW
Director, Child Abuse and Family Violence Institute
California State University, Los Angeles
Department of Communication Disorders
King Hall, Room 109B
California State University, Los Angeles
5150 State University Drive
Los Angeles, California 90032
CFriend@calstatela.edu**

TABLE OF CONTENTS

Executive Summary	1
Introduction/Literature Review	5
Research Design and Methods	9
<i>Research Background Questions and Hypotheses... 10</i>	
<i>Methodology... 14</i>	
<i>Sample Description and Procedures....23</i>	
<i>Data Collection...27</i>	
Results	33
<i>Developmental Sample...33</i>	
<i>Core Study...36</i>	
<i>Discussion of Limitations and Findings...41</i>	
<i>Policy Implications and Future Research...43</i>	
References	46
Tables	
Table 1: Hypotheses Summary...14	
Table 2: Experimental Instruments...22	
Table 3: Key Training Interventions Summarized...25	
Table 4: Interrater Reliability Correlations, per Segment of CWDVISS, for the Developmental Sample...19	
Table 5: Developmental Sample Scores: FRS, CWDVISS, and PPIF Scores...35	
Table 6: Intercorrelations Between Instrument Scores for the Developmental Sample...35	
Table 7: Pretest and Posttest Core Study Scores with Effect Size...38	
Table 8: Practice Quality Rating Rubric for PPQ Rating ...31	
Table 9: Extreme Subjects' per Segment Rank Ordered Skill Scores From CWDVISS...28	
Table 10: PPQ Score "10 Things List" and Audiotape Ratings for Extreme Scores...29	

HELPING CHILD WELFARE WORKERS LEARN INTERVIEWING SKILLS

EXECUTIVE SUMMARY

This CalSWEC research report explores the first two aspects of the Phase II research. It: (a) describes a training, how skill development was measured using instruments and a standardized client (SC); and (b) analyzes how an expert and a novice demonstrated different skills levels. It is also a useful companion for Module III of the *From the Bottom Up* curriculum because it: provides greater detail about actual skill development as a result of training, discusses what we can learn from standardized clients as client proxies, explores the use of a practice rubric in conducting qualitative analysis, and examines how novices and experts (extreme scorers) demonstrate skill differentially. Beyond this, some components of Module V, drawn from the same curriculum, refer to information in this CalSWEC research report.

This report begins by establishing the need for a report such as this: Good interviewing skills are central to the public child welfare worker's duty to determine if a child is safe and if a parent can protect, while simultaneously empowering the parent to take on a more assertive role as a provider and protector. While acknowledging that it is often difficult to do all of this, it is precisely what is expected if, however, unrealistic and inconsistently attained. Moving beyond role play, this report establishes the need for utilizing standardized clients (SCs) as client proxies and as necessary training tools.

More specifically, this report summarizes the research done with a small sample (N = 15) of public child welfare employee subjects who interviewed two SCs pre- and

posttraining. Referred to as “the core experiment,” this is considered a quasi-experimental design. The training in interviewing at the intersection of Child Abuse and Domestic Violence was the independent variable. The report describes the complex process of developing an instrument to capture and measure a “live” public child welfare interview using SCs as no suitable instrument existed. The instrument developed for this purpose is called the Child Welfare Domestic Violence Interview Skill Set (CWDVISS). In the attempt to explore concurrent construct validity for the instrument, utilizing a small developmental sample ($n = 6$), correlations among instruments did not reach statistical significance at $p=.05$. This being an explanatory study, the relationships were explored where p values dropped below .10. Once the validity and reliability of the CWDVISS was examined with a small, separate developmental sample, the subjects’ ($N = 15$) skill demonstration in the core experiment was measured with this instrument. Using t -tests, it was determined that the training made a significant difference in the posttest scores, the mean score being approximately 28 points higher. This led to speculation that the training had an impact on the subjects’ practice. To explore that further, a qualitative content analysis of the highest and lowest scorers (expert vs. novice) was conducted to assess what they were doing differently. In order to do that, a Practice Qualitative Rating was developed (adapted from the literature) that was used to rate both the subjects’ *to do* list, as well as the interview audiotape. That rating was followed by a content analysis that revealed the expert’s very different pattern of conducting early rapport and trust building before exploring difficult subjects; furthermore he consistently delivered non-judgmental feedback. While the novice

trainee was the most improved in the posttest, he was unfortunately brief, leading to speculation that he was attempting to utilize large amounts of new information. He was not consistently non-judgmental, which might have been a new skill he was attempting to practice.

There were several limitations in this study that need to be considered in any discussion of findings. Primarily, it was an exploratory, quasi-experimental design, precluding any strong statement of relationship among the variables. The experiment lacked the resources to measure all the theoretical constructs. Correlations among instruments (used as a validity check) in the developmental sample did not reach statistical significance at the $p=.05$ level. Furthermore, it was difficult to recruit and retain a sample of actual public child welfare workers. These small, voluntary, convenience samples cannot be said to represent the PCW workforce.

Given those understandings, this pilot study is the first study with child welfare workers that actually provides a window into what they do with adult clients (represented by SCs as proxies) and how a brief training might improve that interaction significantly. The study also proposed a methodology for demonstrating and measuring subject skill with a standardized client. As shown here and elsewhere, this method for interview training can be a powerful evaluation tool to improve other kinds of training, especially for the skills needed in social work practice. The instrument developed demonstrated some preliminary reliability and validity, but reuse of this instrument should be preceded with factor analysis, to reduce its complexity and make the coding less labor intensive. In this study it appeared that experts and novices demonstrate

skills differently, thus they may learn differently, which has the potential to inform social work teaching practice. Finally, the study established the complexity involved in PCW interviews, and how proficiency requires nuanced and substantive skill demonstration. After demonstration, given a large disparity between the posttest mean score and the perfect score, the issue of clinical adequacy gets raised. The inclusion of SCs as client proxies allows workers to experiment and make mistakes without doing any actual harm. It is this ability to take the physical step of actually testing that helps the public child welfare worker build feelings of competence and self efficiency that allow him/her to become a better interviewer. We briefly describe the development of a practice rubric to measure the skill demonstration and the cognitive strategies used to train the workers to be better interviewers at the intersection of Domestic Violence and Child Abuse.

A voluntary sample of PCW subjects was offered a 1-day training using Zull's (2002) approach to learning, the Baldwin and Ford model (1988), and cognitive training techniques focusing on how to interview parents. The SCs were used to help evaluate the subjects' transfer of training. This is a skill that must be measured by viewing the interviewer's behavior. The use of SCs allowed for practice and evaluation without the concomitant risk of harm to real clients. An instrument was developed to capture and measure the interaction between the PCW worker and the SC. The SCs were asked to provide feedback regarding their interactions with the PCW subjects. Overall, it was hypothesized that training will make a significant difference in skill demonstration. Beyond the hoped-for difference, levels of clinical change were also assessed. The study's findings may become part of this specific public child welfare agency's efforts to

improve and extend its training, retain workers, and address potential consumer complaints.

INTRODUCTION/LITERATURE REVIEW

If children are our future, then the state has an interest in the creation of a productive citizenry from the ranks of its children. Thus, it has a motive for protecting its youngest citizens from those who would do them harm. The latest national incidence study of child abuse and neglect identified over 1.5 million victims of child abuse or neglect and indicated that 78% of the perpetrators were parents, 10% were other relatives, and the remainder were unrelated (Sedlack & Broadhurst, 1996). With the passage of the Child Abuse Prevention and Treatment Act (CAPTA) of 1974, the federal government took the initiative in establishing a model statute for state child protection programs that mandated standard methods for reporting and investigating child abuse and neglect (Costin, Karger, & Stoesz, 1996).

Some scholars have argued that the vagueness of the definition of child abuse and the reasonable suspicion that reporting thresholds are high may be responsible for dramatic overreporting (Besharov, 1987). In fact, National Incidence Surveys have uncovered nearly 50% more child maltreatment victims than those already known to Child Protective Services (CPS) agencies (Zellman & Fair, 2002). The sentinel survey methodology used in the National Incidence Study included community professionals from non-CPS agencies trained as “lookouts” for maltreated children during the review period. Their discovery of more maltreated children than those reported to CPS lends

support to Finkelhor's (1993) contention that the essential problem is still underreporting, not overreporting.

Central to the federal mandate to investigate child abuse and neglect is the child welfare worker's ability to interview the parents who come to their attention. But there is an inherent adversarial stance between the interviewer and interviewee, since these interviews are the mechanism the state uses to gather information in the exercise of its social control function. Parents who are referred for the coexisting problems of possible child abuse and domestic violence are particularly problematic, and were the focus of this research. Parents who come to the child welfare system's attention correctly perceive the stakes as high, despite the fact that relatively few children are removed from their homes in these encounters (Britner & Mossler, 2002). According to this study, less than 16% of substantiated child victims were removed from their home. However, rates appear to be on the rise as indicated by statistics gathered by the U.S. Department of Health and Human Services (HHS, 2004). The 2003 removal rate was 19%. Practice wisdom validates the likelihood that the child welfare worker will be given very limited disclosures about facts or feelings from defensive parents, making these interviews difficult to conduct. In this encounter, the interviewer must explain the purpose, build rapport, ask a series of difficult questions, de-escalate anger, and manage his/her own emotions. The challenge of balancing all these tasks simultaneously has the potential to hijack the ultimate goal, which is to determine if the child is safe or in need of protection. The temperament of the interviewer must be accounted for as well. An interviewer who avoids conflict or becomes overly aligned with

parents could run the risk of making a false negative assessment on safety issues, while an interviewer who becomes emotionally engaged with hostile parents could, conversely, make a false positive assessment. Some child welfare workers seem to have a talent or skill set for minimizing the power differences; others are unaware, unwilling, or poorly skilled in doing so. It appears that this is a skill set that has not been well identified, trained, practiced, or even evaluated. The only published research on child welfare worker interviewing skills has focused on interviewing child sexual abuse victims, with adults acting as though they were children (Brittain, 2000; Freeman & Morris, 1999; Stevenson, Leung, & Cheung, 1992). Virtually no studies have been published that measure what child welfare workers actually do with parent clients.

Public child welfare (PCW) was once the exclusive domain of Master of Social Work (MSW) trained social workers (National Association of Social Workers, as cited in Perry, 2006). Over time, however, the education and experience requirements for child welfare workers have been considerably reduced, with no MSW requirement at this time, and with the job classification transformed into a generic title that lacks professional specificity or identity. This is further complicated by the variability of requirements from state to state.

A 1981 national survey of child welfare workers' educational backgrounds reported that only 26% of the caseworkers who participated in the survey held a bachelor's degree in social work (Vinokur-Kaplan & Hartman, 1986, as cited in Zell, 2006). Zell could find no recent large-scale studies describing caseworker qualifications. Assuming that the public child welfare workforce typically has little experience and a

variety of educational backgrounds, it is not clear that reinstating the MSW requirement alone would improve the quality of interviewing in the child welfare system. Despite the tradition of training to practice-specific skill sets in MSW programs, social work teachers, researchers, field supervisors, and clients have lamented the little attention given to the practice and evaluation of interviewing skills (Badger & MacNeil, 2002; Carillo, Gallart, & Thyer, 1993; Schinke, Smith, Gilchrist, & Wong, 1978; Linsk & Tunney, 1997). MSW students themselves have reported feeling ill-prepared to negotiate the complexities of the interview situation (Carillo et al., Schinke, Blythe, Gilchrist, & Smith, 1980). Although there has been a promising coordinated effort to draw down one of the last federal entitlements (Title IV-E) for the specific preparation of PCW workers in MSW programs, the jury is still out on the overall effectiveness of an approach that reprofessionalizes the workforce (California Social Work Education Center, 1999).

It seems then that the most effective remedy to address this problem is on-the-job training, yet that brings its own set of issues and complexities. With the availability of Title IV-E funding to support training, there have been a variety of trainings offered to the PCW workforce; however, their content, linked with practice and transfer effectiveness, has been relatively unevaluated (McDonald & McCartney, 1999). Although it is suspected that some of this information is available within state agencies and among privately hired trainers, this information may be kept unpublished, or intentionally confidential, given that as many as 25 state PCW agencies have operated under consent decrees due to poor service delivery (Schwartz & Fishman, 1999). The

work environment presents a heavy caseload demand, not conducive to the practice of skills learned in training. Supervisors have not had time to observe, reinforce, and retrain their subordinates (Freeman & Morris, 1999). Trainers are challenged to address a wide variety of academic backgrounds and fledgling skills in the workshop timeframe. Since 2001, the California Social Work Education Center (CalSWEC) has been disseminating child welfare training and evaluation research on its website. The Children's Bureau website has also become a storehouse for some of this information.

Having escaped public scrutiny and pressure to conduct research on its practices for so long (Gelles, 2000; Lindsey, 2003), the PCW system has now been put on notice by the National Academy to show itself to be more accountable by instituting outcome-oriented, consumer-sensitive, and research-based methods (Chalk & King, 1998). The Adoptions and Safe Families Act (1997; ASFA) has made states responsible and competitive in meeting certain outcomes.

RESEARCH DESIGN AND METHODS

The study's focus on how PCW workers interview clients who come to their attention for allegations of child abuse and domestic violence was particularly salient because: (a) no studies existed on how PCW workers actually relate to adult clients, (b) no studies illustrate how PCW adult consumers perceive these interviews with their workers, and (c) practice wisdom dictates that clients, who are parents reported for the above-mentioned reasons, come to the interview situation defensive and anxious about the possibility of their child's removal.

This study was proposed against the backdrop of intermittent public interest, power disparities, deprofessionalization, and a new outcome/customer orientation stance. Since no published research was available on interviewing adult PCW clients, it was hard to support specific interview interventions with this population, except by an inference from studies on other populations. This study examined, for the first time, exactly what PCW workers actually do in the course of an interview with a parent. These specific parents were represented by standardized clients (SCs) who are actors/actresses trained to stay in character as a type of client. In this experiment, the client was referred to the PCW agency because her child had been reportedly injured in the context of a fight between the parents.

A large urban public child welfare agency was responsive to a request to conduct the proposed research with a small sample of its workforce (henceforth referred to as subjects). The author is grateful to the California Social Work Education Center for providing a curriculum development grant that funded this research. The curriculum developed would then help future PCW workers develop their skills in interviewing adult clients at the intersection of child abuse and domestic violence. The research was also inspired by a 6-year grant to UCLA funded by the United States Department of Health and Human Services (hereafter referred to as UCLA-HHS training) to assist PCW agencies in responding to cases that presented at this intersection.

Research Background Questions and Hypotheses

In preparation for this research, five key propositions from the National Research Council's (NRC, 2002) volume, *How People Learn*, were reviewed and then linked to

three theories that underscored this experiment. The first theory, Baldwin and Ford's (1988) model of transfer noted that subject characteristics, training design, and the work environment were three critical "inputs." Although this model captured the emerging state of training/transfer research almost 20 years ago, much of the current literature on transfer cites this as the foundation for present-day approaches. The review of the training and transfer literature highlighted the ascendancy of Bandura's (1977, 2001) Social Cognitive Learning Theory (SCLT). Key among his assertions was that self-efficacy was essential in facilitating skill retention.

Bandura (1997, 2001) defined self-efficacy as the belief in one's own capacity to organize and execute the courses of action required to manage prospective situations. He proposed that an individual's *expectations* about behavioral reinforcements influence behavior, more than *actual* previous reinforcement. This revolutionary concept emphasized beliefs and perceptions, and challenged reliance on Skinner's (1976) strict behaviorism. Integral to this departure was Bandura's emphasis on personal evaluation as a means of positive reinforcement. He hypothesized that self-respect, self-satisfaction, and belief in one's own competence are all goals and motivators. In essence, it is these anticipatory beliefs and perceptions that link an individual's behavior to good performance.

Kolb's learning cycle states that learning follows this four-part path: concrete experience, reflective observation, abstract hypothesis development, and active testing (1984; Kolb, Boyatzis, & Charalampos, 2000). According to Zull (2002), a biologist who analyzed learning theory and linked it to brain psychology, the key to all of this is the

front and back transmission of brain activity from Kolb's learning cycle that mimics the brain's cycle. When utilized in a balanced approach, learners convert data into their own ideas and actions, experiencing this conversion as learning (National Research Council, 2002; Zull). Zull maintained that in learning, transfer is about taking the "physical action step" of testing. Until we do that, all we have acquired is merely fanciful conjecture; in other words, action is a prerequisite to make the learning cycle complete. Zull also notes that this kind of actual testing helps the learner fill in the details of how to navigate between learning gaps. Transfer takes time for contemplation, action, and even random reaction. To further complicate transfer, according to Zull, learners can have an emotional reaction to the teacher/trainer that impacts motivation. The teacher/trainer must challenge learners to think in the classroom and in other novel situations while supporting self efficacy in order for the transfer to have a future. The teacher/trainer has to be careful not to engender a fear reaction (i.e., I will not be able to do this), which can cause learners to feel overwhelmed.

The training was delivered utilizing these three theories, Baldwin and Ford's transfer model, Bandura's SCLT techniques, and Zull's biological approach to learning. The Baldwin and Ford model of transfer and Zull's Biology of Learning helped identify factors before, during, and after the training that may have influenced the training's transfer. Because a new standardized measure was developed for this experiment, classical measurement and test theories also guided the establishment of validity and reliability for the instrument (DeVellis, 1991).

The specific questions asked during this research were:

- Does brief interview training for PCW workers, using the theories described, lead to skill transfer in a demonstration with a standardized client?
- How can subjects' interview skills, in a demonstration with a standardized client, be measured?
- If skills are transferred, is there a particular pattern in how that takes place that might be based on subject characteristics?

The hypothesis, related to each question, was:

- Interview training, using the theories described, will significantly improve interview skill performance in a sample of PCW subjects.
- An instrument designed to measure interview skill demonstration will demonstrate validity and reliability within a small developmental sample.
- The research will reveal patterns of skill demonstration between highest and lowest level skill demonstrators that will inform future training efforts in the areas of subject characteristics, and training design.

The last hypothesis was examined both quantitatively and qualitatively. A conceptual content analysis was used to determine what skill demonstrations distinguish an expert from a novice in this experiment. This kind of content analysis relies on theory developed by Krippendorff (1980), and is more thoroughly explained in the Colorado State University's (2003) web-based publication on content analysis history and methodology. The relationship between the questions pondered and the proposed hypothesis, as well as the theory and measurements used for this study, are outlined below in Table 1

TABLE 1: HYPOTHESES SUMMARY

Question or hypothesis	How measured	Statistical test	Theory at work
Interview training using social cognitive learning theory will significantly improve interview performance in a set of PCW workers	CWDVISS (2 raters) SC ratings on PPIF	Paired <i>t</i> -tests	Social Cognitive Learning Theory with Baldwin & Ford Model; Zull's Approach to Learning
Instrument designed to measure above will demonstrate reliability within a small sample	Interrater reliability	Correlations	Classic test and Psychometric theory
Instrument designed to measure above will demonstrate validity within a small sample	Correlation between CWDVISS, PPI F and FRS Expert rater concurrence	Correlations	Classic test and Psychometric theory
In order to uncover what might account for very different scores, the highest and lowest scoring interviews will be analyzed at critical junctures to determine what might account for score differentials.	Identification of key skill clusters: 1) Engagement 9) Listening 12) Safety Planning 13) Explaining Options Conduct content analysis of audiotape	Conceptual Content Analysis	Content Analysis

Methodology

This section describes a complex experiment within an experiment. Because no existing instrument could be found to measure what PCW workers actually did with adult clients, one had to be developed. This was deemed research with the developmental sample. That process will be described in detail because it also helps explain how the newly developed instrument's reliability and validity was established.

The new instrument, developed specifically for this research, was titled the Child Welfare Domestic Violence Interview Skills Scales (CWDVISS). There were two sources of concrete assistance in the instrument construction process. The first was an article by Finn and Rose (1982). Their contribution is recapped here in order to ground the discussion of the CWDVISS. Those researchers provided a detailed description of their process in distinguishing between novice and experienced mental health interviewers. They developed the Interview Skills Role Play Test (ISRPT) with subscales; two of these subscales Verbal Following and Seeking Concreteness significantly distinguished between the two groups. A third subscale, Nonjudgmental Responding, was thought to be relevant to the context of the current research, dealing with an allegedly battered woman who was also an involuntary client. Those three subscales became part of a validity check in the efforts to validate the CWDVISS. Beyond this content contribution, Finn and Rose shared the multiple struggles they had in attempting to resolve coding issues, establish reliability, and measure convergent validity. This served as a valuable foundation for proceeding with the investigation.

The other source of assistance in the CWDVISS construction was DeVellis' (1991) classic book on this topic, *Scale Development*. He identifies seven steps that must be undertaken. Here each step is identified and the activities engaged in by the researcher described:

- *Determine what to measure and develop an item pool.* The instrument developed for the UCLA-HHS training actually summarized all the points the training hoped to convey. It was entitled "The Assessment Instrument" and it covered all the issues that the interdisciplinary advisory team determined was key to conducting a comprehensive interview. This instrument is contained in the *From the Bottom Up* curriculum's appendix ("Assessment Instrument and Resource Development

and Safety Plan”). Two researchers extracted the key points from the instrument and agreed that 14 areas stood out as key “training points.” These later became “fields” or “skill clusters.” Because this is a behavioral measure, a group of three practitioners helped to operationalize the specific training points/skills/transfer points into skills that could be measured. Next, an extensive pool of items for each field was generated.

- *Develop a format.* Formatting challenges included these three: how to capture data during a live interview in multiple timeframes, how to code omissions and errors committed, and how to capture skills demonstrated more than once in a given timeframe. Finn and Rose’s (1982) study provided some guidance: They coded all of the above at 1-minute intervals for only the first 10 minutes. Having explained that this level of detail was exhausting for coders, this study settled on a similar level of detail, but a longer 5-minute interval, for nine segments. It was also decided to code for the full length of the interview (i.e., 45 minutes [in 5-minute intervals or segments]). Since this study devised a longer interval for coding, it also had to allow that some errors, as well as skills, might occur more than once. On the other hand, there were some errors that were omissions, and those only occurred once. For example, establishing partnership skills, two types of commission and omission errors were possible: “dominated plan with own focus” and “proscribes or directs,” could be committed multiple times while the error of “not stating the desire to work as a team,” could theoretically only occur as an omission once. The skills that could potentially be demonstrated here were identified as three: “demonstrates honesty,” “asks for trust,” and “asks what the interviewer can do.” It was theorized that all three of these skills could be demonstrated multiple times in each interval or segment. The number of fields, the skill items, and the error items combined to yield a sizable instrument. Because raters would need to have easy access to all the fields simultaneously, the instrument was printed on two 8½ x 11 sheets and taped together.
- *Administer to a developmental sample.* The instrument was piloted with a small developmental sample consisting of four child welfare workers who volunteered. Three volunteers had taken the UCLA-HHS training, and one had not. It was immediately revealed that the process was intense and exhausting, the interval too brief, and an uneven distribution of skills emerged. To address these problems, an adjustment was made with input from the raters: the interval would be elongated to 10 minutes to resolve all three problems. It was projected that a longer interval would make raters feel less pressured. A wider range of situations on both skills and errors could more effectively be observed and coded. Finally, this would lead to a greater likelihood of saturating the instrument, that is to say, with a longer timeframe there would be more opportunity to record more skill and error items. The CWDVISS had allocated boxes for ease in check-off. When skills were demonstrated beyond the three allotted “boxes,” raters agreed to just

keep checking on that “line.” This led to the adoption of a modified set of rules for future coding. It was also discovered that raters needed a 30-minute break between subjects because of the intensity of the task, and the need to be “fresh” for the next observation. A preliminary assessment of validity and reliability will be discussed in a subsequent later in this section.

- *Have experts review the item pool.* Three experts were asked to review the item pool. The experts were the author and two expert faculty members. All three had at least 2 years’ experience in the family violence field. Two had previous experience as child welfare workers. They were apprised of the study’s goals and were asked to freely edit the instrument. Some redundancies were pointed out but purposely retained to provide a check on internal consistency. In accordance with the experts’ recommendations, several behavioral measures were eliminated and some were added; the overall number of fields on skill clusters remained at 14. This expert review helps support an argument for content validity.
- *Include validation items.* DeVellis (1991) explains that validation items are items that are a check on the scales’ final validity. That may include a social desirability scale or items that measure other constructs. No measure of social desirability was included because the instrument seemed to be too lengthy. Several items measuring verbal following, seeking concreteness, and feeling reflection from the ISRPT (Finn & Rose, 1982) were included as a potential validity check. These items served two functions. First, they appeared to capture some aspect of good interview skills. Second, they were items from a measure with established validity. Although it is clear that items extracted from a scale no longer retain the validity attributed to the whole instrument, they were good checks on face validity. For example, these items adopted from the ISRPT included: responds to client comment with nonjudgmental phrases, (e.g., “I see,” “Oh,” etc.) and pauses 2-5 seconds between own and client’s statements.
- *Evaluate items and optimize skill length.* Because some aspects of evaluating items have already been discussed, the focus here will be on length of the instrument. An argument could be advanced that human coders cannot abstract this much data in a “live” interview. Alternatively, this research proposed to determine if there was a pattern of skill acquisition among the 14 skills trained in the intervention. If the fields were edited down, there would be a possibility of being limited in assessing which skills were or were not being demonstrated. It would be better to collect more data now, and collapse it if necessary later. Arguably, a larger developmental sample may have afforded more confidence that certain skill clusters were or were not useful for either skill or error pick-up.
- *Test reliability and validity with a sample.* This was done twice, preliminarily with a group of four volunteers (described here) and later with a group of six, referred

to as a developmental sample. For this first rating of four volunteers, three expert raters were used. These were the same raters who reviewed the instrument's lengthened content. Two raters used the instrument to rate and a third used her expertise to rate the interviewer qualitatively and independently. A positive correlation was anticipated. This plan was a preliminary assessment of validity and interrater reliability issues, and followed the process outlined by Finn and Rose (1982). There was 100% agreement between the expert's ranking of each interviewer's skill demonstration with each interviewer's mean total score on the CWDVISS. In other words, the expert's ranking of first, second, third, and fourth matched the highest to lowest mean scores for the four volunteers using the instrument. Initial interrater reliability was assessed differently. The two raters had practiced rating interview videotapes that were constructed using each other and a volunteer playing a client scenario. In this first "test" administration, an initial overall interrater reliability of .69 was achieved. While Kerlinger's (2000) discussion of reliability discusses .70 as a cutoff score that some experts deem the limit, he allows for a lower score being acceptable when validity is high. When scoring is in this range, he advises reviewing to assure items are unambiguous, adding equal items, and standardizing instruction, with retesting. We focused in on the third option, hoping that this score could be improved via more practice ratings with clear instructions, conducted by raters for the actual study. This was our next step, referred to here as our research with the "developmental sample." The goal was to establish robust correlations between and among the Finn and Rose subscales (FRS), the CWDVISS scores, and the SC's ratings on a standardized measure as a measure of concurrent and construct validity. The standardized measure used here was the Patient-Physician Interaction Form (PPIF). Its use is widespread in California Medical Schools as a tool for SCs to give specific feedback to medical students on their interactive skills (L. O'Grady, Personal Communication, January 8, 2001). This PPIF is discussed in greater detail later in this report.

- Six volunteers participated in the second testing of the instrument, referred to as the "developmental sample." Interrater reliability was calculated per segment (maximum three segments per subject), and is further described in Table 4 below. Validity was checked in two ways. First, the SC rated the subject's ability to meet her need in the interaction on the PPIF. This was correlated with the total CWDVISS instrument score. Second, three subscales from the ISRPT were extracted and compiled into what was renamed the Finn-Rose Subscales (FRS) and correlated with the instrument's first segment score. Ultimately, the final instrument contained 14 fields, rated in three 10-minute observation segments for a total top score of 137 points.

TABLE 4: INTERRATER RELIABILITY CORRELATIONS, PER SEGMENT OF CWDVISS, FOR THE DEVELOPMENTAL SAMPLE

Segment	<i>R</i>	alpha	Standardized	
			alpha	<i>p</i>
1	0.9742	0.9865	0.9869	0.0537
2	0.8503	0.9830	0.9191	0.6714
3	0.9222	0.9545	0.9595	0.0539
4	0.9242	0.9490	0.9606	0.0285
5	0.8969	0.9455	0.9457	0.0366
6	0.8845	0.9374	0.9387	1.0000
7	0.8890	0.9164	0.9412	0.0099
8	0.8413	0.9137	0.9138	0.6859
9	0.8273	0.9041	0.9055	0.0071
10	0.8346	0.8850	0.9099	0.0526
11	0.7264	0.8414	0.8415	0.0961
12	0.6713	0.8096	0.8105	0.8185
13	0.9600	0.7663	0.8166	0.1383
14	0.7524	0.6080	0.8587	0.8557
Total Mean Scores	0.8539	0.8857		

This research had another component, referred to as the core experiment, which utilized a quasi-experimental model, with a pretest, intervention, and posttest ($O_1 \times O_2$) design. The independent variable was the training and the dependent variable was the demonstrated interview skill level. The research utilized two SCs who had been previously employed in the UCLA Medical School's Identified Patient Program. Thus they readily adapted to several hours of training in order to reliably represent the same client who was "reported" to the local public child welfare agency.

The SCs were matched on gender, ethnicity, age, and the ability to stay in character. They were Caucasian females in their early thirties. Two vignettes were utilized, rated at the "moderate level" of skill challenge by three researchers. Both the vignettes and the SCs were switched when the posttest was administered to assure

subjects would not become overly familiar with the particular “client” in question. Two teams operated simultaneously, resulting in data collection on all subjects in a short timeframe. The same process was repeated at time two, less than 2 weeks after the training intervention. Subjects were rated by two raters in the interaction with the SC. The raters were the author, and three Title IV-E stipended MSW students, who were trained together over a 12-hour period, in order to ensure acceptable reliability. All raters coded two practice role plays as a first pilot test. A first pilot test of reliability had been conducted with the developmental sample described earlier.

Overall, five instruments were used to collect data. They are summarized in Table 2 below and briefly described here. First, demographic data on each subject was gathered, on the Demographic Data form (DD). Next, the Phase II, Part I Questionnaire (PPQ) tapped the subjects’ ability to list what specific things they planned to do. It was anticipated that the items on this *to do* list would: (a) reflect what they had learned in the training; and (b) illustrate Bandura’s social cognitive learning theory, where the subject developed confidence and a sense of self efficacy as he/she began to anticipate that he/she could accomplish the *to do* list. The actual interaction of the subject and the SC was coded using the CWDVISS. Its reliability and validity was piloted with a separate and small “developmental sample” of six actual workers. Ultimately, this instrument included 14 fields, or “skill clusters,” that specifically pertain to skills needed for interviewing adult clients whose cases involved both domestic violence and child abuse. These skill clusters scores were scored quantitatively, with points assigned for repeated, specific skill demonstration, and points subtracted for repeated, specific errors. The

CWDVISS served as an anchor for a later qualitative content analysis, which included a post hoc analysis of the audiotaped interview. The SC used the Patient-Physician Interaction Form (PPIF) to rate the subjects. Although there were no published reports of this instrument's validity or reliability, its use is widespread in Identified Patient Programs in California Medical Schools as a tool for SCs to give specific feedback to medical students (L. O'Grady, personal communication, January 8, 2001). It measures the patient's level of satisfaction with the interaction in the following areas: listening, gathering information, establishing rapport, exploring perspective, addressing feelings, and meeting patient needs. Although a 2006 search of the Journal of the American Medical Association yielded no published studies using this specific instrument in research, it was the only instrument found that was subtle and "in use" for capturing the SC's reaction. This instrument has seven fields rated on a five-point Likert scale. It measures the patient's global level of satisfaction during their interaction with the medical student.

TABLE 2: EXPERIMENTAL INSTRUMENTS

Name	Acronym	Purpose
Demographic Data	DD	Asked subject's age, race, education, years of experience, level of previous training. Used in both developmental sample and core study.
Phase II, Part I Questionnaire	PPQ	Asked three preliminary questions to help subjects form a plan used in the core study only. Maximum points: 18.
Child Welfare Domestic Violence Interview Skills Scale	CWDVISS	Measured skill demonstration in both developmental sample, test for reliability and validity, and in the core study; 14 fields: Engagement; Assessing for DV; Demonstrating Priority of Safety; Addressing Potential for Child Removal; Establishing a Partnership; Providing Feedback Nonjudgmentally; Inquiring about Strengths; Inquiring about Injury; Listening; Conducting Threat Assessment; Conducting Social Support Inventory; Engaging in Safety Planning; Explaining Options; Providing Resources. Maximum points: 137.
Patient-Physician Interaction Form	PPIF	Measured the SC's reaction to and appraisal of the subject's skill demonstration; 7 fields: Listening; Gathering Information; Establishing Rapport; Exploring Perspective; Addressing Feelings; Appearing Competent, Meeting Patient Needs. Used in the developmental sample test and core study. Maximum points: 35.
Finn and Rose Subscales of Interview Skills Role-Play Test (ISRPT)	FRS	Measured three subscales (verbal following, seeking concreteness, and nonjudgmental responding) of the Interview Skills Role Play test. Used as a validity check in developmental sample test only. Maximum points: 90.

It was expected that a correlational analysis would reveal a positive relationship between the PPIF score and the CWDVISS. It was important to approach the

experiment with concurrent measures to provide a mechanism for feedback to the PCW worker and a way of involving the SC as a client proxy. While the “borrowing” of the ISRPT’s subscales does not retain the original instrument’s established reliability and validity, what was being sought was a broad check on validity through establishing construct validity, and it therefore served that purpose.

Sample Description and Procedures

Turning our attention to recruitment, there was a challenge in securing enough subjects for these two samples: the first is referred to as the “developmental sample” for testing the CWDVISS’s reliability and validity; the second is the sample for the core study. Subjects were recruited by email and given the following incentives: employment training credit for all levels of participation, a domestic violence book, and a feedback letter (with audiotape if requested). Subjects were assured that their names, individual scores, and audiotaped interviews would be confidential. The original intent of the research was to compare a set of subjects who had previously taken the UCLA-HHS training with a matched set of subjects who had not taken the training, but that plan had to be modified given the low number of subjects recruited, because only three subjects had taken the training. Ultimately, the study involved two small convenience samples, consisting of six child welfare workers in the “developmental sample”, and 15 child welfare workers in the core study. The original core study was set at 19, but two subjects dropped out at the posttest, citing casework emergencies, and two sets of records became unmatched (i.e., it was not clear which was the pre- or posttest for a particular subject), rendering them unusable. In sum, the core study sample was a total

of 15 subjects who were 60% male and 40% female. With regard to ethnicity, 46% of the sample was African American, 34% Caucasian, and 20% Latino. The higher representation of African American males in this sample was typical of the workforce in that particular region of the agency at the time of the study.

The curriculum for this training intervention was adapted from a training previously published by the author as: *Assessment and Case Management of Domestic Violence in Public Child Welfare* (Friend, Mills, & Hoang, 1997). The curriculum was also summarized in a subsequent publication in the Social Workers' Desk Reference (Friend & Mills, 2002). It was also published by the United States Department of Health and Human Services as a training program in 1999. These publications contain either the full or a shorthand version of the original assessment instrument for this earlier training intervention; its content was the foundation of the 14 skill clusters contained on the CWDVISS. The assessment instrument is contained in the appendix in the curriculum. Table 3 outlines process and content points that were extracted from the original training's content to become this experiment's intervention, and are summarized below.

TABLE 3: KEY TRAINING INTERVENTIONS SUMMARIZED

- Acknowledged the tension between domestic violence and public child welfare service providers.
- Addressed potential feeling reactions (fear, overwhelmed, helpless) and normalized them.
- Recognized that higher rates of family violence exposure exist among helping professionals. Addressed how this could be a help or a hindrance in job performance. Discussed what to do if it becomes a hindrance.
- Empathized with workload/organizational demands and their impact on subjects' decision making, addressed paradox of demand to do more work.
- Identified trainer's work history/experience.
- Elicited subjects' experiences and impressions.
- Acknowledged previous academic training was probably not addressing this.
- Utilized visuals (family violence tree, Heart of Intimate Abuse video demonstration) to explain concepts.
- Provided skill demonstration before role-play performance.
- Collaborated on using cognitive techniques such as a *to do* list, and developed group's own mnemonic devices.
- Rotated role-play roles, to facilitate experiencing more than one perspective of the dilemma.
- Provided subjects feedback on strengths demonstrated in role play first and then coaching on other options.
- Solicited subjects' anticipations of benefits to using this method, and appraisal of utility; asked for negative feedback.
- Built on previous knowledge, experiences; elicited what these were.
- Provided conceptual theories/strategies for the development of a framework in the subject: feminist theory, person-in-environment, Motivational Interviewing Principles, and Stages of Change theory.
- Utilized the structure of an instrument to summarize training and guide initial role plays.
- To some extent, the pretest interview with the SC participated in the intervention because subjects "experienced" the SC's reaction to their baseline interviewing.

The difficulty of dealing with domestic violence was acknowledged on four levels.

First, there are tensions between Domestic Violence (DV) service providers and PCW

workers that divide these respective groups on the level of client alignment and mother blaming (Edleson, 1996; Friend, 2000). Second, workers were allowed to normalize their potential feeling reactions to domestic violence which encompassed feelings ranging from fear to helplessness to ambivalence. Third, it was acknowledged that professional helpers have higher rates of childhood exposure to family violence compared with some other professions (Jackson & Nuttall, 1997). Fourth, organizationally, workload and management demands interact to create a short timeframe in decision making. Because the trainer was previously a PCW worker, she was able to align herself with the trainees' experience. The trainees' perceptions and impressions were frequently solicited and built upon whenever possible. Skills were shown being demonstrated on videotape before role-play participation was requested. When the role-play occurred, workers rotated roles to experience more than one perspective. Trainees from this agency were reluctant to engage in role play interviews. Although this could have been attributed to a variety of reasons, it was speculated that the trainees had difficulty implementing a complex skill set, especially when the stakes were high, meaning they were performing in front of a small set of peers. Because trainers understood that trainees would be self-conscious demonstrating and being observed, trainees were given feedback on strengths, followed by coaching on options. As a group, workers brainstormed developing their own mnemonic devices or acronyms for remembering cognitively what to do. The importance of developing a mental or paper *to do* list was emphasized as a means of enhancing self-efficacy. Finally, the trainer was equally solicitous of problems with the utility of the method. In this way,

multiple members of the group with a range of experience and skill levels could be used as problem solvers and reinforcers.

Data Collection

What follows is the data collection procedure for both the core and developmental sample of the study. However, for the developmental sample, there were some differences worth noting. Subjects in the developmental sample were coded by two raters in just one observation, without the benefit of the training intervention, and the audiotape of the interaction was coded that same evening by the same raters utilizing the FRS as an additional validity check on the newly developed CWDVISS.

In the core study, once subjects signed consents, they were given the child abuse hotline/referral on the client. They then viewed a videotape of the SC reacting to the agency hotline report. They were asked to fill out the PPQ, which asked three preliminary questions to help subjects formulate a cognitive *to do* list. They were then told they would have up to ½ hour to interview the identified patient, be observed by two coders, and audiotaped. Subjects were told the audiotape would be made to allow for supplemental coding after the interview concluded. Two teams of two raters each operated simultaneously and separately, rating the subject's interviewing skills on the CWDVISS. Groups were identified as "A" and "B" so that the SC could be switched at time two (the posttest). The literature review advised that previous experimenters detected a "familiarity effect" when subjects encountered the same SC at time two (Badger & MacNeil, 2002; Carillo, et al., 1993). As a protection against such threats to

internal validity, the subjects encountered a new SC before the posttest was administered.

In order to determine if there were unique or shared patterns of skill acquisition, the highest scoring (both pretest and posttest) subject on the CWDVISS and the most improved (from pretest to posttest) subject were examined quantitatively and qualitatively. The total item score was obtained by adding the scores assigned per item. The assumption was that the highest scoring subject (i.e., expert) represented the best preexisting skill and, the latter subject (i.e., novice) the most improved skill. A demographics check showed subject two had less than 1 year of work experience and was a novice in skill and experience level. In sum, the pattern of their CWDVISS scores is found in Table 9, and their scores on the PPQ and a content analysis of the audiotape were tabulated and analyzed in Table 10).

TABLE 9: EXTREME SUBJECTS' PER SEGMENT RANK ORDERED SKILL SCORES FROM CWDVISS

Segment 1								Segment 2							
Subject 2				Subject 5				Subject 2				Subject 5			
Pretest		Posttest		Pretest		Posttest		Pretest		Posttest		Pretest		Posttest	
Skill	Score	Skill	Score	Skill	Score	Skill	Score	Skill	Score	Skill	Score	Skill	Score	Skill	Score
2	9	1	12	1	9	1	16	9	7.5	5	10	9	7	9	9
9	8	2	10	9	9	9	6	1	5	9	9	10	6	1	6
1	3.5	5	6	4	2	2	3	14	4.5	13	7	1	3	10	6
7	3	9	6	6	2	5	2	2	1	1	5	5	3	6	3
11	2.5	14	4	7	2	6	2	11	1	2	5	6	3	2	2
3	0	4	3	11	2	13	2	13	1	4	5	2	2	7	1
14	0	13	3	5	1	3	1	3	0	14	5	13	2	3	0
5	-1	6	2	2	0	7	1	4	-1	6	4	4	1	4	0
12	-1	11	2	3	0	8	1	7	-1	11	3	7	1	5	0
13	-1	7	1	8	0	11	1	12	-1	7	2	11	1	8	0
4	-2	8	1	10	0	4	0	8	-2	10	2	3	0	11	0
6	-2	3	0	12	0	10	0	10	-3	12	1	8	0	12	0
8	-2	10	0	13	0	12	0	5	-4	8	0	12	0	13	0
10	-3	12	0	14	0	14	0	6	-8.5	3	-1	14	0	14	0
Total	14		50	Total	27		35	Total	-0.5		57	Total	29		27

TABLE 9: EXTREME SUBJECTS' PER SEGMENT RANK ORDERED SKILL SCORES FROM CWDVISS (CONT'D)

Segment 3							
Subject 2				Subject 9			
Pretest		Posttest		Pretest		Posttest	
Skill	Score	Skill	Score	Skill	Score	Skill	Score
9	10	--	--	13	10	9	9
1	6	--	--	10	9	1	5
3	3	--	--	1	6	6	5
11	0	--	--	9	6	10	5
2	0	--	--	5	4	2	4
4	-0.5	--	--	6	4	5	4
12	-1	--	--	2	2	13	4
7	-1	--	--	7	2	14	3
14	-1.5	--	--	4	1	3	2
8	-2	--	--	8	1	4	2
13	-3	--	--	12	1	7	1
10	-3	--	--	14	1	8	0
6	-3	--	--	3	0	11	0
5	-9.5	--	--	11	0	12	0
Total	-5.5			Total	47		44

TABLE 10: PPQ SCORE "10 THINGS LIST" AND AUDIOTAPE RATINGS FOR EXTREME SCORERS

	Subject 2				Subject 9			
	PPQ		Audio		PPQ		Audio	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
A	2	2	1	2	2	2	2	2
B	2	2	2	2	2	2	2	2
C	2	2	0	2	2	2	2	2
D	2	2	2	2	2	2	2	2
E	1	2	0	2	2	2	2	2
F	1	2	1	2	2	2	2	2
G	0				2	2	2	2
H					2	2	2	2
I					2	2	2	2
Totals	10	12	6	12	18	18	18	18

Next, the qualitative portion of the research will be described. The researcher attempted to link the "list what 10 things you would like to do" section of the PPQ with what the subject actually did in the interview. The PPQ measure served two theoretical

purposes. In SCLT, cognitively preparing for your interview by naming what you will do is a strategy that boosts perceived self-efficacy which, in turn, serves as a motivator (Bandura, 1977). In Zull's (2002) conceptualization, this thinking of future actions is important, but it derives its highest value when it is accompanied by the actual testing. It is in this actual testing that learners convert a mental process to a physical action where they both see and invent the details of what they need to learn. This is what Zull refers to when he discusses how the brain is changed.

The PPQ *to do* list was rated in accordance with Cournoyer's (2004) practice rubric, as adapted by the author for this experiment (see Table 8 below). A score of two on an item indicated best practice, a score of one indicated acceptable practice, and a score of zero indicated unacceptable practice. For example, if the subject said he was going to "help the client build her self-esteem," that was given a "2," but when that was followed with "help the client build her confidence," the second entry was given a "1" as it was deemed to be repetitive of the earlier entry on self-esteem. One subject said he would, "interview the family together to see dynamics" at the pretest. That was scored "0," as in the raters' judgment, the safety risks that this approach engendered caused it to be contraindicated in good practice. It is specifically not recommended in the professional literature, relegating it to questionable status in the Cournoyer (2004) Practice Quality Rating Rubric. Although the subject probably did not know this in pretest condition, had it appeared in the posttest (it did not), it would have received a "0" score as well because the training clearly focused on the first time intervention, and

stressed the importance of initially interviewing clients alone. Two raters independently rated the PPQ for these two subjects and achieved 100% agreement.

TABLE 8: PRACTICE QUALITY RATING RUBRIC FOR PPQ RATING

Grade	Points	Description
Best or Good Practice	2	<ul style="list-style-type: none"> - Some research-based evidence - May have few RCT studies - Some clinical trial studies - Several case studies and consumer reports - Strong expert or professional association endorsement - Fits with practice guidelines - Most evidence supports safety and effectiveness
Acceptable Practice	1	<ul style="list-style-type: none"> - Some research-based evidence - Few clinical trial studies - Some case studies and consumer reports - Some expert or professional association acceptance - Evidence reflects few risks and some indication of effectiveness
Questionable to Dangerous Practice	0	<ul style="list-style-type: none"> - Lack of research-based evidence - Few or no case studies or consumer reports - Experts or professional associations are neutral or against its use - Evidence reflects some to considerable risks and minimal indication of effectiveness

Adapted by Colleen Friend from Cournoyer, B. (2004). *The evidence-based social work skills book*. Boston, MA: Allyn & Bacon.

Next, a content analysis was conducted on the audiotape of those two subjects. This was both a quantitative and qualitative process as outlined in the Colorado State University's (2003) web-based guidelines and by Krippendorf (1980). Conducting a conceptual content analysis involves five steps outlined below, annotated by what was actually done in the analysis.

- *Decide on the level of analysis:* Phrases and sentences were the “level” of this analysis.

- *Decide how many concepts to code for:* Essentially, we were looking to determine if the subject did what he said he planned to do. Since each subject had set his own “list,” this became a highly individual endeavor.
- *Decide whether to code existence or frequency:* This is important because it can change the coding process and outcome. It was decided to code for existence, which was consistent with our initial guess: could the subject then do what he said he would do? While frequency might have presented a different picture, existence met our need. Having established that, the immediate question was how to determine qualitatively that something existed. This led to the making of rules.
- *Develop rules for coding your texts:* The audiotape was then examined on two levels. First, there was an attempt to evaluate if the subject did what he said he was going to do (existence) and then how well it was done. The scoring methodology was this: 2 points assigned for addressing the issue identified and doing it well; 1 point assigned if the issue was approached but something ranging from inept delivery or unresponsive to a client challenge got in the way; 0 points assigned if the issue was either not attempted or the attempt was counterproductive. For example, when a subject said he would, “talk about supportive services that could help,” and he followed up in the interview by describing these services and asking the client if she would like to do this, he received a score of 2. However, when he seemed to be following up on referring the client to help groups, he tried to push her to volunteer at the shelter. This was categorized as a “1” score, because it was deemed an inept attempt to get her to go to a shelter. When subjects undermined their own plan by making very judgmental statements or not addressing the issue, that received a 0 score. Two raters, this researcher and another MSW student, coded the audiotapes. Each researcher coded the tapes independently. The actual coding was preceded by the coding of three randomly chosen audiotapes from this experiment as practice. After each tape’s coding, agreements and disagreements were discussed. The third tape’s coding was used as a test for reliability where alpha equaled one minus observed disagreement/possible agreement. For example, if a subject said he would do seven things then that became a contribution to the denominator of the equation for total possible agreements. That seven was multiplied by three to reflect the (2, 1, 0) options for how well the item was addressed. Thus the denominator became (7 x 3 = 21). The numerator was the number of times the independent raters agreed. For this content analysis the reliability was .87, which is consistent with Krippendorff’s (1980) discussion of acceptable reliability.
- *Code the texts and analyze the results:* The audiotapes were coded manually and the analysis is presented in the summary findings.

In sum, the process just described involved the use of a SC to measure PCW workers' interviewing skills for two purposes: First, to pilot the reliability and validity of an instrument (CWDVISS) expressly designed for this purpose and experiment with a developmental sample, and second, to evaluate the pre- and posttraining intervention change in interviewing skills with a small sample of actual PCW workers. The core experiment controlled for familiarity by switching the SC and vignette at the posttest. The subjects with extreme scores on the CWDVISS were compared on their scores per segment. A conceptual content analysis was undertaken to further analyze the skill acquisition and change between the extreme scores. The analysis included a qualitative rating of the extreme subjects' *to do* list and a qualitative rating of their skill performance in the attempted execution of the list. What follows is a summary of the findings.

RESULTS

As already noted, this research used a developmental sample ($n = 6$) for instrument reliability and validity, and another sample ($n = 15$) as the core study. This resulted in a complex analysis and therefore will be recapped here in an effort to achieve clarity.

Developmental Sample

First, a pilot test of reliability and validity of the CWDVISS with a small developmental sample discussed earlier, found that the instrument's reliability is comparable to those achieved in other studies where skills are being rated in a live demonstration (Ford, Fallowfield, & Lewis, 1996; Ong, Visser, Van Zuuren, Rietbroek, Lammes, & DeHaes, 1999). Interrater reliability (with two raters) was calculated by

correlating each rater's total score on each 10-minute segment. Of the six subjects, three conducted the interview over three segments, two spoke for two segments, and one only interviewed for one segment. This yielded a total of 14 segments. Table 4 (page 19) illustrates these calculations. Note that most of the correlations are in the .9 to .8 range, only one correlation dipped below .7. This range of correlations is deemed in the "good" range according to Krippendorff (1980). The mean score for the correlations is .85 with a corresponding mean alpha of .88. The alpha is the proportion of the scales' total variance that is attributable to a common source, presumably the true score of the latent variable (DeVellis, 1991). This correlation is an indication that the two raters were fairly consistent in rating the subjects with this instrument and that the instrument was fairly reliable as a tool to measure domestic violence and child welfare interviewing skills in each segment of the interview. Although reliability is necessary, it is not sufficient. Concurrent and construct validity were considered next.

Construct validity of the CWDVISS was established by determining the extent to which correlations among all three instruments (CWDVISS, PPIF, and FRS) led the researcher to believe that the CWDVISS scores behaved the way they were expected in relationship to established measures of other constructs. The difficulty here was that we were measuring an instrument designed to focus on a narrow set of PCW skills, against instruments designed for generic social services interviewing (i.e., FRS) and generic medical practice (i.e., PPIF). Thus, all three instruments attempted to measure the interview skills demonstrated by the subjects, at differing levels of detail. To recap the scoring methodology, scores for the FRS were obtained from the audiotape replay by

one rater, while the CWDVISS score was a mean score derived from two raters who watched the interview. The scores on Segment 1 for each subject are shown in Table 5 below. One of the SCs used in the core experiment was used with the developmental sample; she rated these subjects on the PPIF. An ex post facto power calculation concluded the power level for this sample of six was .09. This suggested that given the effect size of this developmental sample study, a substantially larger sample would need to have been recruited in order to produce more statistically robust findings; hence, this study was a valuable pilot test. Table 6 below shows the bivariate correlation between the mean scores for the FRS, the PPIF, and the CWDVISS segment one and CWDVISS total scores.

TABLE 5: DEVELOPMENTAL SAMPLE SCORES: FRS, CWDVISS, AND PPIF SCORES

Subject	FRS Subscale Total	CWDVISS Mean Segment 1	CWDVISS Mean Total	PPIF
1	10.00	28.00	38	4.0
2	34.00	34.50	47.5	29.0
3	84.00	34.50	108	35.0
4	25.00	26.00	81	19.0
5	45.00	31.00	53	22.0
6	40.00	31.50	31.5	13.0

TABLE 6: INTERCORRELATIONS BETWEEN INSTRUMENT SCORES FOR THE DEVELOPMENTAL SAMPLE

	Segment 1	PPIF
FRS	$r = .72^*$ ($p = .055$)	$r = .82^{**}$ ($p = .027$)
Segment 1 CWDVISS	--	$r = .725^*$ ($p = .052$)
Total CWDVISS		$r = .721^*$ $p = .053$

*Statistically significant at $p < .06$

**Statistically significant at $p < .05$

All four scores correlated in the moderate range (.72, .72, .72, .82), the first three with statistical significance at $p < .06$ and the correlation of the PPIF with the FRS was .82, at $p = .027$. Because the test for statistical significance was set at $p < .05$, it cannot be stated that these correlations were statistically significant. Due to the small sample size and corresponding lack of power to detect differences and similarities between raters, statistical comparisons were explored when p -values dropped below 0.10. Being an exploratory study, we were interested in anything with a p -value $< .10$. The overlap suggests that the instruments were measuring similar concepts. Arguments supporting the construct validity of the CWDVISS can be stated this way: bivariate correlations in the moderate range, among the FRS, the PPIF with the first segment of the CWDVISS, and the PPIF with the total CWDVISS, support that the constructs they each measured have moderate relationships with specific skills in domestic violence and child welfare interviewing.

Core Study

In the core study, t -tests conducted on participants' CWDVISS pretest and posttest scores (average of the two raters) demonstrated that the approximate 28-point difference was statistically significant at $p = .01$. This suggests that the independent variable (training) had a positive impact on the subject's interviewing skills. Effect size calculations, using Cohen's d were estimated at 1.05, indicating a fairly high intervention effect (Counoyer, 2004). Effect size is a statistical indication of the difference between the two groups, which places the difference in a context that helps measure the effectiveness of the intervention.

Conversely, using the PPIF metric, pretest to posttest mean scores improved by 1.4 points, but this improvement was not statistically significant. This suggests that the intervention was not effective as measured by the PPIF. However, there are structural differences between these two measures that might have influenced the findings and should be considered here. The CWDVISS was designed to measure the whole interview and range of interview skills specific to Domestic Violence and Child Welfare needs at the pre- and postintervention level. The PPIF was drawn from a medical practitioner training program, and it may be capturing global levels of patient satisfaction in a generic medical situation, whereas the instrument designed for this public child welfare worker training (CWDVISS) detected pre- and postinterview changes at a more specific level. Despite the limitations of the PPIF in measuring specific change in this domain (domestic violence and child welfare), it was important to approach the experiment with multiple concurrent measures. This leads to the speculation that larger sample testing is necessary before we can draw definitive conclusions about the exact relationship between these instruments and their ability to detect changes in interviewing skills. These results are summarized in Table 7 below.

TABLE 7: PRETEST AND POSTTEST CORE STUDY SCORES WITH EFFECT SIZE

CWDVISS*				PPIF**	
OBS	ID	PRE 1	POST 1	PRE 2	POST 2
1	1	13	97	19	26
2	2	8	102	8	19
3	3	47	90	29	29
4	6	36	99	7	31
5	9	103	106	35	35
6	10	49	29	0	3
7	15	53	106	17	5
8	19	48	37	20	14
9	20	31	36	19	10
10	14	48	61	11	7
11	18	39	90	24	8
12	12	75	71	27	22
13	13	27	67	18	23
14	11	81	64	19	22
15	17	24	45	9	28
Total Mean Score		45.46	73.33	17.40	18.80
Difference		27.87		1.40	
SD		25.53	27.38	9.78	10.30

$$\text{CWDVISS Cohen's } d = \frac{73.33 - 45.46}{\text{pooled } SD (26.47)} = 1.05$$

*Paired *t*-test pretest/posttest *p* = .010

**Paired *t*-test pretest/posttest *p* = .654

Finally, two subjects whose scores were extreme (i.e., lowest at baseline to most improved [Subject 2] and highest baseline [Subject 5]) were analyzed qualitatively and quantitatively, revealing different patterns of skill acquisition in the posttest observation. These subjects were both males, thus they are alternately described as ‘he.’ In sum, this comparison is illustrated in Table 9 (page 28), and can be summarized in this way: Engagement (Field 1) and listening (Field 9) are high scores for both subjects in both

observations, and conducting safety planning is a low score for both. Subject 2 scored well on giving options at the posttest, altogether suggesting that he might be using the mnemonic (cognitive) strategy made up by the subject group (i.e., LEGO—Listen, Explain, Give Options). Both subjects' low score on Field 12 (safety planning) may reflect that this skill cannot be acquired with the intervention the study offered, or it may be consistent with the difficulty many helping professionals have demonstrating this skill (F. Danis personal communication, June 23, 2002).

Subject 2's dramatic improvement in the posttest reveals a shift of increased skill demonstration in establishing a partnership and explaining/giving options. It might mean that Subject 2 began to share power and draw the client into participating, as this is what these two skills have in common. On the other hand, Subject 5 seems to be doing something very different. First, focused almost exclusively on engagement and listening in the first segment, he waits until the second 10-minute segment of the interview to shift into an escalated discussion of the presence of domestic violence and conducts a threat assessment close to the end. He was consistent in the demonstrating of nonjudgmental feedback, something Subject 2 only partially demonstrated at the posttest. Despite being the highest scorer throughout, Subject 5's scores slightly dipped at the posttest's second and third segments, suggesting he was selectively experimenting with some new skills. Lastly, Subject 2's continued brevity (choosing to complete the interview in 20 minutes (i.e., two segments) may have cost him even more points, leaving open to speculation what he might have achieved had he kept at his new skills for 10 more minutes (a third segment) as the other subjects did.

Turning the focus to whether or not the extreme scorers were different on their execution of what they had planned to do, the answer is important. The analysis of the audiotape shows that Subject 2 had difficulty making a plan at the pretest. In execution, he was scored as having errors in the pretest for being very judgmental. That was noticeably improved (but not eliminated) at the posttest, suggesting the intervention helped him. The intervention seemed to also help him come up with a better *to do* list; at the posttest all his items were scored as being within good practice and/or espoused in the training and he was able to execute them at the right level. However, there again, his brevity compromised his total score. Subject 5 had none of these issues at either observation; his audiotape coding revealed perfect execution of a full score plan. An examination of his demographics showed that he had been a child welfare worker for 5 years, while Subject 2 had been hired within the last year. This difference in experience realistically could have accounted for what we were observing; however, the SC's ratings showed that Subject 5 was rated several points higher on the PPIF than his most closely matched-for-experience peers. The analysis supported that the expert appeared to be only slightly modifying his approach, which was likely to enable him to focus on only a few areas of the training. However, we can speculate that the training might have presented the novice with a large amount of new information. Perhaps in his efforts to take it all in and perform it all at one time, while much improved, were hampered by his attempts to do too much, too soon. It is also possible that given Subject 2's inexperience, he greatly benefited from being trained with more experienced colleagues. He may be relying upon the mnemonic device created by the group as he

engages and listens. He also may have taken in the group's insights and problem solving in ways the experiment did not measure, but is possibly contributing to his dramatically improved posttest score.

Discussion of Limitations and Findings

There are several key limitations to this study and identifying them first will help establish the context for the findings discussion. First, this quasi-experimental design precludes causal statements. Second, these small voluntary samples are not necessarily representative of the whole population of PCW workers; the study was meant to be a pilot study. Third, an analysis of the extreme scorers helps us delve into the details of skill acquisition, but it may not represent the whole sample's process (Miller & Crabtree, 1992). Fourth, this experiment lacked the resources to control for or measure all of the multiple constructs that participate in three theories that formed the foundation for the research. The Baldwin and Ford model was a foundation for the design and useful in discussion, but was not operationalized for testing. The design of the training and small sample preclude attributing the findings to any one theoretical application alone. Fifth, the correlations among the instruments did not reach statistical significance at traditional p -levels. Sixth, an argument can be made that the instrument developed to measure skill in this research (CWDVISS) was too complicated to allow for actual measurement and the standardized measure (PPIF) was too global to capture what was common between them.

Given those understandings, two of the three key theoretical principles were given some support, an instrument was developed and piloted for use with an SC,

cognitive strategies of developing a *to do* list and using mnemonic devices, etc. were developed to assist with self-efficacy, and the scores of the extreme scorers were analyzed to determine what they were doing differently. First, the physical step active of testing (Zull, 2002) offered by the in-training role play and the pretest SC interview probably allowed the subjects to develop a sense of anticipatory beliefs and competence (Bandura, 2001) to do significantly better at the posttest as a group. Second, by addressing the content, processing, and management of feelings generated in the training, the trainer may have struck an alliance with the subjects that inspired self efficacy (Bandura) and most importantly, did not engender fear (Zull). Third, this research may have “hit” the appropriate timeframe for significant skill acquisition, retention, and demonstration for a number of reasons: the approximately 2-week period between the beginning and ending of the experiment may have been short enough to facilitate retention and recall, and possibly, long enough to allow for testing in other practice settings. Fourth, it appeared that the most improved (novice) subject may have been using the metacognitive strategies such as LEGO (Listen, Explain, Give Options) that paralleled the strong 1-9 profile that both extreme scorers showed. The novice also was able to develop and execute a better *to do* list at the posttest, suggesting that he was able to utilize cognitive strategies more effectively. Lastly, the training of both the experienced and novice subjects together did expose those inexperienced subjects to advanced strategies.

Policy Implications and Future Research

Nevertheless, the findings suggest that this could be a useful tool for exploring how PCW workers actually apply their training. It was speculated that the novice might have been able to use the cognitive strategies more effectively in the posttest because of his exposure to the more experienced peers in the training group. Future experiments with a focus on this aspect of the experiment could tease out the cost benefit analysis of training novices and experts together. Although the 1-day training provided as a part of this study seemed to make a significant difference in the quantitative measurement of the interviewing skills on the CWDVISS, the qualitative analysis and the SCs evaluation suggests subjects appear to need even more training. This raises the issue of the relationship of statistical significance to clinical adequacy. While it does seem to say that PCW workers need more training to be clinically adequate, there is an issue of the individual subjects' characteristics (Baldwin & Ford, 1988) that plays a role in training's transfer. This was the inspiration for analyzing the extreme scorers. The expert subject was doing something very differently from the novice: the expert built the relationship before he explored difficult subjects, so his strength appeared to be in his early rapport and trust building. This is an important finding that tends to reinforce what seasoned social work teachers already teach (Carrillo, et al., 1993).

The inclusion of SCs in this research as a proxy for clients has implications for the future role of current or former clients in this kind of skill training and research design. Increasingly, PCW agencies are moving toward a long-delayed consumer consciousness. The Institute for the Advancement of Social Worker Research (IASWR)

has recently addressed this issue in its Workforce and Accountability report, which identified agencies where families are being engaged in the ASFA outcomes review process and are more aware of their rights (IASWR, 2004). Finding opportunities for workers and clients to work in mutually designed training and research partnerships participates in a form of reciprocity that can help reduce the adversarial nature and power disparity that surfaces in these interviews (Freire, 1993; Garcia, Sivak, & Tibrewal, 2003). Ultimately, it is the relationship between the worker and the parent that allows any appraisal of the parent's protective capacity to take place; thus, interviewing skills that help PCW workers to develop a relationship with their clients are integral to PCW's mission and success.

Given looming federal budget deficits and contention over ongoing war funding, congress may once again threaten to restructure Title IV-E funding, the heart of PCW service and training funds. An overhaul may combine and "cap" direct service funds with training funds. This, in turn, could affect both PCW training and MSW education as they would then be likely to receive fewer funds over time to prepare the workforce for navigating the nuanced and substantive skills described in this research. Unfortunately, as this research indicates, subjects need more, not less training. What is recommended is that PCW workers and MSW students who are preparing for PCW work be trained in ways that allow for measurable skill demonstration, and that SCs or clients be included in the training, research, and outcome review process. These organizations might then be able to enhance their case for insuring ongoing stable training funds in this important public policy area.

It is imperative that CalSWEC continue to publicize its efforts to evaluate PCW training and establish training's impact on practice outcomes (Johnson, 2005). While this experiment established that a brief training did have a significant relationship to the small trained group's skill improvement, we need more research, larger studies, and more forums to disseminate this information. Database and web-based searches for PCW training evaluation outcome research yield notably few results, until one visits the CalSWEC website. Hopefully, other states and research universities will follow this example and create a place where states can share their training practice and policy dilemmas honestly, in the hope that we will become more proficient in connecting measurable outcomes to training.

REFERENCES

- Badger, L., & MacNeil, G. (2002). Standardized clients in the classroom: A novel instructional technique for social work educators. *Research on Social Work Practice, 12*(3), 364-374.
- Baldwin, T., & Ford, J. (1988). Transfer of training: A review and directions for future research. *Personal Psychology, 41*(1), 63-105.
- Bandura, A. (1977). Self-efficacy: Toward a unified theory of behavioral change. *Psychological Review, 84*, 191-215.
- Bandura, A. (2001). Social cognitive learning theory: An angentic perspective [Electronic version]. *Annual Reviews Psychology, 52*, 1-26.
- Besharov, D. (1987). *Child abuse and neglect reporting and investigation: Policy guidelines for decision making*. Washington, DC: American Bar Association.
- Britner, P. A., & Mossler, D. G. (2002). Professionals' decision-making about out-of-home placements following instances of child abuse. *Child Abuse and Neglect, 26*(4), 317-332.
- Brittain, C. (2000). The effect of a supportive organizational environment on transfer of training in child welfare programs. Doctoral dissertation (unpublished). University of Colorado at Denver.
- California Social Work Education Center. (1999). *Tenth year anniversary report*. Retrieved December 9, 2006 from University of California at Berkeley, California Social Work Education Center Web site: <http://calswec.berkeley.edu>
- Carillo, D., Gallart, J., & Thyer, B. (1993). Training MSW students in interviewing skills: An empirical assessment. *Arete, 18*(1), 12-19.
- Chalk, R., & King, P. (Eds.). (1998). *Assessing family violence interventions: Linking programs to research-based strategies*. Washington, DC: National Academy Press.
- Colorado State University. (2003). *Writing @ CSU writing guidelines. Overview: Content analysis*. Retrieved May 15, 2003 from <http://writing.colostate.edu/references/research/content/index.efm>
- Friend, C. (2009). *Helping child welfare workers learn interviewing skills: A research report*. Berkeley: University of California at Berkeley, California Social Work Education Center.

- Costin, L., Karger, H., & Stoesz, D. (1996). *The politics of child abuse in America*. New York: Oxford University Press.
- Cournoyer, B. (2004). *The evidence-based social work skills book*. Boston: Allyn & Bacon.
- DeVellis, R. (1991). *Scale development: Theory and applications*. Sage Applied Social Research Methods Series, 26. Newbury Park, CA: Sage.
- Edleson, J. L. (1996). Controversy and change in batterers' treatment programs. In J. L. Edleson & Z. C. Eisikovits (Eds.), *Future interventions with battered women and their families*. Thousand Oaks, CA: Sage.
- Finkelhor, D. (1993). The main problem is still underreporting, not over-reporting. In R. Gelles & D. Loske (Eds.), *Current controversies in family violence*. Newbury Park, CA: Sage.
- Finn, J., & Rose, S. (1982). Development and validation of the interview skills role-play test. *Social Work Research & Abstracts*, 18, 21-27.
- Ford, S., Fallowfield, L., & Lewis, S. (1996). Doctor patient interactions in oncology. *Social Science Medicine*, 42, 1511-1519.
- Freeman, K., & Morris, T. (1999). Investigative interviewing with children: Evaluation of the effectiveness of a training program for child protective service workers. *Child Abuse & Neglect*, 23, 701-713.
- Freire, P. (1993). *Pedagogy of the oppressed*. New York: Continuum Publishing.
- Friend, C. (2000). Woman abuse and child protection: A tumultuous marriage. *Children and Youth Services Review*, 22, 309-314.
- Friend, C., & Mills, L., (2002). *Domestic violence and child protection services: Risk assessment*. In A. Roberts and G. Greene (Eds.), *Social workers' desk reference*. New York: Oxford University Press.
- Friend, C., Mills, L., & Hoang, P. (1997). Assessment and case management of domestic violence in public child welfare. Berkeley: University of California at Berkeley, California Social Work Education Center.
- Friend, C., Mills, L., Hoang, P., Maxwell, E., & Rubin, J. J. (1999). *Interventions in domestic violence and child maltreatment: An innovative training program for child welfare workers*. Washington, DC: United States Department of Health and Human Services.

- Gelles, R. (2000). How evaluation research can help reform and improve the child welfare system. *Journal of Aggression Maltreatment & Trauma*, 3(1), 7-28.
- Institute for the Advancement of Social Work Research. (2004). *Workforce and accountability: Child and family services reviews—Implications for child welfare practice, and IASWR child welfare workforce initiative*. Retrieved March 24, 2004, from www.iaswresearch.org
- Jackson, H., & Nuttall, R. (1997). *Childhood abuse: Effects on clinicians' personal and professional lives*. Thousand Oaks, CA: Sage.
- Johnson, B. (2004). *An introduction to the eighth annual national human services training evaluation symposium*. Retrieved January 6, 2007 from: <http://calswec.berkeley.edu/CalSWEC/Publications.html>
- Kerlinger, L. (2000). *Foundations of behavioral research* (4th ed.). Orlando, FL: Harcourt College Publishers.
- Kolb, D. (1984). *Experiential learning: Experience as the sources of learning and development*. Englewood Cliffs, New Jersey: Prentice Hall.
- Kolb, D., Boyatzis, R., & Charalampos, M. (2000). Experiential theory: Previous research and new directions. In R. Sternberg & L. Zhang (Eds.), *Perspective on cognitive learning and thinking styles*. Mahwah, NJ: Lawrence Erlbaum Publishing.
- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Newbury Park, CA: Sage.
- Lindsey, D. (2003). *The welfare of children* (2nd ed.). New York: Oxford Press.
- Linsk, N., & Tunney, K. (1997). Learning to care: Use of practice simulation to train health social workers. *Journal of Social Work Education*, 33(3), 473-489.
- McDonald, J., & McCartney, B. (1999). *Effective partnership models between the state agencies, community, the university and the community service providers in changing paradigms of child welfare practice: Responding to opportunities and challenges*. Monograph from symposium held June, 1999, 43-72. Washington, DC.
- Miller, W., & Crabtree, B. (1992). Primary care research: A multi-method typology and qualitative road map. In B. Crabtree & W. Miller (Eds.), *Doing qualitative research*. Sage: Newbury Park, CA.

- National Research Council. (2002). *How people learn: Brain, mind, experience and school*. Washington, DC: National Academy Press.
- Ong, L., Visser, M., Van Zuuren, F., Rietbroek, R., Lammes, F. & De Haes, J. (1999). Cancer patients' coping styles and doctor-patient communication. *Psycho-oncology*, 8, 155-166.
- Garcia, J., Sivak, P., & Tibrewal, S. (2003). Transforming relationships in practice and research: What is the Stanislaus model?, *Protecting Children*, 18(2), 22-29.
- Perry, R. E. (2006). Do social workers make better child welfare workers than non-social workers? *Research on Social Work Practice*, 16(4), 392-405.
- Schinke, S., Blythe, B., Gilchrist, L., & Smith, T. (1980). Developing intake-interviewing skills. *Social Work Research & Abstracts*, 16(4), 29-34.
- Schinke, S., Smith, T., Gilchrist, L., & Wong, S. (1978). Interviewing-skills training: An empirical evaluation. *Journal of Social Service Research*, 1(4), 391-401.
- Schwartz, I., & Fishman, G. (1999). *Kids raised by the government*. Westport, CT: Praeger.
- Skinner, B. F. (1976). *About behaviorism*. New York: Random House.
- Sedlack, A., & Broadhurst, D. (1996). *Third national incidence study of child abuse and neglect*. Washington, DC: U.S. Department of Health and Human Services.
- Stevenson, K. M., Leung, P., & Cheung, K. M. (1992). Competency based evaluation of interviewing skills in child sexual abuse cases. *Social Work Research & Abstracts*, 28, 11-16.
- US Department of Health and Human Services, Administration on Children, Youth, and Families. (2004). *Child maltreatment 2004: Reports from the states to the National Child Abuse and Neglect Data System*. Retrieved December 10, 2006, from <http://www.acf.hhs.gov/programs/cb/pubs/cm04/chaptersix.htm#post>
- Zellman, G., & Fair, C. (2002). Preventing and reporting abuse. In J. E. B. Myers, L. Berliner, J. N. Briere, C. T. Hendrix, T. A. Reid, & C. A. Jenny (Eds.), *The APSAC handbook on child maltreatment*. Thousand Oaks, CA: Sage.
- Zell, M. C. (2006). Child welfare workers: Who they are and how they view the child welfare system. *Child Welfare*, 85, 83-103.

Zull, J. (2002). *The art of changing the brain: Enriching the practice of teaching by exploring the biology of learning*. Sterling, VA: Stylus Publishing.

Friend, C. (2009). *Helping child welfare workers learn interviewing skills: A research report*. Berkeley: University of California at Berkeley, California Social Work Education Center.