TECHNICAL REPORT

Evaluating Child Welfare Programs

CONTENTS

FEBRUARY 1, 2018

Bruce A. Thyer, PhD, LCSW, BCBA-D
Distinguished Research Professor

Florida State University College of Social Work
296 Champions Way, Tallahassee, FL 32306
Bthyer@fsu.edu | 850-645-4792

KEY WORDS

child welfare, program evaluation, outcome studies, evidence-based practice

Funded through a contract with the Florida Institute for Child Welfare

FLORIDA INSTITUTE FOR CHILD WELFARE
AT FLORIDA STATE UNIVERSITY

Executive Summary

This technical report reviews the importance of individual child welfare agencies conducting periodic program evaluations of their own services' outcomes. It is important that the evaluations make use of reliable and valid outcome measures that directly pertain to the agency's purposes, and be as specific as possible. Variables that are very broad and general are not as useful as more direct measures of the issues facing clients. This technical report provides information on how to locate suitable outcome measures. This report offers a review of practical pre-experimental and quasi-experimental research designs that are widely used in program evaluations. The designs are illustrated with published examples of their use. Fundamental information on conducting statistical analyses of program outcomes is also provided. Experimental designs, studies that involved randomly assigning clients or families to differing treatment conditions are briefly described, but their use is less emphasized in favor of the more practical pre-experimental and quasi-experimental evaluation studies. Information is provided on how to format program evaluation reports. A policy recommendation is made that each child welfare agency should attempt to systematically collect pre- and post-service outcome measures on their clients' functioning, periodically assess this aggregated information, and use this data to make empirically-based decisions about the need to revise, expand, contract or terminate existing services. The approach outlined in this report is consistent with large-scale initiatives on the part of the federal and state governments, community-based care lead agencies (CBC), providers, and the United Way to request empirical outcomes data on funded services. Faculty at nearby universities can usually be contacted to provide free or low-cost consultation to child welfare agencies in the design and conduct of program evaluation studies.

What is Known?

Child welfare agencies tend to focus their evaluation efforts on process and service delivery measures, data involving the number of clients served, time to close cases or complete investigations, new intakes, discharges, number of clients who complete treatment, and so forth. Accreditation reviews similarly focus on process features of service delivery, factors such as the completeness of client records, the qualifications of human service providers, duration of waiting lists, premature terminations, etc. Such a focus on inputs and process is similarly found when child welfare agencies are asked by the Florida Department of Child and Families (DCF) to evaluate agency programs. Recently, there is increased attention given to issues related to evaluation *outcomes*, not only processes and service delivery features. This has not filtered down to the level of agencies and programs so that they regularly conduct evaluations of outcomes.

What is Important – Why it is Important?

In a very real sense, the features surrounding child welfare processes and service delivery measures are surrogates or indirect measures of outcomes evaluations. It can be assumed that if the process and service delivery measures are poor, clinical outcomes will not be good. However, the converse is not true. An agency could be achieving high-quality measures of process and service delivery, but the outcomes of these services could nevertheless be unsatisfactory, e.g., children and families could not be improving as assessed by measures pertaining to their presenting problem. Analogous to a client who completes mandated substance abuse treatment, but who continues to abuse alcohol or other drugs, parents deemed at high risk for abuse could complete a parent-training program and still remain at high risk to abuse their child. Or, foster parents could complete all the training needed to become certified foster parents, yet still treat their foster youth in an aversive manner. Such instances illustrate why it is important that both public and private agencies engage in regular evaluations of the outcomes of their services in addition to collecting data on process and service delivery. At the national and some state levels, agencies and programs are being asked to collect outcome measures on their services. Occasionally these mandates involve an imposed template—outcome measures are pre-specified, as is who must collect the data, what clients must be contacted, and when data is gathered—all established by the funding agency and reported periodically. In such instances, child welfare agencies have little latitude in terms of deviating from the imposed program evaluation design. Sometimes agencies may wish to go beyond a state or federally-mandated evaluation plan and conduct an assessment of program outcomes. This may be desired when an idiosyncratic program does not neatly fit into the mandated evaluation, or when the mandated evaluation is deemed to be inadequate by local agency managers or program administrators. In this case, while faithfully carrying out the mandated reporting of process and services, additional efforts may be undertaken at the local level. These efforts can be well developed, designed, carried out and reported, but many times they have limitations. There are many barriers to conducting an effective evaluation of program outcomes. Some of these barriers are structural, or involve a lack of resources. Other times they may involve a lack of local expertise, or child welfare agency administrators may simply not possess sufficient knowledge or skills to design and successfully carry out a local-level evaluation of outcomes. In the worst case, a poorly designed or carried out evaluation plan may produce inaccurate results leading to incorrect conclusions about program outcomes.

The professional codes of ethics of the major helping professions all assert the responsibility of practitioners to engage in the systematic evaluation of their practice outcomes at the clinical and programmatic level.

For example, the *Code of Ethics of the National Association of Social Workers* states:

> **5.02 Evaluation and Research**
>
> **a)** a) Social workers should monitor and evaluate policies, the implementation of programs, and practice interventions.
>
> **b)** Social workers should promote and facilitate evaluation and research to contribute to the development of knowledge.
>
> **c)** Social workers should critically examine and keep current with emerging knowledge relevant to social work and fully use evaluation and research evidence in their professional practice.

The *Principles of Medical Ethics* as they pertain to psychiatry state that:

> **Section 5:** A physician shall continue to study, apply, and advance scientific knowledge, maintain a commitment to medical education, make relevant information available to patients, colleagues, and the public, obtain consultation, and use the talents of other health professionals when indicated.

The *Code of Ethics for Nurses* asserts that:

> All nurses must participate in the advancement of the profession through knowledge development, evaluation, dissemination, and application to practice. Knowledge development relies chiefly, though not exclusively, upon research and scholarly inquiry. Nurses engage in scholarly inquiry in order to expand the body of knowledge that forms and advances the theory and practice of the discipline in all its spheres…Dissemination of research findings, regardless of results, is an essential part of respect for participants… dissemination of findings is fundamental to ongoing disciplinary discourse and knowledge development.

Thus, the theme of evaluating program outcomes seems very important across a range of human service and health care professions. This echoes sentiments expressed by former President Barak Obama in his first inaugural address in 2009:

> The question today is not whether our government is too big or too small, but whether it works…Where the answer is yes, we intend to move forward. Where the answer is no, programs will end.

At the state level, DCF Secretary Mike Carroll notes:

> While the work we do in communities will never be done, we are constantly focused on ensuring that our resources are directed to the right places at the right times based on the greatest need, and *that our programs are operating as efficiently and effectively as possible* (italics added, c.f., http://floridafiscalportal.state.fl.us/Document.aspx?ID=14581&DocType=PDF).

Florida is implementing a Results-Oriented Accountability Program, intended to help the state, districts, CBCs, and individual agencies and programs determine if child welfare goals are being met:

> The Results-Oriented Accountability Program will provide the resources and tools Florida needs to improve the lives of the children and families it serves. The Program, which requires quantitative and qualitative data to measure desired outcomes, will enable the Child Welfare system to build a stronger and more evidence-informed operating model. In order to hold stakeholders accountable, they must be measured against the outcomes they are charged with achieving. By measuring and monitoring outcomes over time, the State will have insight into whether its Child Welfare programs and services are having a positive impact on the safety, permanency and well-being of children. Furthermore, through the use of data reported at the system and stakeholder levels, both the Child Welfare system as a whole, and the individual participants, can make better decisions about the interventions most effective in driving outcomes (c.f. http://floridafiscalportal.state.fl.us/Document.aspx?ID=14581&DocType=PDF, p. 22)

Among the core competencies advocated by DCF, we find:

> **Data Analytics:**
>
> Everything we do must be outcome-based and solution-focused. We must analyze data and information in multidimensional ways to gain deep understanding of system issues and challenges. (c.f. emphasis added, http://www.myflfamilies.com/about-us/office-secretary/mission-vision-values)

More historically, the British Scientist Lord Kelvin claimed, "If you cannot measure it you cannot improve it."

The above quotations clearly illustrate what is important—evaluating outcomes of child welfare services. Why is it important? To help improve the quality of care provided to children and their families. Unfortunately, all interventions, including some widely used in child welfare, do not work well. Some have actually been shown to be harmful to children.[1] Thus, evaluating child welfare services in terms of their outcomes is a desirable activity for agencies to undertake.

Purpose of this Technical Report

This technical report provides practical guidance for state officials and local child welfare agency administrators to:

- develop answerable questions related to program outcomes
- locate and select reliable and valid outcome measures
- select an adequate research design to answer the specified question(s)
- conduct the evaluation
- prepare an intelligible report that conveys the results in an understandable manner and can provide useful feedback to the agency regarding the quality and outcomes of the services they provide

Types of Questions that are Important for Agencies to Ask and Answer

Outcome questions can be viewed as falling along a continuum of complexity, ranging from simple questions (requiring simple evaluation designs), to more complex questions relating to making causal inferences (requiring more sophisticated evaluation designs). Here are some questions that can be addressed in program evaluations of service outcomes:

- How *satisfied* are clients with the outcomes of the services our child welfare agency provides?
- How are clients (children, families, caregivers) functioning *after they received* our services?
- Do clients *improve* after receiving our services?
- If clients improve immediately after receiving our services, do the positive results *persist over time*?
- Do clients *improve more* with our services than they would have if they received no services?
- Do clients improve more following receipt of our new or novel service, than they could have if they received *treatment as usual*?
- If clients improved following our services, how confident can we be that these improvements *were caused* by our services, as opposed to being the result of other factors?

The focus of this report is on the design of quantitative outcome studies, as these are the type of designs most commonly used to evaluate program outcomes involving large numbers of clients. Agencies need to select one or more clear questions as the first step in designing an outcome evaluation and it is recommended that they initially attempt some of the simple evaluation designs and successfully complete them, before undertaking a more complex design intended to answer more complex questions.

How to Locate and Choose Reliable and Valid Outcome Measures

One of the first steps in conducting an outcome evaluation measure is determining what to measure across all clients within a given program. When possible, it is best to select one or more measures that can be used to assess ALL clients receiving a common service expected to produce some favorable result. If an agency provides different types of programs having differing goals, then each program should be evaluated using measures appropriate for the services provided. Having such measures completed by the client is one common approach, for example an adolescent who completes a standardized self-esteem measure, who will be participating in a mentoring program that is expected to enhance self-esteem. Sometimes measures can be completed by parents or caregivers—as in a parent rating of child behavior. Other times, outcome measures can be employed that reflect existing data. School performance can be assessed by a child's grade point average the prior term, their absenteeism or tardiness, or by the number of disciplinary referrals. When completed by youth themselves, or by parents/caregivers, measures should be brief, easy to understand, easy to score, sensitive to change, and lend themselves to repeated assessment. A screening instrument may not be

sensitive to change, especially if it contains items that read along the lines of "Have you ever….." Answers to such questions will not usually change, and thus scores on screening measures are insensitive to possible clinical improvements. They may be used for their intended purpose—to screen for eligibility for instance—but not be used for *repeated* assessment purposes to assess *changes*. The Drug Abuse Screening Test for Adolescents noted below is an example of an excellent pre-treatment screening tool but is unsuitable for use as an outcome measure due to its insensitivity to change. Some of its questions include: "Have you used drugs other than those required for medical reasons?" "Have you engaged in illegal activities to obtain drugs?" "Have you ever been in a hospital for medical problems related to your drug use?" A youth's answers to such screening questions would not change, even after months of successful drug abstinence following treatment.

Be careful of using measures that are too long or time intensive. Formal intelligence tests require a skilled test administrator and may take an hour or more to complete and score, and are thus impractical for use as a child welfare agency outcome measure. Similarly, the Minnesota Multiphasic Personality Inventory - Adolescents measure has over 400 items. Again, this measure's length precludes its practicality within the context of most child welfare agency settings.

In conducting an outcome study, it is usually wise to avoid inventing a new outcome measure of client functioning. For such measures to have any scientific credibility they must possess acceptable levels of reliability and validity, and demonstrating this is an ambitious and difficult undertaking, usually beyond the capacity of a given agency. It is always a good practice to carefully search the existing literature for potential outcome measures, giving preference to measures previously published in credible professional journals or books. Outcome measures available solely from online websites, or through internal reports published by some centers, do not have the same level of credibility of measures published in reputable books and journals. This caveat refers to locating and using outcome measures of client functioning that assess some higher level construct such as self-esteem, depression, anxiety, family attachment, etc. Agencies can use more simple and straightforward factors as outcome measures when these are appropriate. Examples may include "Number of children adopted per month", "Number of cases successfully closed", "Number of adoptions which failed within one month", etc. Adolescent arrests, drug test results, days absent from school, and last term's grade point average, all represent examples of important outcome measures that do not need elaborate methods of assessment.

Many online sources can be used to help locate legitimate outcome measures for child welfare agency evaluation purposes. Some of these sources screen and describe available measures, and explain how to locate them, while others provide actual access to the measures themselves. Here are some useful sources for locating outcome measures:

The Child Outcomes Research Consortium
(http://www.corc.uk.net/measurestable.html)

The Consortium provides access to a wide array of free outcome measures, including various factors of relevance to child welfare agencies. Table 1 presents examples of assessment tools that can be used.

Table 1: Assessment Tools and Their Utilization

ASSESSMENT TOOL	AGE RANGE	ASSESSES FOR	ADMINISTRATION	TERMS OF USE
Children's Global Assessment Scale	6-17	Psychological and social functioning	Completed by clinician	Free
Revised Children's Anxiety and Depression Scale	8-18	Frequency of anxiety and low mood	5-10 minutes	Copyrighted
The Strengths and Difficulties Questionnaire	11-17	Emotional and behavioral screening	5-10 minutes	License required can be obtained from NHS Digital
Patient Health Questionnaire	18-99	Mental health disorders	Self-administered	Free
Children's Revised Impact of Events Scale	8-18	Children at risk for Post-Traumatic Stress Disorder	Self-reported	Free
Me and My Feelings (Me and My School)	8+	Child mental health	10 minutes	License required can be obtained from NHS Digital
Beck Youth Inventories	7-18	Symptoms of depression, anxiety, anger, disruptive behavior, self-concept	Paper/pencil/ verbal	Purchase the BYI manual, kit and/or booklets.
Brief Parental Self-Efficacy Scale	Parents	Confidence in raising child	Self-administered	Free
Mood and Feelings Questionnaire	6-17	Depression in children	Child self report	Free
Drug Abuse Screening Test	Any	Drug use not alcohol or tobacco	< 8 minutes, self-report or by a clinician	Free
PedsQL Measurement Model for the Pediatric Quality of Life Inventory	2-4, 5-7, 8-12, 13-18	Quality of life in children	5 minutes, self-administered	License fees vary according to study type and financing
How Are Things	Parents	Behavioral difficulties	Paper questionnaire	License required can be obtained from NHS Digital
Students Life Satisfaction Scale	8-18	General statements about their life	Self-administered	Free
The WHO (Five) Well-being Index (WHO-5)	9 and above	Mental well-being	Variety of settings	Free
Student Resilience Survey	7 and older	Students perceptions of their individual characteristics as well as protective factors embedded in the environment	Child-reported version	Free
Eating Disorder Examination	14+	Severity of Eating Disorders	Self-reporting questionnaire	Copyright-free
Systematic Clinical Outcome and Routine Evaluation	12+	Family life and the need for therapy and therapeutic Change	At or before relevant sessions	License required can be obtained from NHS Digital

An example of a free source which provides a direct link to one outcome measure only, in this case the Hamilton Depression Rating Scale, can be found at: http://healthnet.umassmed.edu/mhealth/HAMD.pdf

Several measures of *Child and Youth Resilience* can be obtained via this website: http://cyrm.resilienceresearch.org

The *PsycTESTS Database* is a paid subscription available at http://www.apa.org/pubs/databases/psyctests/field-guide.aspx and contains information on thousands of psychosocial assessment instruments, including copies of the scales. Evaluators can search the database by age (e.g., child and adolescents) and by problem area. The included measures are scientifically credible and summaries are provided about the evidence supporting their use. The local university library or a university faculty member may be able to provide agency staff with access to this site. The agency can also purchase access to it.

The *PsycINFO Database* http://www.apa.org/pubs/databases/psycinfo/index.aspx is a user-friendly way to search through thousands of journals and books. It is maintained by the American Psychological Association. Search terms can be entered (e.g., scale, assessment measure, rating, test, inventory, etc.) for a specific problem or issue (anxiety, trauma, self-esteem, parent relationship, life skills, etc.), a date range, (e.g., published since 2010), and an article based on search criteria will appear on a list. Click on the link and the article should come up as a PDF. *PsycTESTS* and *PsycINFO* is a subscription-based service, but some can access it free at a local university library or public library.

Sage Journals

Some of the larger commercial publishers, which usually charge a fee to access their online journal content (and most are online these days), offer a limited window of free access. One publisher, *Sage Publications*, produces over 700 journals, many of which are in the areas of child welfare, social work, psychology, family therapy, mental health counseling, etc. Usually in November of each year, Sage allows free and unlimited access to their entire journals' database, found here: http://journals.sagepub.com/search/advanced?SeriesKey=rswa. Evaluators should plan ahead and enter their system and conduct searches for assessment measures, and print out the PDF articles at no charge during this one month access period. This resource is more limited than *PsycINFO*, which covers almost all publishers whereas, those maintained by individual publishers only provide access to the journals they produce. But some journals have a variety of scales. For example, the journal *Research on Social Work Practice*, produced by Sage Publications, has published hundreds of articles about assessment measures. One recent article is a systematic review of social-emotional screening instruments for young children in foster care.[2] The article contains useful information about the scientific credibility and use of 24 different screening instruments intended for use with children (ages 10 and under) in foster care. Child welfare agencies would find this information valuable and should become familiar with the instruments. Agency leadership and evaluators can browse all of the issues of this journal, and find citations and abstracts for free, here: http://journals.sagepub.com/loi/rsw

Psychological Assessment Resources (PAR) (see http://www4.parinc.com/)

A number of commercial firms sell copies of scoring guides for various screening and assessment measures. One of these, given as an example only, and not intended as a specific endorsement, is *Psychological Assessment Resources*. They have hundreds of scales for sale, and a number of these have applicability in child welfare settings. In their 'child abuse/custody' section there are measures such as the *Checklist for Child Abuse Evaluation*, *Child Abuse Potential Inventory*, *Child Sexual Behavior Inventory*, *Parenting Alliance Measures*, *Parenting Stress Index, Stress Index for Parents of Youth Children*, and the *Trauma Symptom Checklist for Children* (see http://www4.parinc.com/products/ProductListByCategory.aspx?Category=FORENSIC&SubCategory=CHILD_ABUSE/CUSTODY#)

PAR offers free online training in using their measures (see http://www4.parinc.com/page/traininglist.aspx

It is necessary to pay a modest fee to acquire official copies of these copyrighted assessment measures, with pricing approximately $70 for a booklet of 50 scales, and another $70 for a user and scoring manual. The price depends on the instrument and its complexity. However, by incorporating assessment measures with good research evidence to support their use, a child welfare agency will have taken a big step forward in making legitimate program evaluation projects possible. It may be worth it for an agency to acquire one or more scales and pilot their use.

Assessment and Measurement Books

One excellent book that contains useful outcome measures is titled *Measures for Clinical Practice*.[3] One of the two volumes contains copies of dozens of outcome measures for use in evaluating practice outcomes with children, youth, and families, along with scoring instructions and citations to the published research supporting each scale (a separate volume contains measures for use with adults). Many child welfare agencies would find it useful to obtain a copy of this book and make it available for its service providers to review and locate practice measures.

A more focused resource book is titled *Practitioner's Guide to Empirically-based Measures of Depression*.[4] The Association for Behavioral and Cognitive Therapies sponsors several problem-focused books that present additional compilations of assessment measures dealing with a given issue. Available separate titles include empirically-based measures of anger, aggression and violence, anxiety, social skills, and school functioning. These are available here: https://www.abctcentral.org/eStore/index.cfm?mz=110&prid=75&s_category_id=4

Agency administrators, managers, and practitioners should not assume that there are no existing measures available for use in evaluating one's program. There is usually something suitable. It just may require some searching. If options are limited, then a request for help from faculty with child welfare interests at a nearby university may prove useful. Many useful instruments are free; some must be purchased individually. Some include an option to purchase an agency license to use a set amount of copies of a given scale per year.

Types of Designs that Can Answer Important Questions About Outcomes

Program evaluation outcome questions can be answered using different quantitative research designs. For the purposes of diagramming these designs, the letter O is used to indicate a time when clients are observed or assessed, and X indicates the provision of services. These designs can be grouped as follows:

Pre-experimental Designs

Examples of the post-test-only design, diagrammed as X – O_1

This design simply involves an agency selecting one or more reliable and valid outcome measures that clearly relate to client outcomes, and assessing these after the receipt of services. In the above diagram, X refers to the clients receiving the intervention, and O_1 refers to a time when the clients exposed to the program are assessed or their functioning is measured. If an agency claims that its purpose is to help achieve some specified goal with its clients, then it is necessary to empirically determine how the clients are doing—perhaps immediately after services have been terminated, or after some reasonable amount of time later. A practice example would be to systematically assess the durability of foster care placements, or adoptions, some reasonable time after a child has been placed. If an agency finds that 50 percent of placements failed after one year, this would likely be seen as a non-satisfactory outcome. If 95 percent of the placements were maintained after one year, this would be seen as a far better outcome. Another example would be if an agency is tasked with helping homeless families find long-term, safe, affordable housing, and considers a family's case 'closed' when they have placed a family in a home, it would be useful to determine the housing status of each family after one year. If the placement rate remains at 90 percent 12 months later, this can be seen as a good outcome. A placement rate of 30 percent, much less so. This design is also widely used by agencies to conduct client satisfaction studies.

Here are examples of the use of the post-test-only design by child welfare agencies. It should be mentioned that the studies described are intended to illustrate the use of program evaluation designs, not to suggest that interventions described should be adopted by child welfare agencies.

The *Marcus Institute for Development and Learning* in Atlanta, Georgia provided comprehensive interdisciplinary team evaluations for children with disabilities and their families. Following the evaluation, their services included providing the caregivers with practical information on how to obtain social, therapeutic, family, and medical services. The evaluations and subsequent recommendations took over four hours to complete and the agency staff were understandably interested in the extent to which the families actually followed up on recommended services. If the families actually obtained few of the recommended services, then the evaluation process could be seen as a waste of time. Efforts were made to contact all 51 families evaluated in the prior year (none more recently than four months previously) to get their permission to be interviewed by telephone about their follow-up on recommended services. Fourteen families could not be reached and one declined to participate, leaving a sample of 36 who were interviewed. The average age of the child member of the participating families was 21 months, and the sample of toddlers consisted of 25 boys and 11 girls. Caregivers were interviewed by telephone as to whether they had obtained services for their child. Overall, 79 percent of the recommendations had been followed and the families were receiving the suggested services. Specifically, 89 percent of the recommended medical services had been obtained, 84 percent of the educational services, and 46 percent of the recommended social services were being received. Some barriers to the families obtaining needed services were also identified. Overall, the staff at the Marcus Center were very pleased with the families' positive follow-through with the agency recommendations. The identification of barriers encountered by the families enabled them to anticipate such problems among the subsequent families they served.[5]

Examples of the pre-test-post-test design, diagrammed as O_1 – X – O_2

In this design, clients (parents, children, caregivers, teachers, other stakeholders) are asked to complete one or more measures intended to assess current functioning, prior to enrollment into some child welfare service. At the point of planned termination of services, discharge, or unexpected dropout, the client is asked to provide another identical measure of their functioning. In this diagram, O_1 refers to the first assessment, in this instance before the clients' receipt of services, and O_2 is the second assessment, conducted after services were completed. Alternatively, a second (and third, etc.) assessment could be undertaken after some duration of time (after completing three months of treatment). Over time, the numbers of clients served by the child welfare agency can be aggregated, and the pre-and post-treatment measures summarized in some manner (e.g., an average is computed for the first assessment and again for the second time, third, etc.). The desired sample size for a study like this should be at least 12 or so clients, all receiving the same agency's services, and ideally presenting with similar problems. Each client, for this design to make sense, should be assessed using the same outcome measure(s). These overall pre- and post-test measures can be subjected to simple statistical analysis to see if any observed changes were statistically significant or not. If they were statistically significant, an effect size measure can be calculated to help determine the clinical impact of the change. In the case of data measured with mean scores pre-test and post-test, the appropriate statistical test would be a paired sample *t* test.

The author has conducted such studies with a wide array of agencies over the years, often with the help of various graduate students interning within the agency. One such evaluation took place at a private psychiatric hospital located in Macon, Georgia.[6] Fifteen clinically depressed adolescents who were consecutively admitted to the unit were asked to complete three measures of mental health (two measures of depression and one measure of self-esteem) at the time of admission. At discharge, after an average stay of 28 days, these same measures were completed by the youth once again. Statistically significant and clinically meaningful improvements were found on all three outcome measures. Thus, the question, "Do adolescents improve over the course of their hospitalization?" could be answered in the affirmative. Remember, most agencies cannot answer this simple question with actual data, so an elementary investigation of this type is a good first step in undertaking a systematic program evaluation. Simple studies such as this are also a good way for

practitioners to learn the skills needed to evaluate outcomes, in a relatively uncomplicated manner. If an agency lacks sufficient internal staff resources to do a study like this, outreach to a local university's departments of social work, psychology, family therapy, or nursing may help locate faculty or graduate students interested in taking the lead on a project of this nature.

The study described above set the stage for a later, larger study conducted on a different psychiatric unit in a different city.[7] Thirty-six consecutively-admitted child patients completed the well-known *Child Behavior Checklist* (CBCL) upon admission and again at discharge, after an average stay of 55 days. Both the CBCL internalizing and externalizing behavior subscales demonstrated statistically significant and clinically meaningful improvements over time. The children presented with an array of mental health diagnoses and received the usual complex menu of services during their stay. It is not possible, based on these data to say exactly what components of the treatment program they received may have been responsible for their improvements. Indeed, they may have improved simply because of the passage of time. But the question posed did not address those issues. It was more simply "Do the children we treat improve over the course of their stay?" which was answered in the affirmative.

A school social worker in New Orleans was working with a population of predominantly Hispanic youth, many of whom had been exposed to traumatic events, and as a consequence, developed symptoms of depression and post-traumatic stress disorder. The social worker, herself Hispanic and fluent in Spanish, had received training in a brief, standardized skill-based group intervention intended for traumatized youth, a program called Cognitive Behavioral Intervention for Trauma in Schools, or CBITS. As a school social worker, she was providing CBITS to students needing trauma therapy. After obtaining informed consent from parents and informed assent from the youth, she had the children complete the Child Post-Traumatic Symptom Scale, and the Short Mood and Feelings Questionnaire as pre-treatment measures of trauma symptoms. Children were enrolled in her 10-week group therapy sessions according to grade. Sessions were conducted in Spanish, and Spanish translations of the CBITS handouts and workbook were used. At the conclusion of the program, she had pre- and post-therapy data on 23 Latino youth, of whom 61 percent were girls, and the mean age was 12 years old. Post-treatment, the youth on average displayed statistically and clinically significant reductions in trauma symptoms and improvements in mood. By building these validated measures of trauma into the clinical services she was already providing, the school social worker was able to answer, with some credibility, the question "Do my Hispanic child clients with serious trauma symptoms improve after they receive the CBITS services I deliver?" The answer was yes. The time required for her to score the child-completed measures was relatively brief, as was the effort needed to conduct the simple statistical analysis. There was no external funding needed to undertake this project. Further details can be found in Allison and Ferreira.[8]

A foster care agency in Chicago was well established, but despite opportunities to transition a foster care placement into a permanent adoption of the child by the foster family, such transitions (seen as a more desirable outcome than long-term foster care) were not occurring. Harold Briggs, a social worker with considerable experience in this area was hired as a consultant to promote the foster care workers' successful transition of children on their caseloads from the status as a child in foster care, to permanently adopted by that same child's foster family. In the three years prior to Mr. Briggs' position as a consultant, there were zero foster care to adoption transitions. In the first year of Mr. Briggs' consultancy, there was one such placement. The next year there were 5, and 15 in the third year. The program he designed to increase adoptions involved some staff training (above and beyond the usual training provided by the agency) related to adoption policies, staff time management skills and case planning; foster parent training; clinical consultations; and specialized monthly supervision.

Thirty-one clients, with a mean age of 12 years old, were consecutively enrolled in a Trauma Recovery Program (TRP), operated in an urban child welfare agency. The Trauma Symptom Checklist for Children (TSCC) was used to assess the children's trauma symptomatology. The checklist includes subscales evaluating anxiety, depression, PTSD, dissociation, anger, and sexual concerns. The clients were assessed upon admission to the program, and again after three months of participation. Assessments continued after every three months while the children remained in therapy. This project reported the changes from pre-treatment to about three months post-treatment, (after the child completed TRP care). The duration of treatment varied widely across the youth, from 2 to 26 months of care. It should be noted that during the time participants were recruited for this study, a total of 319 children were referred to the TRP and 184 cases were opened. Of these, 142 (77%) completed the intake process (four sessions), and 122 remained in treatment long enough to be eligible for reporting (attending at least 12 sessions or three months of services). Of the 122, pre-test and post-test data for only 31 children were available in the agency's records. The results were positive. Each TSCC subscale displayed statistically significant and clinically important improvements, suggesting that the traumatized youth actually improved over the course of their participation in the TRP. This large rate of attrition is common among evaluation projects conducted in child welfare agencies and should not be used as a rationale to not attempt to undertake such evaluations. Some data is better than no data, and learning how all 31 youth who received a sufficient number of clinical services to be considered a legitimate test of the program is certainly desirable. Remember that the purpose of an evaluation study, such as is described in this technical report, is to ascertain outcomes at the agency-level, not to try and make generalizable conclusions about a given form of therapy that potentially extends beyond one immediate practice setting. The latter endeavor usually requires a randomly sampled study from a larger population of interest (e.g., all traumatized youth) and this is not usually possible in conducting agency-based outcome studies.[9]

Lauren's Kids, a non-profit organization, developed the Safer, Smarter Kids program and provides training to teachers in various Florida schools to learn how to teach the sexual abuse prevention curriculum to young children. It consists of six 30-minute sessions and uses video materials, structured learning, and class exercises. The program is available for free to all Florida public schools, and a website provides curriculum support for teachers. Schools in Florida were asked to participate in an evaluation of the program, and 86 schools in 4 counties were included. Teachers, counselors and school social workers attended a live webinar training. Students in the selected classrooms were administered an 11-item evaluation questionnaire. 1,169 kindergarteners attended all six sessions of the Safer, Smarter Kids program and completed pre-tests and post-tests. A simple paired-sample *t*-test was used to compare possible differences in post-test scores, compared to the pre-test.

Gains were fairly substantial, with increases about 77 percent, as well as statistically significant, reflecting increased knowledge of key prevention concepts.[10]

Researchers[11] used this form of design to evaluate a Florida, urban outpatient treatment program (Crisis Center of Tampa Bay) for youth with problematic sexual behaviors. A total of 28 cases were retrieved from agency files. The youths' ages ranged from 11-17 and they were enrolled in the Youth Sexual Behavior Problem (YSBP) outpatient program. All were male. There were three outcome measures, a rating form estimating risk of adolescent sexual offending, the Child Global Assessment Scale (both completed by treating clinicians), and the Parenting Stress Index. Significant improvements in sexually related issues were obtained on 5 of 7 outcomes.

Youth (aged 10-17) receiving juvenile justice services in Florida participated in a Prodigy Cultural Arts Program located in Tampa. The program provided training in the visual, performing, musical, media and theater arts, with classes taught by master artists from the local community. The program's purpose is to promote healthy development and discourage harmful behavior. The length of the program lasts 8 weeks, with about 3 hours a week of activities. Reliable and valid measures were taken of the youths' mental health, delinquency, and family functioning pre- and post-treatment. Paired *t*-tests were used to assess improvements on the outcome measures for approximately 350 participants, youth, and their parents. Statistically significant and clinically meaningful improvements were found on all measures, arguing well for the potential effectiveness of the program.

As can be seen, this $O_1 – X – O_2$ pre-test / post-test design is useful in evaluating child welfare interventions. It is capable of answering simple questions, such as "Do clients receiving our services improve?" and "Do they get worse?" Usually when dealing with data obtained from groups of individuals, some type of statistical analysis is needed to help remove human bias from judging whether or not change occurred. Child welfare agencies may have in-house staff with expertise in conducting appropriate statistical tests. But many common inferential tests (those with testing whether or not change occurred pre-post-treatment, or if two groups differ after treatment), wherein the pre-test and post-test data are presented in terms of means (average scores) at each of these two time points, are available using easy-to-use online statistics calculators. For the $O_1 – X – O_2$ design, a common test is called the t-test for paired samples, and a staff member can input the data (either individual scores, or previously calculated mean scores) into the online calculator, click a button and obtain the result. Here are two online calculators that can be used for this purpose:

t-Test Online Calculators

http://vassarstats.net/tu.html

https://www.graphpad.com/quickcalcs/ttest1/?Format=SD

Between two groups

http://www.socscistatistics.com/tests/studentttest/Default2.aspx

https://select-statistics.co.uk/calculators/two-sample-t-test-calculator

If the data represent frequencies, not means, of the data, then the Chi-Square (χ_2) test would be an appropriate statistical test. Online calculators for this measure can be found here:

http://www.socscistatistics.com/tests/chisquare/Default2.aspx

http://www.socscistatistics.com/tests/chisquare2/Default2.aspx

The pre-test / post-test design can be improved using several methods. One is to have more than one pre-test. This allows a better determination of client functioning before they received the formal child welfare intervention. This design could be diagrammed as $O_1 – O_2 – X – O_3$. Another way to improve this design is to have more than one post-test, in other words, a follow-up period. This could look like $O_1 – X – O_2 – O_3$. Follow-up assessments are needed to see if any initial improvements were maintained—as sometimes interventions produce immediate improvements that wear off soon—which is not a desirable state of affairs. Or sometimes it takes a while for an intervention to yield an effect, which could only be detected if a follow-up assessment covering a suitable time period was undertaken. A researcher conducted a study of the effectiveness of suicide gatekeeper training for child welfare workers, educators, and other individuals whose work brought them into close contact with youth.[12] The training program called the Tennessee Lives Count (TLC) was delivered to over 14,000 workers across the state. A sample of 630 gatekeepers drawn from the entire trained group was assessed before the gatekeeper training, immediately after, and six months later. There were three self-report outcome measures dealing with knowledge, self-efficacy in dealing with a suicidal youth, and beliefs about the inevitability of suicide. All measures improved immediately post-training, with some reductions in these improvements at six months, but still remaining well-above initial scores.

One can combine multiple pre-tests and multiple post-tests to create a study that looks like this: $O_1 –O_2 – X – O_3 – O_4$, a nice improvement over the $O_1 – X – O_2$ design. The pre-test / post-test design can also be improved by using more than one outcome measure. For example, a child-completed measure of depression could be supplemented by a parent's rating of their child's apparent depression, taken at about the same time period. Or an adolescent self-report measure of drug use could be supplemented by a urine screen at the same time. Obviously, direct measures of behavior are of greater credibility than self-reports of the same behavior (e.g., drug use, sexual activities) and should be used whenever possible as outcome measures. However, sometimes direct measures of behavior are not practical or the issue is subjective such as a feeling state like depression, anxiety, or self-esteem, variables that may not lend themselves readily to behavioral measurement. Another way to augment the value of a pre-test / post-test design is to increase the sample size. If an agency can obtain pre-and post-treatment data for 15 of its 100 clients, that is good. But if they can get similar data for 30, or 50, or even all 100 treated clients, that is even better. However, the practical inability to obtain complete data for all participants in a program should not preclude efforts to obtain what data one can. Some data is almost always better than none when it comes to program evaluation.

Quasi-experimental Designs

The pre-experimental designs described above involve the systematic assessment of the status of one group of child welfare clients who received an intervention—the $X - O_1$ design, and the assessment of one group of clients before and after treatment, the $O_1 - X - O_2$ design—and its variants with multiple pre-treatment assessments and/or post-treatment assessments. While a practical approach to seeing how clients are doing, or if they are improving, these designs are limited in their ability to answer more complex questions that relate to whether or not clients improve because of treatment. The goal is not only to ascertain if changes happened, but to try to plausibly determine the sources of such changes. Trying to answer this latter question is a much more ambitious undertaking. In order to attempt this, we need to control for various potential sources of bias or error, which cannot usually be accounted for using the pre-experimental designs. Technically, these potentially confounding factors are called threats to internal validity. Internal validity refers to the confidence one can have in determining the causes of observed changes. Some of the issues that can cloud causal inference are discussed below.

The Passage of Time

Some problems facing our clients tend to go away on their own. If a group of clients is assessed, receive an intervention, and are reassessed later, there may be improvements. Are these due to the intervention, or the passage of time? It is sometimes hard to tell. For example, if 100 teenagers enroll in the job-finding program, and 3 months after the program, 70 of the youth are now employed, it could be the effect of the program, or of the teenagers' independent initiatives at locating work. Finding and reporting that 70 percent of clients obtained work after participating in the program would be a great outcome to know and report, but be cautious and use conservative language in making conclusions. Reporting that "70 out of 100 unemployed teenagers seeking work found employment three months after participating in our agency's job-finding program" is a truthful statement. It would not be truthful to say "Our program resulted in 70 out of 100 youth finding employment." Agencies cannot honestly make this latter claim with the pre-test / post-test design.

Maturation

Some problems resolve, as one grows older, irrespective of treatment. Suppose a mental health agency offered group therapy for enuretic children, (children who wet the bed) aged 5 or older. At the beginning of the program, all the children were regularly wetting the bed. After 6 months of group therapy, only 30 percent were still wetting the bed. The unsophisticated child welfare worker might be tempted to claim, "My group therapy program cured 70 percent of the children with bedwetting!" The problem with this claim (inference) is that many five-year old children who wet the bed now will grow out of this problem as they develop. The sophisticated child welfare worker may truthfully claim that "70 percent of our enuretic children stopped wetting the bed after 6 months of treatment."

Concurrent Events

Things happen in people's lives outside of child welfare intervention, such as events with local, regional, or even national ramifications. Think of terrorist attacks, hurricanes, factory closings, presidential elections, or even the seasons of the year. In some Asian countries, adolescent suicide spikes near examination times, or when college acceptance (and rejection) letters are mailed. The Christmas season is different than other times of the year. Thus, if a child welfare agency begins a therapy program and after three months sees changes in its clients receiving services - are any positive changes due to treatment or something else going on in the clients' lives? This question is very hard to answer using only a pre-experimental design with just one group of clients.

Placebo Effects

For many conditions, especially those related to mental health and mood, clients respond positively to: attention, being listened to, the opportunity to have their concerns heard in a non-judgmental way, the opportunity to ventilate, and other similar factors. If you add to the picture a confident therapist with experience and diplomas hanging on the wall, a pleasing demeanor, trained in a therapy, and a plausible explanation for the problem, the stage is set for some possibly remarkable improvements that have literally nothing to do with the effectiveness of the treatment itself. The term placebo means "to please" and it refers not simply to sugar pills but more broadly to any intervention, medical or psychosocial. In order for a given intervention to be considered genuinely effective, it must produce effects significantly greater than effects induced by a similar treatment known to be simply a placebo. Placebo treatments are rarely used in evaluating child welfare interventions, given the seriousness of many of the situations our clients find themselves in. Placebo effects nevertheless need to be taken into account when appraising the possible effects of a given treatment.

Desire to Please the Therapist

Clients receiving child welfare services, whether child, youth, caregiver or parent, are usually aware of the efforts made on their behalf to assist them by a staff member or therapist. They know the investment of time and emotion made by counselors in providing services. When services are concluded and clients are asked how they are doing now, even if using some sort of standardized rating scale or measure for this purpose, the clients may provide artificially high appraisals of their feelings or functioning, in order to please the therapist. This can cloud understanding the 'true' effects of any treatment.

Regression to the Mean

It is common for clients to seek help, or to become involved with child welfare services, when their problems are at the worst, or following some crisis, such an episode of abuse, a drug overdose, or a criminal act. When an issue is thus 'peaked', it can be expected in the ensuing weeks or months for things to return to their more normative (even if problematic) level of functioning. This complicates interpreting the outcomes of a pre-test / post-test design. Clients are assessed pre-treatment and things are severe. They are assessed post-treatment and things are better. Is this due to the intervention or to the client/system functioning 'regressing' to the previous average level? At times, this is difficult to determine.

Collectively, these threats to internal validity make it complicated to determine the actual effects of treatment, above and beyond non-treatment-related influences. Lilienfeld et al. provides an overview of these issues which explains why we cannot simply ascribe client improvements solely to the results of therapy.[13] Yes, an apparent positive result can be attributed to the treatment. But, could it also be due to placebo, passage of time, regression to the mean, etc? Unless these factors are somehow controlled for, our ability to unambiguously claim that child welfare services caused positive outcomes is compromised, and, scientifically, should not be asserted. How can these threats to internal validity be controlled for? One good answer is to use a control group.

Control Groups

Control groups can potentially help us take into account some of these threats to internal validity. Using similar assessment protocols, comparing an untreated group of clients that is demographically comparable and has similar problems and severity; to the group receiving treatment; enables the program evaluator to better determine what caused the changes in the control group. In effect, the evaluator is able to 'subtract' any changes seen within the control group from any changes seen within the treatment group. This process allows a better estimate of the real effects of treatment. Collectively, program evaluation designs that make use of one or more control groups are labeled *quasi-experimental designs*. The term quasi is used because these designs are not quite true experiments. In true experiments, control groups are created by randomly assigning clients to various conditions (real treatment versus a waiting list control condition, for example). In quasi-experiments, groups are developed or created *without* using random assignment. Here are some examples of quasi-experimental evaluation designs.

Examples of the post-test-only no-treatment control group design

In this simple quasi-experimental design, one group receives treatment and is assessed using one or more reliable and valid outcome measures. A second group of clients, similar to the first, does not receive treatment but is assessed similarly to the first, or treatment group. This design is diagrammed as follows:

$$\begin{array}{c} X - O_1 \\ - - - - \\ O_1 \end{array}$$

Here is a hypothetical example. High school students are offered a class to help them prepare to take the Scholastic Aptitude Test (SAT). On a voluntary basis, some take the class, and some do not. Everyone takes the SAT in the fall. The school counselors examine the scores for all of the students, sorting them into the group that took the SAT prep course (the X group) and those that did not (the no-treatment group). They look to see if those who attended the SAT prep class had better average SAT scores than those who did not attend. If this hypothesis was supported, it would provide some evidence that the SAT prep course was effective. If there was no difference, it would be evidence that the SAT prep course was *not* effective in raising scores. If those who did not take the course did better than those who took it, that might suggest that taking the prep class actually negatively impacted SAT performance. But, none of these conclusions can be seen as very strong. Because, it may be that the higher-motivated students from higher socio-economic backgrounds were more likely to take the prep class, and their better scores were due to pre-existing advantages, and not because of the effectiveness of the class. In a study like this, with two groups and outcome measures scaled as mean scores, an independent sample *t*-test can be used to see if the two groups differ at the post-treatment assessment. A simple online calculator to conduct this type of test can be found here: https://www.graphpad.com/quickcalcs/ttest1/?Format=SD.

Inputting the previously calculated sample size, mean, and standard deviation of your outcome measure for your two groups, and selecting 'calculate' will provide the results. For example, taking the hypothetical example above, assume you have two groups of 100 teenagers each. The teens who attended the SAT prep class scored, on average, 1200 (standard deviation of 100) and those who did not go to the prep course scored 900 (standard deviation of 100). Entering these figures into the online calculator reveals that this difference is statistically significant at less than the .001 level, meaning that the SAT prep group's scores were indeed better off than the non-prepped students.

Rosenbaum used an example of this type of evaluation design by comparing the outcomes of 289 adolescents aged 15 and older, who had taken virginity pledges as a part of an abstinence-based sex education program.[14] The study also included outcomes of 645 non-pledgers. Outcomes included measures of premarital sex, sexually transmitted diseases, and participation in oral and anal sex, among other factors. This study found no differences between the students who took pledges to abstain from pre-marital sex compared to those who did not make pledges. Given that high school students could not be randomly assigned to take a virginity pledge, this *post-test only no-treatment design* was a very practical way to evaluate this element of abstinence-oriented sex education. An identical design was used to evaluate virginity pledges and found a modestly protective effect for pledges.[15] At the three-year follow-up, 42 percent of the non-pledgers had initiated sexual intercourse, compared to 34 percent of the pledgers, which was a small, yet real difference.

Having a no-treatment control group like this partially helps to control for some threats to internal validity, such as the passage of time, maturation, and concurrent events. It cannot control for placebo influences or the desire to please the therapist.

Examples of the post-test only comparison control group design

In this approach, one group of child welfare clients receives a new intervention of interest and a comparable group of clients receives the standard program currently being offered by the agency. Both groups receive the intervention and are assessed post-treatment at about the same points in time. This design would look as follows:

$$\begin{array}{c} X - O_1 \\ - - - - \\ Y - O_1 \end{array}$$

The idea here is that if clients who received X had better outcomes than those that received Y, X might be considered a better intervention than the current, usual practice. Other factors need to be taken into account, such as the relative cost of X compared to Y; the amount of training and expertise needed to implement X versus Y; was X equally acceptable to the clientele as was Y; and so forth.

This design was used by Georgiades to evaluate comparative outcomes of an Independent Living (IL) program for foster youth in Florida's District 11.[16] Forty-nine youth who completed the IL program (the X group in the above diagram) were contacted some years into adulthood. Eighteen youth who did not participate in the IL program (the Y group) but received regular foster care transition services were also contacted. Most participants were African-American women and had spent on average 8 years in 7 different foster care placements. These adult 'alumni' of the Florida foster care system were contacted by mail and they returned a completed measure of life skills. This measure assessed constructs such as money management, job seeking and maintenance skills, and social skills. The IL group were found to be better educated, employed more, living more independently, earning a higher monthly income, and more likely to have a driver's license, compared to the non-IL group. These outcomes suggest that the IL program was indeed better in assisting youth transition from foster care in Florida to attain adult independence.

An expanded application of this design was used to evaluate outcomes of various methods of dealing with child abuse investigations. Researchers sampled from Orlando, Florida, and obtained records from closed child abuse and neglect cases that were opened over a five-year period. Some of the cases (*n* = 59) were investigated using a Child Advocacy Center (CAC) model involving a multidisciplinary team of professionals from the fields of law enforcement, social services, mental health care, and child protection. Other cases were investigated using a Child Protection Team (CPT, *n* = 72 cases), and others were processed using a traditional child abuse investigation approach operated by Florida's Department of Children and Family (DCF, *n* = 53 cases). Each case was assigned one approach to conduct the investigation non-randomly, which makes this a quasi-experimental study. Overall, the children had an average age of 8 years, ranging from birth to age 18. The design can be diagrammed like this:

CAC Model (*n* = 59)	$X - O_1$
	– – –
CPT Model (*n* = 72)	$Y - O_1$
	– – –
DCF Model (*n* = 53)	$Z - O_1$

Data were collected on the number of substantiated (e.g., verified) cases, number of days in the system until substantiation status was determined, and arrest of perpetrator in substantiated cases. Another variable collected was reports of re-victimization within two years of case closure. This study is an example of what could be called *clinical data-mining*, going to existing agency records and extracting information. It did not involve collecting new data from children and their families. Sometimes data like this are available via paper files, but agency data are usually maintained in electronic records. By examining outcomes of routinely offered agency services, broken down by different types of services provided, one can empirically examine which approaches appear to be more effective in daily practice. By gathering information from many cases, the program evaluator is in a better position to obtain an objective appraisal of service outcomes, transcending to some extent personal bias, preferences, opinion, and conflicts of interest. The outcomes of this study found that the interdisciplinary models, CAC and CPT had higher levels of subsequent substantiated abuse and neglect than the traditional DCF model. Cases were closed more rapidly with the CAC and CPT models. Perpetrator arrests in subsequent substantiated cases did not differ across models.

Sometimes data are presented in the form of numbers or percentages, not means. In such instances, a Chi-square test (aka χ^2) may be an appropriate inferential statistical test to use. Take for instance, a group of substance abusing teenagers—some of whom participate in a drug rehabilitation program and others do not. Assume that there are 100 teens per group (the numbers do not have to be equal size), and that after the treatment group has been in the program for three months, all participants are required to complete urine tests. If 20 of the treatment teens turn up with positive results (i.e., they used drugs) but 80 of the teens non-treatment group had positive results, one can use a simple online χ^2 test calculator to determine if this difference is statistically significant (it is, favoring the treatment group). See http://www.socscistatistics.com/tests/chisquare/Default2.aspx

Examples of the pre-test / post-test no-treatment control group design

In this design, one group of clients receives a child welfare intervention and a second group does not receive it. Each group is assessed using one or more reliable and valid outcome measures pertinent to their problem, at about the same point in time before treatment. The treatment group then receives the intervention, and the no-treatment group does not. Then, at an appropriate time after the treatment group has completed treatment, both groups are reassessed using the same measure. This design can be diagrammed as follows:

$$O_1 - X - O_2$$
$$- - - - - -$$
$$O_1 \qquad O_2$$

This design was used to evaluate the potential effects of an approach to foster parent training known as the Model Approach to Partnerships in Parenting (MAPP). Child welfare staff who had been certified to provide the 10-week-long MAPP training program were used in this study. Seventeen participants had volunteered to become foster parents and 12 members of the local county foster parents' association served as the no-treatment control condition. All participants completed the Adolescent/Adult Parenting Inventory at the same two points in time, before and after the MAPP training for the treatment group. Simple inferential statistics were used to determine if the two groups were similar at the pre-treatment assessment, and if the MAPP training group had improved post-training. Any child welfare agency providing foster parent training can readily incorporate the measurement of their aspiring foster parents of their parenting skills and attitudes to assess whether their skills improve post-training. In selected instances, a suitable comparison group that does not receive the training can be recruited.

Another real-life example of this design was used to evaluate an infant simulator program (Baby Think It Over) designed to help teenagers obtain a more realistic sense of the demands that parenting an infant would require of them. Six teenagers received the infant doll, along with appropriate supplies, diapers, diaper bag, clothing, bottles, etc. This small sample size was dictated because there were only six dolls available. Periodically the life-size doll would cry, requiring action on the part of the teen to alleviate the crying (e.g., taking the baby's temperature with a fake thermometer). The six were assisted with the simulated infant care by 17 other student helpers. Twenty-five other student

volunteers did not get assigned the simulated baby (but were offered it after the study was over). The outcome measure was the Parenting Attitude Scale, a ten-item scale. Pre-treatment, the two groups were similar, but post-treatment the six teens and their helpers had much more realistic expectations about the demands and responsibilities of childcare, the intended purpose of the exercise.

The Switching Replications Delayed Treatment Control Group Design

If there is concern about not providing a child welfare intervention to a no-treatment control group, evaluators could make use of this Switching Replications design, diagrammed as follows:

$$\begin{array}{ll} O_1 - X - O_2 & O_3 \\ - - - - - - - - - - - - \\ O_1 & O_2 - X - O_3 \end{array}$$

The logic of this design follows these lines. At O_1, the pre-treatment assessment, the two groups are equivalent on both demographics and scores on the outcome measure(s). The immediate treatment group gets treatment and the delayed treatment group gets nothing. After the first group completes intervention, both groups are reassessed at time O_2, finding ideally, Group 1 has improved and Group 2 has not changed. Then Group 2 gets the same intervention, and both groups are reassessed a third time, at O_3, hopefully finding that Group 2 has now improved to an extent equivalent to the improvements seen at O_2 by Group 1 and that Group 1's gains seen at O_2 are maintained at time O_3. This is an elegant design that mitigates concerns of using a no-treatment control group. By demonstrating an apparent effect of treatment twice within the same study, internal validity is enhanced. It is not uncommon for a child welfare agency to have a group of clients on its waiting list for services. Sometimes program evaluators can take advantage of this fact to create a waiting list control group design such as this.

The Pre-test / Post-test Comparison Group Design

This type of design is used to compare outcomes of two or more treatments provided by a child welfare agency. It can be diagrammed as follows:

$$\begin{array}{l} O_1 - X - O_2 \\ - - - - - - \\ O_1 - Y - O_2 \end{array}$$

This design was used to evaluate the Parenting with Love and Limits (PLL) program for juvenile offenders. 155 youth between the ages of 14-18 were referred for treatment and received PLL. A comparison group of 155 similar youth received treatment as usual with the state juvenile justice system. The outcome measures included the Child Behavior Checklist (CBCL) and recidivism. The PLL treatment lasted about six weeks. CBCL scores dramatically improved after PLL treatment and one-year recidivism rates were comparatively better among the PLL group. Thus in this study, PLL could be considered the X treatment in the above diagram, and treatment as usual, the Y comparative treatment condition. No youth was denied care, and this study showed that PLL produced greater improvements than standard therapy.[19]

Experimental Designs

While the above quasi-experimental designs are an improvement over the one-group pre-experimental designs, they are still not as tight as could be desired in terms of controlling for threats to internal validity. A substantial improvement over quasi-experiments is to conduct a true experiment. The sole distinction of being in a true experiment is that the groups are deliberately constructed by the program evaluator using random assignment methods. In random assignment, one develops methods to assure that each potential participant has an equal chance to be assigned to any of the two or more conditions (e.g., active treatment, treatment as usual, no treatment, placebo treatment, etc.). One can do this simply by tossing a dice. There are also random number generators online, and tables of random numbers found in statistics texts that can be used for this purpose. It is essential that some non-arbitrary method be employed to conduct random assignment—you cannot simply assign every other new client to each group, or Monday's clients to group one and Tuesday's clients to group 2, even though that may seem 'random'.

Randomized experiments have been given a bad reputation in the human services. It is frequently claimed that these designs are very rare, impractical, unethical, insensitive, or too time consuming. None of this is true. A recent bibliography of true experiments published within the field of social work located over 740 such studies, many of which took place in child welfare contexts.[20] The earliest such true experiment conducted by social workers was published in 1949, and involved a juvenile delinquency prevention program that began in the 1930s. Thus, the history of using randomized experiments in the social services rivals the duration of their use in other fields such as medicine and clinical psychology.

Randomized experiments are diagrammed exactly as are quasi-experiments, except the letter R is placed in front of each group, and the dashed line separating the groups is omitted. For example:

The post-test-only no-treatment control group design, is diagrammed as:

$$\begin{array}{ll} R & X - O_1 \\ R & O_1 \end{array}$$

The post-test only comparison control group design:

$$\begin{array}{ll} R & X - O_1 \\ R & Y - O_1 \end{array}$$

Each of these designs can be conducted retrospectively (involving the analysis of existing agency data) or prospectively (involving the analysis of new data gathered in a planned manner for the purposes of evaluation). Each can include follow-up assessments completed sometime after assessments and completed immediately post-treatment, to evaluate the durability of any gains.

The Pre-test / Post-test No Treatment Control Group Design:

$$R \quad O_1 - X - O_2$$
$$R \quad O_1 \quad O_2$$

The Pre-test / Post-test Comparison Group Design

$$R \quad O_1 - X - O_2$$
$$R \quad O_1 - Y - O_2$$

The Switching Replications Delayed Treatment Group Design

$$R \quad O_1 - X - O_2 \quad O_3$$
$$R \quad O_1 \quad O_2 - X - O_3$$

As with quasi-experiments, these true experimental designs variations can build upon each other, use more than two groups (e.g., a new active treatment condition, a standard care condition, a placebo condition, a delayed treatment group), and have more than one pre-test and/or post-test assessment (e.g., a follow-up period). The methods of statistical analysis are generally the same, t-tests for independent groups looking at mean differences on outcomes for the two groups post-treatment only study, or using χ^2 tests with frequency-based data. With three or more groups, and outcome measures scored as mean values, a more complicated inferential test is appropriate—something called an Analysis of Variance (ANOVA). With a pre-test / post-test no treatment control group design, and outcome measures calculated as means, a 2 (meaning two groups) by 2 (meaning two time periods) ANOVA would be used. If there had been three conditions—new treatment, treatment as usual, and no treatment—and assessments conducted twice, before, and after treatment, the test would be a 3 (groups or treatments) by 2 (time periods when assessment occurred) ANOVA.

If there was a pre-test and post-test, a follow-up assessment, and two experimental conditions, the test would be a 2 (groups) by 3 (time periods) ANOVA. But, most program evaluations conducted within child welfare agencies are considerably more simple than these latter designs. Those that do undertake more ambitious projects, such as true experiments with conditions created using random assignment, usually have a statistical consultant (perhaps a nearby university) on hand to help with the analysis.

How to Prepare a Useful Report of a Local Outcome Evaluation

The best way to learn how to write a program evaluation article is to carefully read program evaluation articles that report outcomes of investigated services that your agency is interested in evaluating. By doing there is opportunity to learn about writing style, how to structure an evaluation report, how much detail to include, or not include, and so forth. The *Publication Manual of the American Psychological Association*[21] contains a section at the very end of the book called the *Journal Article Reporting Standards* (JARS, pp. 247-250). The JARS are checklists of content that needs to be included in each section of a research report—the title page, abstract, the introduction, description of methods, participants, sampling method, outcome measures, results, and discussion. Each child welfare agency seeking to write program evaluation reports should have a hard copy of the APA manual at hand. The APA also has extensive help available online (see http://www.apastyle.org/index.aspx?_ga=2.145534516.1226729163.1502900938-2093779991.1494782591), including a link to the JARS (see http://www.apastyle.org/manual/related/JARS-MARS.pdf)

Reviewing the JARS as before writing each section of the report, can ensure that all essential information is included. Once the report is completed, give it to a colleague, along with the JARS, so it can be double checked.

Do not use a published paper as an infallible guide to adhering to APA formatting. APA guidelines are for preparing *manuscripts* for publication, and are not always adhered to the final published product. For example, in APA style, we are told that tables should be placed, one per page, after the reference list. But in an article published in a journal that uses APA style, the tables are usually imbedded in the text, at their relevant location and not after the references at the end. Read published papers to gain a sense of writing style, but follow the APA manual to format the manuscript itself.

There are many books and articles that describe how to write research reports and a few of these are listed in the Appendix.

Recommendation for Policy and Practice

It is clear that both the federal and state governments focus on obtaining measurable outcomes of the services they provide or purchase. This reflects a true sense of caring about the children and families served by the child welfare system and fiscal prudence. It is irresponsible to spend large sums of money on services that do not work, or worse yet, may have harmful effects. Private foundations, the United Way, and non-governmental organizations are all promoting the mantra of measure, evaluate, improve. The model of evidence-based practice is increasingly relied upon in the areas of health and social care. This means not only making use of existing research evidence to help guide the selection of services provided (combined with considering clients' preferences and values, professional ethics, available resources, costs, and other equally important considerations), but also the expectation that funded agencies engage in regular, systematic, and objective methods of outcomes evaluation. This information should be readily available to stakeholders and used to inform and revise existing services.

The examples contained in this report illustrate that the potential to conduct pre-experimental and quasi-experimental studies of outcomes in child welfare is within the scope of any agency. The following policy is offered for individual child welfare agencies to consider:

> *Regardless of other reporting mandates from Community-based Care entities, the Florida Department of Children and Families, or accrediting bodies, every child welfare agency should select one or more credible measures that assess aspects of child*

or family functioning that agency hopes to positively change. Efforts should be made to administer this/these measures during the intake/assessment period, before the client becomes fully engaged in treatment or care. A reasonable benchmark or time frame should be chosen for re-administering these measures, such as at the termination of care, discharge, emancipation, or after every three months in cases of continuing care. Each agency could choose suitable benchmarks appropriate for its clientele and region of the state. Periodically, perhaps every six months, or annually, these pre-test and post-test data are aggregated for each program of an agency and a clearly written report describing the results is prepared and made available to all stakeholders. The results would be reviewed by agency administrators and used to make decisions about the revision, expansion or contraction of a particular program. Each agency could decide to attempt to publish their data in a suitable professional journal.

As this report illustrates, many potential outcome measures already exist and are waiting to be used by child welfare agencies. However, barriers may also exist. Clients may not wish to complete an agency's measures. Staff may forget to complete a pre-assessment or post-test assessment. Such instances will result in incomplete data. This should not deter you from making the effort, and reporting the available results, while acknowledging information gaps.

Summary

This technical report has reviewed the importance of child welfare agencies undertaking periodic program evaluations of their outcomes. It is important that such evaluations make use of reliable and valid outcome measures that directly pertain to the agency's purposes, and be as specific as possible. Variables that are very broad and general are not as useful as a more direct measure of clients' serious issues. Information was provided on how to locate suitable outcome measures. This was followed by a review of relatively simple and practical pre-experimental and quasi-experimental research designs that are widely used in program evaluations. Fundamental information on conducting statistical analyses of program outcomes was also provided. Experimental designs, studies involving randomly assigning clients or family to differing treatment conditions were briefly described, but their use is less emphasized in favor on the more practical pre-experimental and quasi-experimental evaluation studies. A policy recommendation was made that each child welfare agency should attempt to systematically collect pre- and post-service outcome measures on their clients' functioning, periodically assess this aggregated information, and use this to make empirically-based decisions about the need to revise, expand, contract or terminate existing services. The approach outlined in this report is consistent with large-scale initiatives on the part of the federal and state governments, CBCs, non-profits, and the United Way to request empirical outcomes data on funded services. Faculty at nearby universities are available to provide free or low-cost consultation to child welfare agencies in the design and conduct of program evaluation studies. The approach outlined in this report is not new. Similar suggestions have been widely made in many different fields of health and social care.

Appendices

Resources on Writing up Program Evaluation Studies

Furman, R. & Kinn, J. T. (2012). *Practical tips for publishing scholarly articles*. New York: Oxford.

Grinnell, R., Gabor, P. & Unrau, Y. (2012). Program evaluation for social workers (6th edition). New York: Oxford University Press.

Pietzak, J., Ramler, M., Renner, T., Ford, L. & Gilbert, N. (1990). Practical program evaluation: Examples from child abuse prevention. Thousand Oaks, CA: Sage.

Pyrczak, F. (2014). Writing empirical research reports. New York: Routledge.

Rosnow, R. L. (2011). Writing papers in psychology (9th edition). Belmont, CA: Wadsworth.

Rossi, P. & Williams, W. (1991). Evaluating social programs. New York: Seminar Press.

Royse, D., Thyer, B. A. & Padgett, D. (2016). Program evaluation: An introduction to an evidence-based approach (6th edition). Belmont, CA: Cengage.

Shadish, W. R. Cook, T. D. & Leviton, L. C. (1991). Foundations of program evaluation. Thousand Oaks, CA: Sage.

Silvia, P. J. (2007). How to write a lot: A practical guide for productive academic writing. Washington, DC: American Psychological Association.

Silvia, P. J. (2014). Write it up: Practical strategies for writing and publishing journal articles. Washington DC: American Psychological Association.

Thyer, B. A. (1994). Successful publishing in scholarly journals. Thousand Oaks, CA: Sage.

Thyer, B. A. (2008). Preparing research articles. New York: Oxford University Press.

Thyer, B. A. (2012). Quasi-experimental research designs. New York: Oxford University Press.

Thyer, B. A. (2015). Evaluating school social work. In P. Allen-Meares (Ed.). Social work services in schools (7th edition, pp. 297-326, 401-404). New York: Pearson.

Thyer, B. A. & Myers, L. L. (2007). A social worker's guide to evaluating practice outcomes. Alexandria, VA: Council on Social Work Education.

Thyer, B. A. & Myers, L. L. (2016). Linking assessment to outcome evaluation using single-system and group research designs. In C. Franklin & C Jordan (Eds). Clinical assessment for social workers: Quantitative and qualitative methods (4th edition, 345-366). Chicago: Lyceum.

References

1. Mercer, J. (2017). Evidence of potentially harmful psychological treatments for children and adolescents. *Child and Adolescent Social Work Journal, 34*, 107-125.

2. McCrae, J. S. & Brown, S. M. (in press). Systematic review of social-emotional screening instruments for young children in child welfare. *Research on Social Work Practice*.

3. Corcoran, K. & Fischer, J. (Eds.) (2013). *Measures for clinical practice and research*. New York: Oxford University Press.

4. Nezu, A. M., Ronan, G. F., Meadows, E. A. & McClure, K. S. (Eds.) (2000). *Practitioner's guide to empirically based measures of depression*. New York: Springer

5. Pabian, W. E. et al. (2000). Do the families of children with developmental disabilities obtain recommended services? A follow-up study. *Journal of Human Behavior in the Social Environment, 3*(1), 45-58.

6. Robinson, R. M., Powers, J. M., Cleveland, P. H. & Thyer, B. A. (1990). Inpatient psychiatric treatment for depressed children and adolescents: Preliminary evaluations. *The Psychiatric Hospital, 21*(3), 107-112.

7. Gerardot, R. J., Thyer, B. A., Mabe, P. A. & Poston, P. M. (1992). The effects of psychiatric hospitalization on behaviorally disordered children: A preliminary evaluation. *The Psychiatric Hospital, 23*(1), 65-68.

8 Allison A. & Ferreira, R. J. (2017). Implementing Cognitive Behavioral Intervention for Trauma in Schools (CBITS) with Latino youth. *Child and Adolescent Social Work Journal, 34*, 181-189.

9 Dauber, S., Lotsos, K., & Pulido, M. L. (2015). Treatment of complex trauma in the front lines: A preliminary looks at child outcomes in an agency sample. *Child & Adolescent Social Work Journal, 32*, 529-543.

10 Brown, D. M. (2017). Evaluation of Safer, Smarter Kids: Child sexual abuse prevention curriculum for kindergarteners. *Child and Adolescent Social Work Journal, 34*, 213-222.

11 Greaves, J. R. & Salloum, A. (2015). Evaluation of a Youth with Sexual Behavior Problems (YSBP) outpatient treatment program. *Child and Adolescent Social Work Journal, 32*, 177-185.

12 Keller, D. P., Schut, L. J., Puddy, R. W., Williams, L., Stephens, P. L., McKeon, R. & Lubell, K. (2009). Tennessee Lives Count: Statewide gatekeeper training for youth suicide prevention. *Professional Psychology, Research and Practice, 40*, 126-133.

13 Lilienfeld, S. O., Ritschel, L A., Lynn, S. J., Cuatin, R. L. & Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. *Perspectives in Psychological Science, 9*, 355-387.

14 Rosenbaum, J. E. (2008). Patient teenagers? A comparison of the sexual behavior of virginity pledgers and matched nonpledgers. *Pediatrics*, 123, e110-e120

15 Martino, S. C., Elliott, M. N., Collins, R. L., Kanouse, D. E. & Berry, S. H. (2008). Virginity pledges among the willing: Delays in first intercourse and consistency in condom use. *Journal of Adolescent Health, 43*, 341-348.

16 Georgiades, S. (2005). A multi-outcome evaluation of an independent living program. *Child and Adolescent Social Work Journal, 22*, 417-439.

17 Wolfteich P. & Loggins, B. (2007). Evaluation of the Children's Advocacy Center model: Efficiency, legal and revictimization outcomes. *Child and Adolescent Social Work Journal, 24*, 333-352.

18 Lee, J. H. & Holland, T. P. (1991). Evaluating the effectiveness of foster parent training. *Research on Social Work Practice, 1*, 162-174.

19 Karam, E. A., Sterrett, E. M. and Kaier, L. (2015). The integration of family and group therapy as an alternative to juvenile incarceration: A quasi-experimental evaluation using Parenting with Love and Limits. *Family Process, 56*, 331-347.

20 Thyer, B. A. (2015). A bibliography of randomized controlled experiments in social work (1949 – 2013): *Solvitur Ambulando. Research on Social Work Practice, 25*, 753-793.

21 American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th edition). Washington, DC: Author.

22 Collins-Camargo, C., Sullivan, D. & Murphy, A. (2011). Use of data to assess performance and promote outcome achievement by public and private child welfare agency staff. *Children & Youth Services Review, 33*, 330-339.